Binary classification of HRV signals

Rebecca Lütz

June 17, 2019

Populärvetenskaplig sammanfattning

HRV är en förkortning av det engelska begreppet ”Heart rate variability” och beräknas med hjälp av data från ett elektrokardiogram, EKG. Detta mäter elektriska signaler som skickas ut från hjärtat. En tydlig sammandragning av hjärtat kallas för puls och det syns tydlig på ett EKG som en ökning i den elektriska signalen. I motsats till pulsen, är det inte själva hjärtslagen som är viktiga när man beräknar HRV utan det som händer emellan dem. HRV kan därför definieras som variansen av den elektriska signalen från hjärtat mellan två tydliga sammandragningar som vi kallar puls. Dessa intervaller mellan hjärtslagen är inte identiska vilket leder till en naturlig varians. Men stora avvikelser inom den naturliga variansen kan tyda på att någonting så som stress eller sjukdom påverkar kroppen. Till exempel leder stress till att variansen minskar.

Datan som har använts i denna uppsats är tagen från en studie i Kristanstad där 53 personer placerade sin hand i iskallt vatten. Detta skulle simulera stress och visa hur stress påverkar HRV. För att få ett kontrollset utfördes en liknande test med varmt vatten vilket visade HRV i viloläget.

Denna uppsats behandlar binära klassificeringar vilket innebär att data blir antingen klassificerad som positiv eller negativ. I vårt fall betyder en positiv klassificiering att signalen har blivit klassificierad som kall. Den negativa klassificieringen betyder således att signalen har klassificierats som varm. Målet med denna uppsats är att få fram en metod för lyckad klassificiering som är baserad på en frekvensanalys. Detta betyder att man analyserar HRV signalen med fokus på vilka frekvenser som finns och hur mycket energi som har ackumulerats vid dessa frekvenser.

Medan olika metoder leder till olika bra resultat är det ändå väldigt tydligt att det verkar finnas parametrar utifrån vilka man kan korrekt klassificiera två okända HRV signaler. Denna slutsatsen kan dras eftersom alla metoder klassificierar minst 50% av datan rätt.

Abstract

Heart rate variability, commonly abbreviated as HRV, displays the variance between consecutive heartbeats. This variance occurs naturally but can change due to stress and problems with the cardiac system. HRV is therefore widely used for medical research. The goal of this thesis is to correctly classify two HRV signals where one is obtained at a resting state, the warm signal, while the cold signal is obtained during a simulation of stress. The use of spectral estimation methods leads to the analysis of the high frequency range (0.12 - 0.4 Hz) as well as the analysis of a more narrow frequency band around the respiratory maximum. The analysis of those frequency ranges is done by using linear models as well as studying how the energy of the cold and the warm signal is distributed. All approaches lead to binary classification with more than 50% accuracy. However, the best results are obtained when analyzing the frequency band around the respiratory maximum located at 0.2 Hz or higher. When using a linear model for changes in energy over time, dividing the data into four sets leads to 93.4% correct classification. When analyzing the energy that is present in the first 90 s of each signal, 96.23% correct classification is obtained.

Contents

1 Introduction

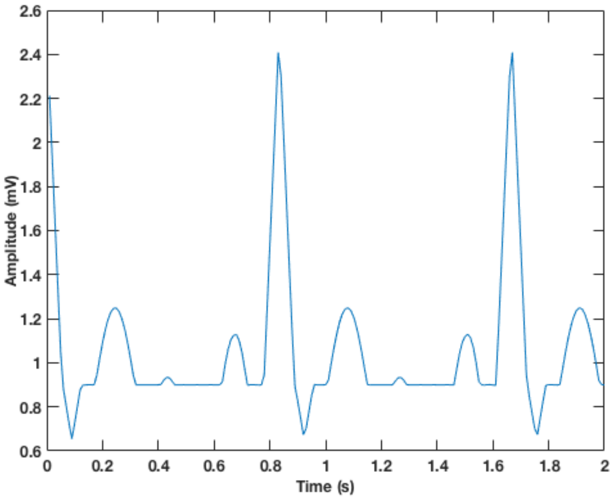

Figure 1: An example of a RR-interval obtained by an ECG-simulation.

Heart rate variability, or short HRV, has been a well researched topic for many years now. It deals with the interval between two consecutive heart beats, also known as the RR-interval. This means that we do not actually look at each heartbeat but what happens in between. A typical RR-interval is shown in Figure 1. The two peaks that are shown are also known as R-peaks and they indicate two consecutive heartbeats. The data to display a RR-interval is easily obtained from an electrocardiogram (ECG) signal, which measures the electric signals produced by the heart. We can then use the ECG signal to calculate the HRV. This is done by computing the variance between two consecutive heartbeats. Since the variance between two consecutive heartbeats is constant the plot is typically staircase shaped. An example is shown in Figure 2.

HRV first became of interest in 1965 when it was noted that the monitoring of HRV could foreshadow fetal distress, since changes within HRV could be observed long before changes in the actual heartbeat [1]. This lead to HRV being studied intensively.

The naturally occurring variation within HRV takes into account both the health of the autonomic nervous system as well as cardiac health [2]. It is therefore seen as a good measure of overall health, which has lead to it being the focal point of research about several illnesses. The idea behind this research is to pay attention to any arising abnormalities within the variation and analyze these, since some sicknesses lead to specific changes of the HRV. Smoking and drugs

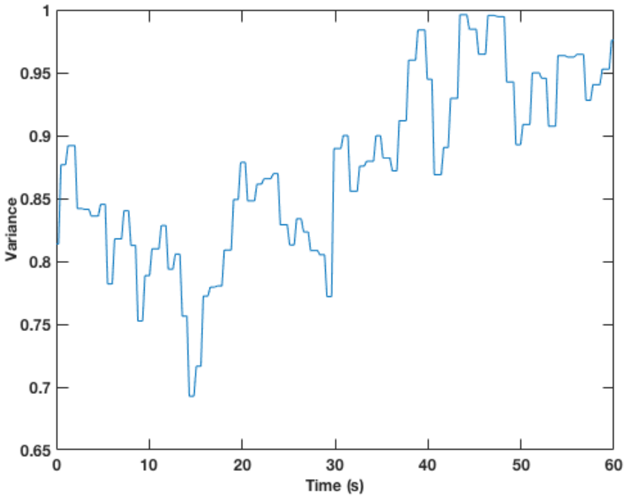

Figure 2: An example of an HRV signal obtained at resting state.

can also affect the normally occurring variation and HRV can therefore be used to assess even those damages [2].

When analyzing an HRV signal, you usually focus on either the low frequency (LF) range, the high frequency (HF) range or the ratio of those two, LF/HF [1]. For adults, everything between 0.04 - 0.12 Hz is defined to be in the LF range and is associated with sympathetic activity, for example exercise, which leads to an increased heart rate. Parasympathetic activity, on the other hand, is mostly reflected in the HF range, which is defined as 0.12 - 0.4 Hz. Parasympathetic activity can be described as the bodies' resting activities and thus leads to a decreased heart rate. Examples include the function of internal organs such as the digestion system [2]. However, it should be noted that HRV cannot be used to exactly measure either of those activities and at most it should be used as an indicator. Furthermore, research is still conducted on how much each frequency range really is affected by sympathetic and parasympathetic activity. However, in previous studies it has been observed that exposing your skin to a cold stimuli affects the HRV signal in the HF range [1]. In order to conduct successful research based upon the analysis of HRV data, breathing is often controlled during experiments due to its significant impact on the lengths of the intervals between each heart beat. During inhalation heart rate increases which leads to a shorter interval while the interval becomes longer during expiration due to decreased heart rate [3].

The data that is being used for this thesis comes from a study in Kristanstad, Sweden. 90 participants between the ages of 19 to 31, carried out a so called Cold Pressure Test where your hand is placed into ice-cold water for three minutes. During this time, measurements were taken continuously and they were then down sampled to a sampling frequency of 4 Hz. However, due to not everyone finishing the Cold Pressure Test, only the data of 53 subjects is actually being used. The mean age of those 53 participants is 23.23 and the variance is 7.4. For each of those participants, there also exists a control set where the same procedure was repeated with lukewarm water. During both tests, an ECG was taken and respiration measured. However, it is important to note that no guidelines on respiration were given. Thus, a control set containing a warm HRV signal and corresponding respiratory data was obtained as well as the HRV signal and respiratory data from the Cold Pressure Test. Given these two HRV signals, the goal of this thesis is to find a method that successfully classifies those as either warm or cold.

2 Theory

In this section, the theory that is needed for all further analysis of the data is discussed. This includes the definition of the spectral density and how to estimate it, the definition of a measure to judge the goodness of a binary classification as well as the definition of linear regression and its limitations.

2.1 Spectral density

The Fourier of the covariance function of a stationary process is known as the spectral density. The variance of the process is the total power of it. Thus, the spectral density tells us how the total power is distributed at specific frequencies. Hence, we have obtained a characterization of a given stationary process in the frequency domain.

Given $r(\tau)$, the covariance function of a discrete stationary process $x(t)$ for $t = 0, \pm 1, \pm 2, \ldots$, we can calculate its spectral density

$$R(f) = \sum_{i=-\infty}^{\infty} e^{-i2\pi f\tau} r(\tau),$$

where f is the frequency [4]. Here, we assume that $R(f)$ is symmetric, integrable as well as positive. However, since in reality we can not obtain infinitely many samples to compute the spectral density, we need to estimate it in some way.

2.1.1 Periodogram

The periodogram is used to estimate the spectral density of a zero mean process, $x(t)$ for $t = 0, 1, \ldots, n-1$ [4]. It is defined as

$$\hat{R}_X(f) = \frac{1}{n}|\chi(f)|^2,$$

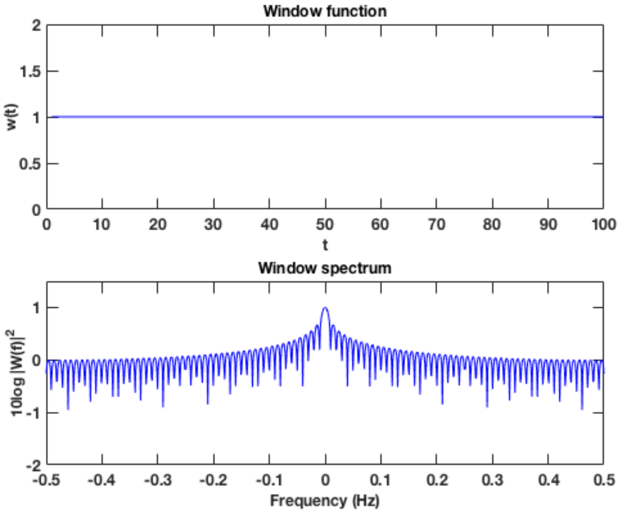

Figure 3: The rectangular window function and its corresponding window spectrum centered around 0 Hz. Here $W(f)^2$ denotes the Fourier transform of the window function $w(t)$.

where n is the length of the given data vector. The Fourier transform of $x(t)$ is $\chi(f)$ and is defined as

$$\chi(f) = \sum_{t=0}^{n-1} x(t)e^{-i2\pi ft}.$$

While the periodogram is a good starting point for spectral estimation, there are some disadvantages such as high variance and leakage of the side lobes. This leakage is very apparent when comparing the height of the side lobes of the window spectrum of the periodogram, seen in Figure 3, to the height of the side lobes obtained when using the Hanning window, shown in Figure 4. Here, both window spectra have been normalized to make comparison easier. Due to those problems, other methods, such as the modified periodogram and the Welch method, are more commonly used for spectral estimation.

2.1.2 Modified periodogram

The Hanning window, sometimes also called the Hann window, can be used to modify the periodogram [4]. The spectral estimate can then be rewritten as

$$\hat{R}_w(f) = \frac{1}{n}|\sum_{t=0}^{n-1} x(t)w(t)e^{-i2\pi ft}|^2,$$

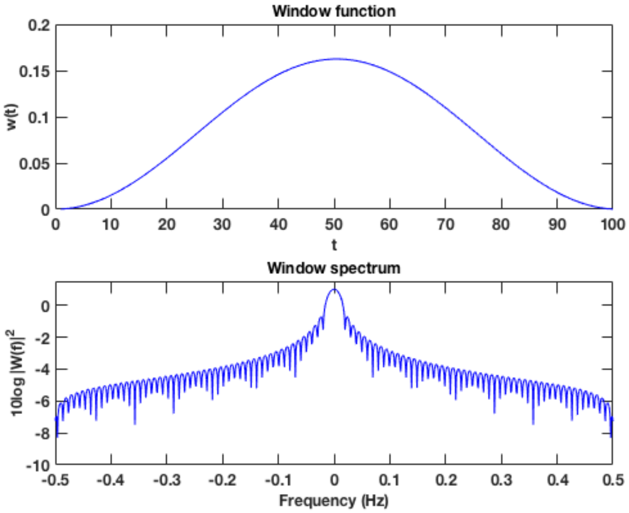

Figure 4: The Hanning window function and its corresponding window spectrum centered around 0 Hz, where $W(f)^2$ denotes the Fourier transform of the window function $w(t)$.

where $w(t)$ is the window function. When using the Hanning window

$$w(t) = \frac{1}{2} - \frac{1}{2}\cos(\frac{2\pi t}{n-1}),$$

where $t = 0, 1, \ldots, n-1$. Its main advantage over the periodogram are the lower side lobes in frequency, even if they come at the cost of the main lobe being twice as wide. Meaning that its width increases from $\frac{2}{n}$ to $\frac{4}{n}$. However, this is still very advantageous since the lower side lobes lead to a reduction of bias [4]. Both the increased width of the main lobe and the lower side lobes can be seen when comparing Figure 3 and Figure 4.

2.1.3 Welch method

Another often used method to estimate the spectral density is the Welch method [4]. The idea is to reduce the variance, since this leads to a more consistent estimate of the spectral density. This is done by dividing n data points into K sequences, each containing only L data points and where the sequences overlap p percent. The estimate of the averages therefore becomes

$$\hat{R}_{av}(f) = \frac{1}{K}\sum_{k=1}^{K} \hat{R}_{x,k}(f),$$

for $k = 1, \ldots, K$. Here, $\hat{R}_{x,k}(f)$ is the spectral estimate of the k^{th} data sequence. However, variance will only be reduced if the individual spectral estimates do

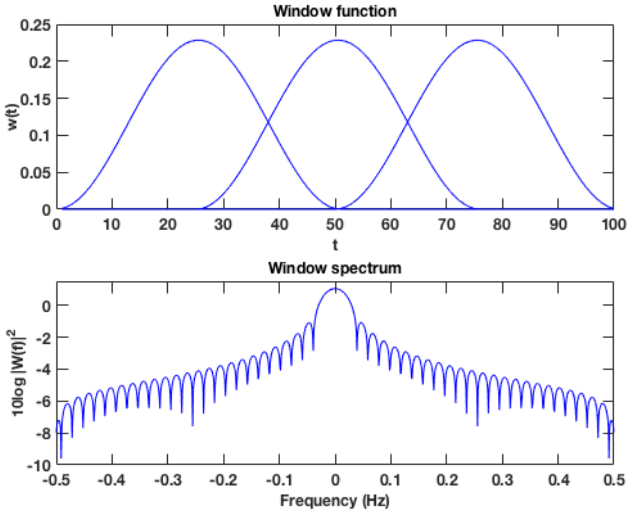

Figure 5: Window function and corresponding spectrum of three Hanning windows that overlap with 50%.

not have a high correlation. Compared to the periodogram, when having no overlap the Welch method gives a higher bias as well as a K times wider main lobe [4]. Using the Hanning window with a 50% overlap is shown to give the best results for the Welch method, meaning that we reduce the variance as much as possible while at the same time increasing the bias as little as possible [4]. An example of the window function and its spectrum when using three Hanning windows that are overlapping 50% is shown in Figure 5. It can also be noted that comparing the Welch method with no overlap to the Welch method with overlapping windows, the width of the main lobe decreases as the percentage of overlap increases.

2.2 Matthews correlation coefficient

When classifying data with a binary classification algorithm there are four possible outcomes. If the data belongs to the class "positive" and is predicted as "positive", we say that the result is true positive or TP. Similarly we say that an outcome is true negative or TN, if the data belongs to "negative" and is predicted as such. However, if something is falsely classified as "negative" even though it actually is "positive", the result is called false negative or FN. Vice versa the result is called false positive or FP, if data from "negative" gets classified as "positive". Those possible outcomes can be displayed in a confusion matrix, seen in Table 1.

Table 1: Confusion matrix.

	Positive	Negative
Predicted positive	TP	FP
Predicted negative	FN	TN

The Matthews correlation coefficient,

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

is based on the values of such a confusion matrix and is widely used in machine learning to measure the goodness of a binary classification algorithm [5]. The MCC takes on values between -1 and 1, where -1 is the worst, 0 can be interpreted as the classification being random and 1 means perfect classification. In order to achieve a high score for a binary classification algorithm, the algorithm needs to do well on both positive and negative predictions, which becomes increasingly important if the sets ”positive” and ”negative” differ greatly in size. This is also the main advantage of the MCC in comparison to other summarizing statistics [5].

2.3 Linear regression

Given data x and y, where y is real valued and called response variable and x are the explanatory variables, we want to find a linear model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i,$$

for $i = 1, \ldots, n$. In matrix form we thus get

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \ldots & x_{pn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{pmatrix},$$

where ϵ_i are Normal distributed independent identical random variables with mean 0 and constant variance σ^2. Using the least squares estimator we can solve for β such that

$$\hat{\beta} = \underset{\beta}{\text{argmin}} (\mathbf{Y} - \mathbf{X}\beta)^{\mathrm{T}} (\mathbf{X} - \mathbf{Y}\beta),$$

which has solution

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

Sometimes the residuals ϵ_i do not follow a Normal distribution. This can be due to a small data set or if there is a lot of variance within the given data. When this is the case, we can not use a linear regression model. However, we can use a linear model that is based on linear regression. Such a linear model then gives us a line of best fit,

$$y = \beta_0 + \beta_1 x.$$

3 Method

In order to be able to correctly classify two given signals, different approaches were tested. All of them used spectral estimates of the HRV signals and they were based on there being a common pattern within the data of the cold signal that was different to some pattern within the data of the warm signal. Those patterns on the other hand, were easier to identify once variance was reduced. It was therefore decided to estimate the spectral density by the Welch method using a Hanning window with 50% overlap. This resulted in smooth curves only showing the most important frequencies as shown in Figure 6.

Before implementing any methods, the mean was subtracted from the signals to get a zero-mean process. Additionally, the data was normalized so that the results later could be compared. This resulted in

$$X_{\mathrm{norm}} = \frac{X - \bar{X}}{\sqrt{(X - \bar{X})^{\mathrm{T}}(X - \bar{X})}},$$

where X is a matrix containing the data that was used and $\bar{X}$ the vector of the column averages of X. This normalization was done for both the HRV data and the respiratory data. All further calculations of the spectral estimate were carried out with the normalized data X_{norm}. However, the optimal number of windows, meaning how smooth the estimate eventually would be, changed for different methods. The possible range out of which the optimal number of windows was chosen, was determined by looking at the number of data points that each set contained as well as the lowest frequency that would be analyzed. In order to be able to draw reasonable conclusions, we wanted to see more than one period. While one period could seem sufficient, this could lead to loss of information which should be avoided. This loss of information could occur due to window functions, such as the Hanning window, not prioritizing the data on the edges of the set in the same way as the data in the middle due to the Hanning window being shaped like a bell curve. When calculating the maximum number of windows, it was thus chosen to try to show at least one and half periods, optimally two periods. The maximum number of windows, W, was calculated as

$$W = \frac{2n}{L} - 1.$$

Here, n denoted the length of the set that was used and L the number of data points in two periods. All results that were obtained for W were however rounded up resulting in not always showing two periods.

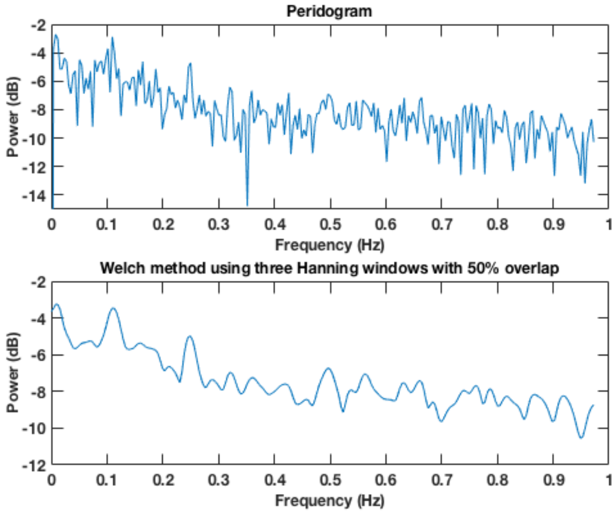

Figure 6: Comparison of periodogram and the Welch method of the same data set.

Using a linear regression model to analyze the data was initially considered. Here, the energy of each set was seen as the response variable, y, and the number of sets as the explanatory variable, x. When dividing the data into ten sets we thus got

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 10 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix},$$

where y_i denotes the energy in the i^{th} set. Solving for $\hat{\beta}$ then led to $\hat{y}_i$, the estimate of y_i. But in order to be able to implement this linear regression model, it needed to be checked that the residuals $\epsilon_i = \hat{y}_i - y_i$ were Normal distributed with mean 0 and constant variance σ^2. To check this, the residuals were plotted with the help of a QQ-plot, as shown in Figure 7. Plotting the residuals did not imply that the residuals of the whole data set were $N(0, \sigma^2)$ for some σ^2. They deviated greatly from the line showing the Normal distribution for this specific mean and variance. It became evident that there was a big variance within the data. This variance could possibly be the result of the data coming from different people where every person might have a different resting state HRV. Since the assumption of the residuals being $N(0, \sigma^2)$ was not true, a linear regression model could not be implemented. Hence, a linear model based on a line of best fit for each person was implemented instead.

All approaches described in the section ”Binary classification” were based on the

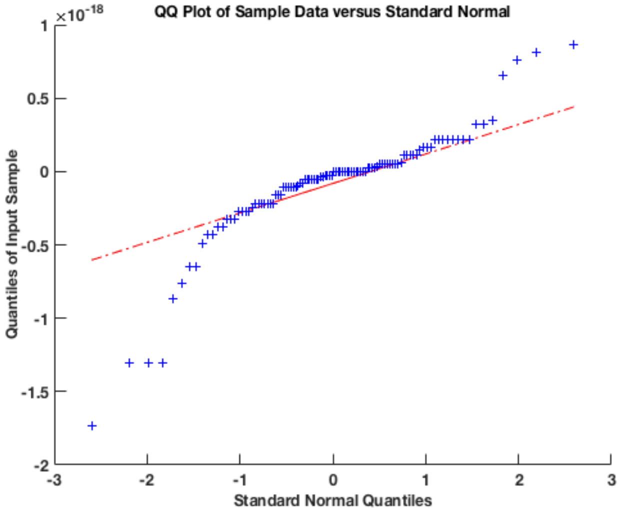

Figure 7: QQ-plot of the residuals for the linear regression model over the HF range of the cold HRV signal when splitting the data into two sets.

assumption that the warm and the cold signal have different energy distributions and that the energy changes over time. This is shown in Figure 8. Here the energy of the cold signal is higher in the second set, while the energy of the warm signal is lower in the second set. Energy was calculated as the area under the graph of the spectral estimate,

$$e_i = \sum_{j=1}^{n} \hat{R}^i_{av}(f_j),$$

where f_j is the frequency vector containing n entries and $\hat{R}^i_{av}$ is the spectral estimate of set i using the Welch method with a 50% overlap. Since the data was sampled discretely over time, the energy was obtained by summation instead of the calculation of an integral.

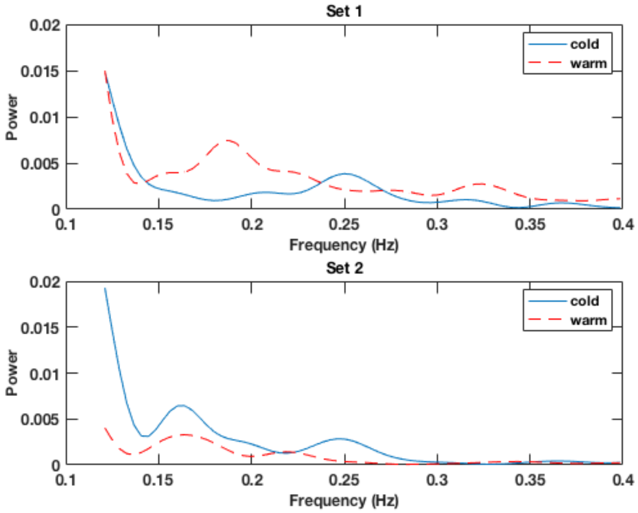

Figure 8: Difference in energy distribution of subject 1 over the HF-range. Here, set 1 is made up of the first 90 s of the signals while set 2 is made up of the last 90 s.

3.1 HF-band analysis

In this section, the different frequency bands, that were used, are explained. The general HF-band, which for adults is defined as 0.12 - 0.4 Hz, was the widest band that was analyzed with a width of 0.28 Hz. The individual frequency bands were more narrow and had at most width 0.2 Hz.

3.1.1 General HF-band

After an initial assessment of the data, the conclusion was reached that the biggest differences between the warm and the cold signals could be seen in the HF-range. This has also been observed in previous studies where it has been noted that the exposure to cold stimuli can effect the HRV signal in HF range [1]. For further analysis of this frequency band, the HRV data was then split up in time in up to ten sets.

When analyzing this frequency band, it was at first tried to split the data into a training set and a validation set. The training set consisted of the data of 36 to 46 randomly selected subjects. The data of the remaining individuals made up the validation set. The algorithm was then trained on the training set. Here, training the algorithm meant that it would find the number of windows that led to the highest percentage of correct classification. This number of windows was then assumed to be optimal. Using this optimal number of windows then for the validation set as well resulted in a MCC-value and percentage of correct classification for the validation set. This was tested a total of 25 times. However, different sizes of training sets resulted in a different number of optimal windows. Furthermore, different training and validation sets of the same size with the same number of optimal windows did not achieve the same classification results and therefore did not lead to the same MCC-value. Due to the results on average being better than random but having a quite big variation, it was decided to instead treat the data of all 53 persons as the training set.

3.1.2 Obtaining an individual frequency band by finding the respiratory maximum

As opposed to the general HF-band, not only HRV data was used for this approach but also respiratory data. All data was normalized and then divided into two, four or ten sets. But for this approach we did not look at the whole HF-range since this can show more than just the most important information. Instead we obtained an individual frequency band for each person. Even for the same person, this frequency band did not need to be identical for different sets. The frequency band,

$$f_{\max} \pm \delta_f,$$

was centered at $f_{\max}$, the frequency where the maximum of the respiratory spectral estimate was located. This approach was tested for both $\delta_f = 0.05$ Hz and $\delta_f = 0.1$ Hz where the only restriction was that the frequency band should not start below 0.05 Hz. This translates to the maximum lying above 0.1 Hz respectively 0.15 Hz. This was done since we ideally only wanted to analyze the HF-range and everything below 0.12 Hz is part of the LF-range. However, since the wider frequency band did not consistently lead to better results, this was not tested for different $f_{\max}$. The more narrow frequency band, however, was also tested for $f_{\max} \geq 0.15$ as well as $f_{\max} \geq 0.2$. Setting a higher threshold for $f_{\max}$ was done to see how much of an impact the noise at the lower frequencies actually might have. Once this frequency band was found for each subject, the energy of the HRV spectrum could be computed. This was advantageous since this frequency band could be seen as each subjects personal high frequency range meaning that only the relevant parts of the HRV signal would be taken into account.

3.2 Binary classification

Different approaches were tried to find out the most successful method of classification. When deciding which method to implement, not only the percentage of correct classification was relevant but also what kind of data is available. Some methods used only the HRV data, while others also made use of the respiratory data.

3.2.1 Energy distribution

For this method, the HRV signal was split into two sets each 90 s long. The goal of this method was to correctly classify whether the data came from a cold or a warm HRV-signal. This classification was based on the assumption of there being a difference in energy over time between the warm and the cold signal. More specifically, we compared the energy of the first 90 s to the energy of the last 90 s. However, only the data of the cold signal was used to obtain the optimal number of windows, since we only classified based on the energy distribution of the cold signal. Hence, the energy of each set was calculated. It was assumed that the energy of the cold signal was not evenly distributed but increased over time. Given that the cold signal had less energy in the beginning than in the end, the algorithm then chose the optimal amount of windows from a range which depended on what the lowest frequency in each frequency band was. Here, the optimal amount of windows was defined as the one that leads to the maximum number of energy increases over time for the cold signal, meaning that it was counted how many times the energy in the second set was higher than in the first. The window that led to this number being the highest was then said to be optimal. If certain number of windows led to the same maximum, the one that on average had the highest difference between the beginning and the end of the signals, was said to be optimal. Thus, based on the optimal number of windows, the energy of both sets was calculated for the warm and the cold signal. If the energy in the second set was larger than the energy in the first set, the HRV signal would then be classified as "cold". If, on the other hand, the energy in the second set was not higher, the signal would then be classified as "warm".

3.2.2 Linear model

Based on the assumption of different energy distributions over time, a linear model was implemented. However, this time the algorithm trained both on cold and warm HRV data to obtain the optimal number of windows. We not only wanted the slopes of the signals to be different, but since energy over time increased for the cold signal we wanted the slope of the linear model to be positive. Since the warm signal behaved in the opposite way, i.e. decreased over time, we wanted the slope to be negative. The training algorithm then chose the optimal number of windows based on this linear model. However, this time the optimal number of windows was the one that maximizes the number of correct classifications for both the cold and the warm HRV data. Meaning that the optimal window leads to both the highest number of non-negative slopes for the cold HRV data, $\hat{\beta}_{\text{cold}} \geq 0$, while still having negative slope for the warm HRV data, $\hat{\beta}_{\text{warm}} < 0$. Thus, a signal's classification was only dependent on the slope of the line of best fit, $\hat{\beta}$. A signal would be classified as cold if $\hat{\beta} \geq 0$ and as warm if $\hat{\beta} < 0$. Different to the previous method, the algorithm was not only implemented with two sets but also with four and ten sets. The window range was adjusted to the set size as well as to the lower bound of each frequency band. The more sets we divided the data into, the smaller the window range became.

3.2.3 Comparison of absolute energy

In this method, classification was based on the total amount of energy, rather than the distribution of the energy. This means that we did not analyze how the amount of energy changed over time. Instead we just compared the total energy of warm to the total energy of cold over the given length of the data set and the given frequency range. It should be noted that due to this comparison, this method does not treat the problem of binary classification but rather the problem of pairwise classification. The difference is that with binary classification one signal is classified as either ”positive” or ”negative”, while with pairwise classification two signals are compared to each other and then one is classified as ”positive” and the other one as ”negative”. However, we could thus not classify one signal on its own with this approach. The idea was that the total energy of the warm signal was higher than the total energy of the cold signal. This pattern however was more distinct when looking at the first 90 s. As shown in Figure 9, the total energy of cold over the whole signal was higher than the total energy of warm. But when only looking at the first 90 s, warm had the most energy. This observation lead to an additional step for this method, since the biggest difference seemed to be seen in the first 90s of the signal. Hence, the absolute energy of cold and warm over the first 90 s was analyzed as well as over all 180 s.

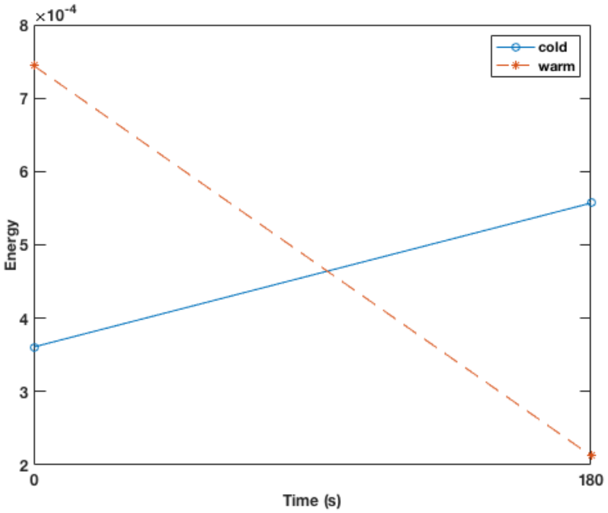

Figure 9: Energy distribution of subject 70.

4 Results

In this section, the results of the above mentioned methods are presented. The tables all contain the confusion matrix of the training set, the MCC-value, the percentage of correct classification as well as the optimal number of windows when using the Welch method with a 50% overlap. Here, P denotes positive and cold is seen as the positive classification. N on the other hand stands for negative and the negative classification is in this case warm. For further visualization of the results, plots of the confidence interval for the linear models are also presented.

4.1 Analysis of the general HF range

In this section, the results of analyzing the general HF-band are presented. Due to the general HF-band being quite wide, the results are worse than when analyzing each individual's HF range. In Table 2, the maximum number of windows for each method is shown.

Table 2: Maximum number of windows for the HF range.

# Sets	Method	# Windows
2	Energy distribution	10
2	Linear model	10
2	Energy comparison of the last 180 s	10
2	Energy comparison of the first 90 s	5
4	Linear model	5
10	Linear model	2

4.1.1 Energy distribution over HF-HRV

When training only on cold HF-HRV data, while using the Welch method with a 50% overlap, this leads to the results presented in Table 3.

Table 3: Energy comparison over HF-HRV.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	39	19	0.1019	54.72	6
	N	14	34			

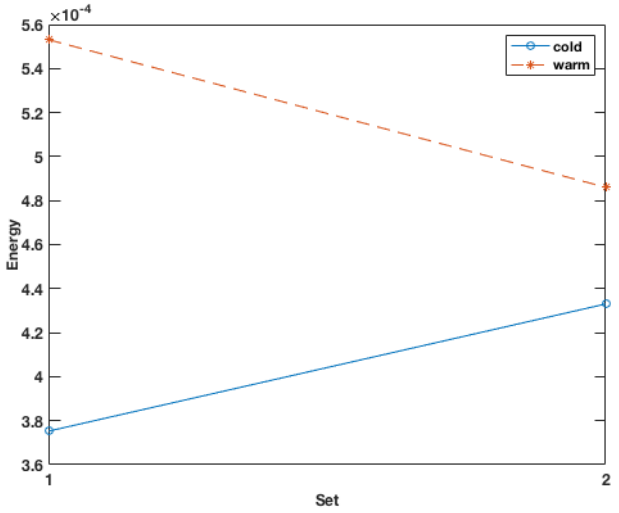

Figure 10: Mean energy over HF-range.

The confusion matrix in Table 3 shows that there are quite many false positive classifications. However, when looking at the mean energy of all subjects for both the cold and the warm signal, shown in Figure 10, it becomes clear why this approach was chosen.

4.1.2 Linear model over HF-HRV

A linear model over the same frequency band leads to better classification results, shown in Table 4, than the ones obtained by analyzing the energy distribution presented in Table 3. The percentage of classification increases by at least 9 percentage points compared to the previous results.

Table 4: Linear model over HF-HRV.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	37	17	0.3774	68.87	2
	N	16	36			
4	P	40	19	0.3988	69.81	2
	N	13	34			
10	P	41	26	0.2934	64.15	2
	N	12	27			

When dividing the data into four sets the highest correct classification is obtained by this approach. A subject which was classified correctly with this

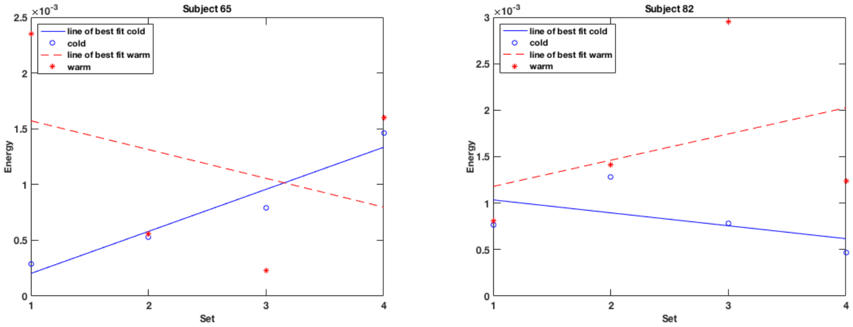

Figure 11: Linear model over HF-range of subject 65 and subject 82.

method is for example subject 65. The lines of best fit for both the cold and the warm signal of subject 65 as well as the actual energy are shown on the left side in Figure 11. While the splitting into 45 s intervals, as opposed to 90 s or 18 s, over the HF range still achieves the best results and also shows the most improvement to the previous approach, false classification can still occur. A good example is shown on the right side in Figure 11 which shows the lines of best fit of subject 82. Here $\hat{\beta}_{\text{cold}} < 0$ and $\hat{\beta}_{\text{warm}} > 0$ which is the opposite of what we want and therefore does not lead to correct classification.

Since we use a linear model here, we can not only compute the percentage of correct classification and the MCC-value. We can even look at the measure that we classify the data with, namely the slope of the cold HRV-signal, $\hat{\beta}_{\text{cold}}$, and of the warm HRV-signal, $\hat{\beta}_{\text{warm}}$. In Figure 12, the 95% confidence intervals of the linear model for the HF-HRV signals are shown. For each person we have a line of best fit,

$$y = \beta_0 + \beta_1 x,$$

for both the cold and the warm signal. Here, β_0 denotes the y-intercept and β_1 the slope of the line of best fit. Thus the 95% confidence interval for the slope β_1 is

$$I_{\hat{\beta}_1} = (\hat{\beta}_1 - \frac{1}{\sqrt{53}}\lambda_{0.025}, \hat{\beta}_1 + \frac{1}{\sqrt{53}}\lambda_{0.025}) = (\hat{\beta}_1^{\text{lower}}, \hat{\beta}_1^{\text{upper}}),$$

where $\lambda_{0.025}$ denotes the upper 0.025 quantile [6]. The thick lines show the average slope,

$$y^{\text{average}} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

The dotted lines show the lower bound of the confidence interval,

$$y^{\text{lower}} = \hat{\beta}_0 + \hat{\beta}_1^{\text{lower}} x$$

respectively the upper bound,

$$y^{\text{upper}} = \hat{\beta}_0 + \hat{\beta}_1^{\text{upper}} x.$$

As previously described, x specifies how many sets we divide the data into. Thus, the ticks on the x-axis indicate the number of sets that were used to calculate the intervals. While the three different means of $\hat{\beta}_{\text{cold}}$ as well as the corresponding upper bounds are non-negative as wished, this is not the case for the three corresponding lower bounds. For the slope of the warm HRV signal, $\hat{\beta}_{\text{warm}}$, we have a similar situation. The lower bounds and the mean are negative as needed. But only the upper bound when dividing into two sets is negative. The other two upper bounds are positive and therefore have the wrong sign.

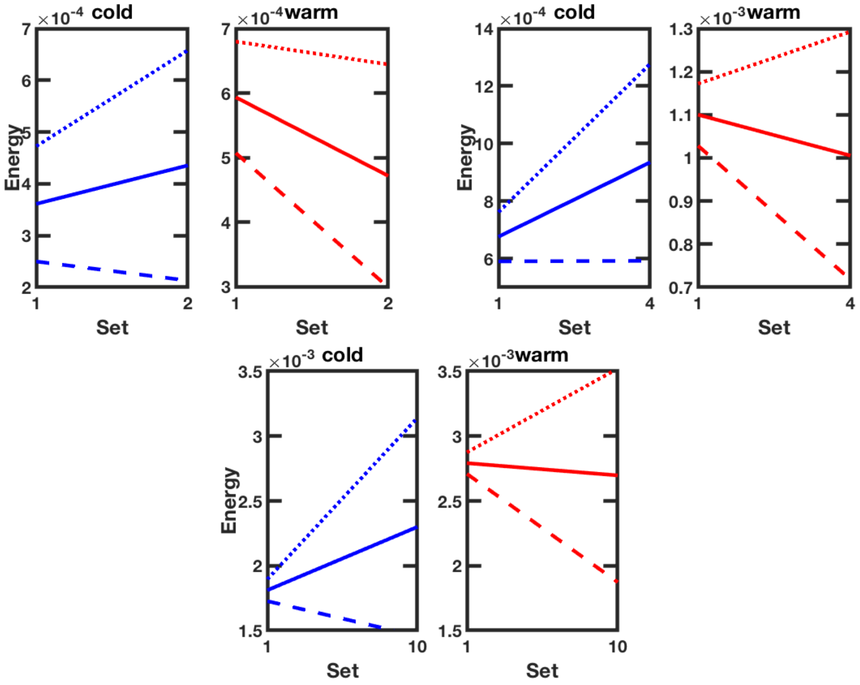

Figure 12: 95% confidence interval over HF-HRV for the division into two, four and ten sets.

4.1.3 Energy comparison over the HF-range

The best results over the HF range are obtained by classifying based on the total energy of either the whole signal or the first 90 s. Those results are shown in Table 5. Both the MCC-values as well as the classification percentage is by far higher than for the other two methods. However, since we now deal with pairwise binary classification, this is an easier problem.

Table 5: Energy comparison over the HF-range.

Length of data set	Predicted	P	N	MCC	Classification [%]	Windows
180 s	P	42	11	0.5849	79.25	10
	N	11	42			
90 s	P	44	9	0.6604	83.02	3
	N	9	44			

4.2 Analysis of the maximum respiratory frequency band

In this section, the results of all three methods over all individual frequency bands are presented. Thus each subsection contains the results of four different frequency bands. In Table 6, the maximum number of windows for each frequency band is presented. Due to the different frequency bands having different lower bounds, this table has one more column then Table 2.

Table 6: Maximum number of windows for the HF range.

Lower bound [Hz]	# Sets	Method	# Windows
0.05	2	Energy distribution	4
0.05	2	Linear model	4
0.05	2	Energy comparison of the last 180 s	4
0.05	2	Energy comparison of the first 90 s	1
0.05	4	Linear model	1
0.05	10	Linear model	1
0.1	2	Energy distribution	8
0.1	2	Linear model	8
0.1	2	Energy comparison of the last 180 s	8
0.1	2	Energy comparison of the first 90 s	4
0.1	4	Linear model	4
0.1	10	Linear model	1
0.15	2	Energy distribution	13
0.15	2	Linear model	13
0.15	2	Energy comparison of the last 180 s	13
0.15	2	Energy comparison of the first 90 s	6
0.15	4	Linear model	6
0.15	10	Linear model	2

4.2.1 Energy distribution

The results for the analysis of energy distribution over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.1$ are presented in Table 7. Compared to the results over the HF range shown in Table 3, the percentage of correct classification shows an improvement of almost 20 percentage points.

Table 7: Energy distribution over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.1$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	34	7	0.523	75.47	4
	N	19	46			

Setting the restriction that $f_{\max} \geq 0.15$ leads to more true positive classifications, which is shown in Table 8. While the number of true negatives decreases, this does not negatively affect the MCC-value or the classification percentage, since this decrease is smaller than the increase in true positive predictions.

Table 8: Energy distribution over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.15$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	36	8	0.5361	76.42	8
	N	17	45			

Table 9 shows the wider frequency band centered around $f_{\max} \geq 0.15$. While the negative classifications remain unchanged, more correct positive classifications lead even here to better overall results.

Table 9: Energy distribution over $f_{\max} \pm 0.1$ for $f_{\max} \geq 0.15$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	40	8	0.6065	80.19	4
	N	13	45			

For this method, the best results are obtained for a narrow frequency band centered around $f_{\max} \geq 0.2$, as shown in Table 10. Less than 15% get falsely classified, which is an increase of over 30 percentage points compared to the energy distribution method over the HF range shown in Table 3.

Table 10: Energy distribution over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.2$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	42	4	0.7233	85.85	8
	N	11	49			

4.2.2 Linear model over the maximum respiratory frequency band

As with the previous linear model, the more sets we divided our data into, the smaller the range of possible windows becomes. This smaller range of windows does however not necessarily lead to worse results. Comparing the results of the linear model over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.1$, each person's own HF range, to the results from the linear model over 0.12 - 0.4 Hz, we see an improvement no matter how many sets we divide our data into. The results for this frequency band are shown in Table 11.

Table 11: Linear model over $f_{\max} \pm 0.05$, $f_{\max} \geq 0.1$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	34	7	0.523	75.47	4
	N	19	46			
4	P	31	9	0.4282	70.75	1
	N	22	44			
10	P	26	3	0.4867	71.7	1
	N	27	50			

Once again, we can analyze our measure of classification. Compared to the confidence intervals obtained from the linear model over HF-HRV, shown in Figure 12, the ones in Figure 13 are much better. As previously, the mean is shown as one straight line while the lower and upper bounds are dotted lines. We still have that the lower bound, when dividing into four and ten sets, has a negative slope for the cold signal. However, both the lower and the upper bounds for the warm signal have a negative slope, which is exactly the result that we wished for.

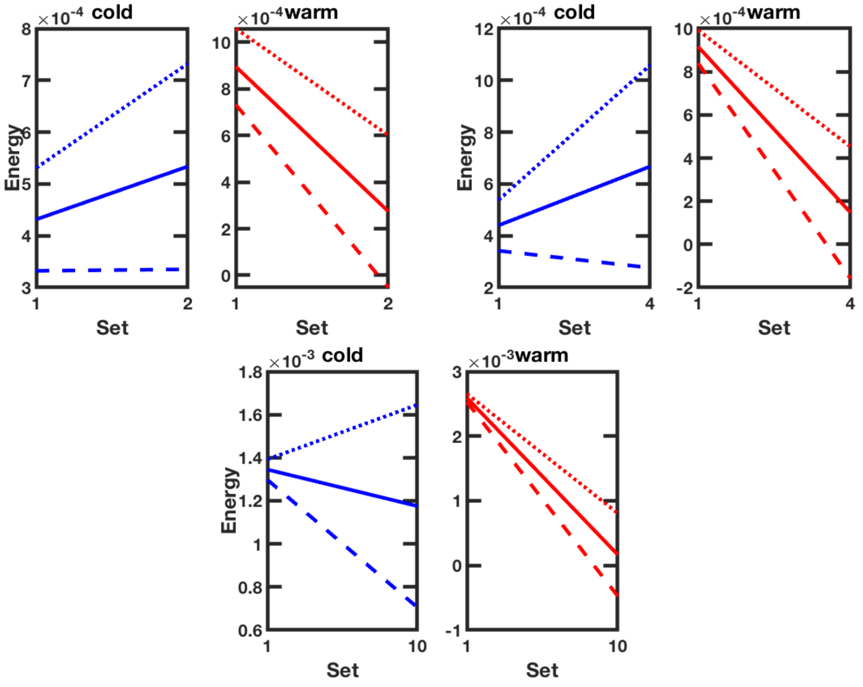

Figure 13: 95% confidence interval over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.1$.

For the division into two and four sets, changing the possible frequency band to be centered around $f_{\max} \geq 0.15$ Hz does lead to better results, as shown in Table 12, than when having a lower threshold such as $f_{\max} \geq 0.1$ Hz. This improvement is however not visible for the division into ten sets.

Table 12: Linear model over $f_{\max} \pm 0.05, f_{\max} \geq 0.15$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	36	8	0.5361	76.42	8
	N	17	45			
4	P	40	5	0.668	83.02	2
	N	13	48			
10	P	49	28	0.4444	69.81	1
	N	4	25			

Increasing the threshold for $f_{\max}$ from 0.1 Hz to 0.15 Hz leads to improved confidence intervals for the first two divisions, as shown in Figure 14. As for $f_{\max} \geq 0.1$, the means as well as the 95% confidence interval bounds for the warm signal have the slope that we want, meaning $\hat{\beta}_{\text{warm}} < 0$. For the cold signal, the slopes of 95% confidence interval when dividing into two and four sets are positive as needed. However, when dividing into ten sets, both the mean and the lower bound of the cold signal are negative, meaning that they have the wrong sign.

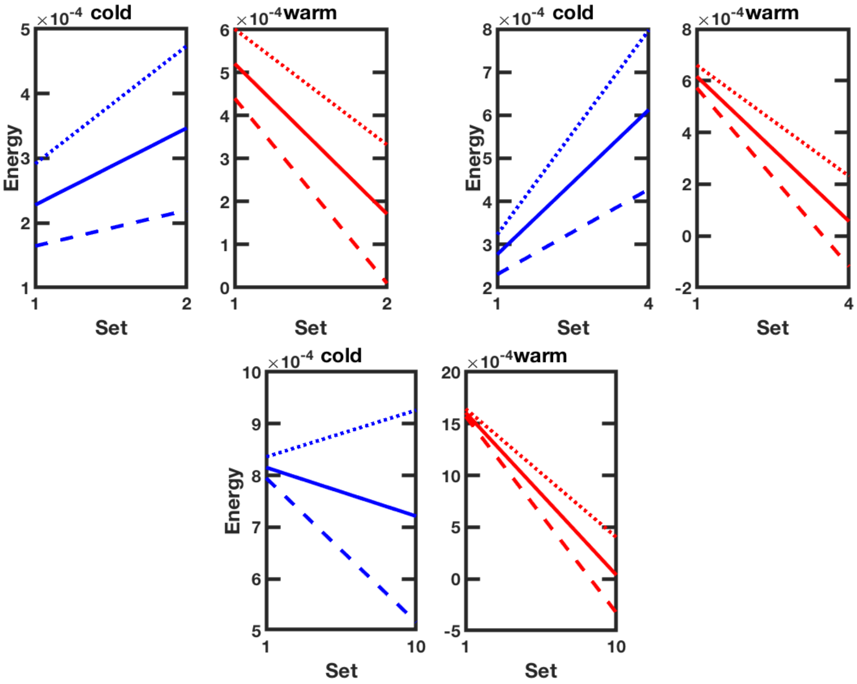

Figure 14: 95% confidence interval over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.15$.

Compared to the more narrow frequency band centered around the same $f_{\max}$, this wider frequency band leads to better classification and MCC-results when dividing into two or ten set, as shown in Table 11.

Table 13: Linear model over $f_{\max} \pm 0.1$, $f_{\max} \geq 0.15$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	40	8	0.6065	80.19	4
	N	13	45			
4	P	34	10	0.4595	72.64	1
	N	19	43			
10	P	28	2	0.5445	74.53	1
	N	25	51			

Figure 15 shows the 95% confidence intervals corresponding to the linear model over $f_{\max} \pm 0.1$ for $f_{\max} \geq 0.15$. Here, only the lower bound of the cold signal when dividing into ten sets has the wrong slope. Otherwise, the remaining lower and upper bounds as well as all means are correct.

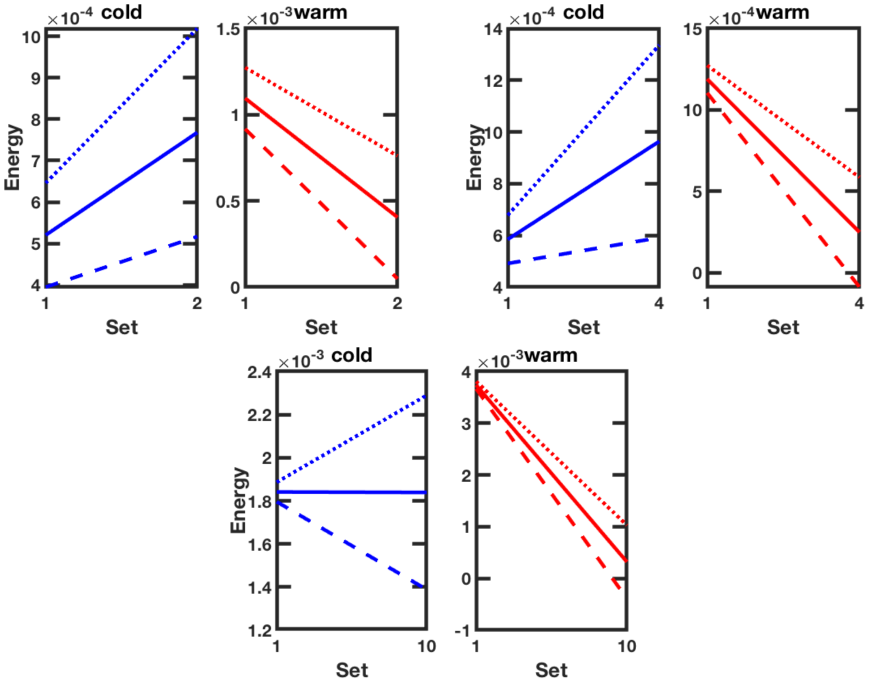

Figure 15: 95% confidence interval over $f_{\max} \pm 0.1$ for $f_{\max} \geq 0.15$.

The best results, when classifying the data based on a linear model, are obtained for $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.2$. For the division into four sets, less than 7% get falsely classified as seen in Table 14.

Table 14: Linear model over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.2$.

# Sets	Predicted	P	N	MCC	Classification [%]	# Windows
2	P	41	2	0.7493	86.79	13
	N	12	51			
4	P	46	0	0.8756	93.4	4
	N	7	53			
10	P	32	0	0.6576	80.19	2
	N	21	53			

Analyzing the confidence interval for this method, shows great results even here. However, even for this threshold not all bounds are correct. Once again, the lower bound of the cold signal is negative when dividing the data into ten sets. Those results are shown in Figure 16.

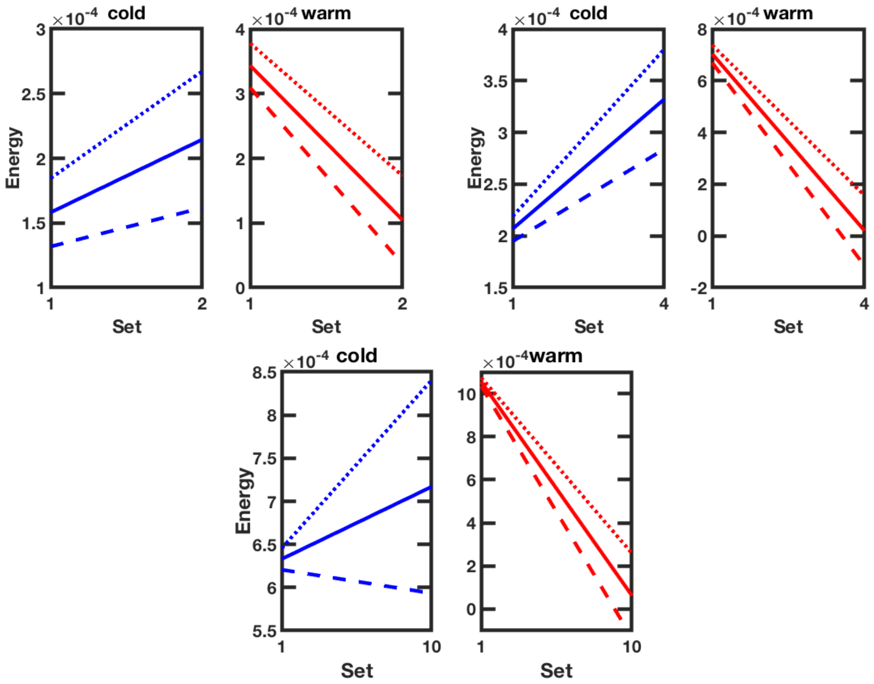

Figure 16: 95% confidence interval over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.2$.

4.2.3 Energy comparison of the maximum respiratory frequency band

The comparison of total energy over $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.1$ leads to the results presented in Table 15. For this frequency band, quite a big difference can be seen between taking into account all 180 s of the data set or only the first 90 s.

Table 15: Energy comparison over $f_{\max} \pm 0.05, f_{\max} \geq 0.1$.

Length of data set	Predicted	P	N	MCC	Classification [%]	Windows
180 s	P	39	14	0.4717	73.58	1
	N	14	39			
90 s	P	43	10	0.6226	81.13	1
	N	10	43			

The confusion matrix that is obtained for $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.15$ when taking into account the first 90 s of data, shown in Table 16, is the same as for $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.1$ which was shown in Table 15. However, this frequency band leads to a different optimal number of windows for both the 180 s and the 90 s. When looking at the results of the first two frequency bands presented in this section, $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.1$ and $f_{\max} \pm 0.05$ for $f_{\max} \geq 0.15$, it already becomes apparent that the first 90 s of data lead to better classification results.

Table 16: Energy comparison over $f_{\max} \pm 0.05, f_{\max} \geq 0.15$.

Length of data set	Predicted	P	N	MCC	Classification [%]	Window
180 s	P	38	15	0.434	71.7	8
	N	15	38			
90 s	P	43	10	0.6226	81.13	4
	N	10	43			

The analysis of a wider frequency band, $f_{\max} \pm 0.1$ for $f_{\max} \geq 0.15$ compared to the more narrow one centered around the same $f_{\max}$, leads to the same results when looking at all 180 s. However, when only looking at the first 90 s it leads to worse results. Those results are shown in Table 17. This is also the reason why $f_{\max} \pm 0.1$ is not carried out for an increased $f_{\max}$ threshold. The first 90 s generally lead to better classification results than when taking into account all 180 s. However, for this wider frequency band, the percentage of correct classification decreases now to below 80% when looking at the first 90 s.

Table 17: Energy comparison over $f_{\max} \pm 0.1, f_{\max} \geq 0.15$.

Length of data set	Predicted	P	N	MCC	Classification [%]	Window
180 s	P	38	15	0.434	71.7	2
	N	15	38			
90 s	P	40	13	0.5084	75.47	1
	N	13	40			

Not considering $f_{\max}$ below 0.2 Hz leads to the best classification results for this model. The results for both the whole data set as well as those for the first half are shown in Table 18.

Table 18: Energy comparison over $f_{\max} \pm 0.05, f_{\max} \geq 0.2$.

Length of data set	Predicted	P	N	MCC	Classification [%]	Window
180 s	P	43	10	0.6226	81.13	11
	N	10	43			
90 s	P	51	2	0.9245	96.23	1
	N	2	51			

5 Discussion

While some initial data analysis was done using both the periodogram and the Hanning window, all the results that are presented here were obtained using the Welch method using the Hanning window with 50% overlap. The Welch method was chosen, since we wanted as little variance as possible due to there being already great variance between the individual subjects. However, when having quite few data points using the Welch method with many windows needs to be done with caution since the spectral estimates will become very smooth. If the spectral estimate is too smooth, important information can get lost which in return can lead to false conclusions. After all, the spectral estimate is the basis of all the binary classification methods in this thesis. In order to avoid the loss of important information, the range of the possible number of windows was adjusted according to the length of each set as well as the lowest frequency that needed to be displayed. However, when dividing the data into ten sets that usually meant that only one window was used. This could be the explanation to why the division into ten sets does not always lead to the best results. Since there were no windows to choose from one window became by default the optimal number.

Because only the data sets of 53 subjects were available to implement different methods on, it was decided to use all data as training data. Previous trials with randomly assigning subjects to either the training or the validation set did lead to better than random classification of the validation set. While this is positive, there were some problems with the random assignment to these two sets, since the classification results seemed to highly depend on what subjects the validation set contained. For instance, if the training set contained many subjects whose HRV signals follow the general trend while the validation set contained mostly those outliers that do not follow the trend, this could lead to problems. When testing several methods on the same training and validation set, some discrepancies could be made between which method was the best. However, when running each method on its own on several validation sets, the results within one method varied greatly. But the results still varied between slightly better than random classification and almost perfect classification. The main reason for this great variance seems to be the great variation within the data sets of different individuals.

When analyzing the mean of the data, a general trend could easily be found. However, those trends did not always hold for all the subjects, since there often were quite many outliers. On the other hand, this means that it is not necessarily surprising that many methods lead to false predictions. It is hard to make a model that fits all the data, when there is relatively little data and such a great variation between individuals. This also became a problem when trying to implement a linear regression model for all subjects. The assumption that the residuals follow a Normal distribution does not hold for this data set which means that an individual linear model with a line of best fit needed to be implemented instead. That being said, even though a great variation between the individuals' data can be observed, all the tested methods lead to a binary classification of the training set that is better than random with correct classification up to 96.23%. It remains to be seen how well these classification methods work when tested on unknown data, since not having a validation set means that the methods actually could perform quite differently with a different set of data.

An important observation is the improvement of the maximum respiratory frequency band analysis once the threshold was set to $f_{\max} \geq 0.2$. With this threshold, the best results are obtained both comparing the total energy as well as for the linear model. This improvement might be due to noise in the lower frequencies of each signal which makes correct classification harder. With noise we mean in this case elements from the LF range, since it was decided to only analyze the HF range. This applies both for the general HF range or the specific one for each individual. This noise could also be an explanation why the linear model over the whole HF-HRV range does not perform as well.

6 Conclusion

In conclusion, it can be said that the binary classification of warm and cold HRV signals is possible. Here, the warm signal is the HRV at resting state while the cold signal is obtained during a stress simulation, a so-called Cold Pressure Test. If no respiratory data is available, 83.02% correct classification can be obtained by comparing the energy over the general HF range of the first 90 s of the signal. When also having respiratory data available, more than 80% correct classification is easily obtained, since then an analysis over each subject's personal HF range can be carried out. It is thus recommended to monitor breathing during an ECG so that respiratory data is available for further analysis. Correctly classifying more than 80% can be achieved by all three methods described in the report. However, classifying more than 90% correct occurs only when using a narrow frequency band of 0.05 Hz and setting the threshold to $f_{\max} \geq 0.2$. Dividing the data into four time sets in that frequency band leads to the linear model classifying 93.4% correctly. However, the most successful method, that was obtained during this project, was comparing the total energy over this frequency range. Here we get 96.23% correct classification when taking the first 90 s of the signal into account.

References

[1] M. Malik et al. "Heart rate variability". In: *European Heart Journal* 17 (1996), pp. 354–381.

[2] U. Rajendra Acharya et al. "Heart rate variability: a review". In: *Medical & Biological Engineering & Computing* 44 (2016-11-17), pp. 1031–1051.

[3] G. E. Billman. "Heart rate variability - a historical perspective". In: *Frontiers in Physiology* 2.86 (2011-11-29).

[4] G. Lindgren, H. Rootzén, and M. Sandsten. *Stationary stochastic processes for scientists and engineers*. CRC Press, 2013.

[5] D. Chicco. "Ten quick tips for machine learning in computational biology". In: *BioData Mining* 10 (2017).

[6] D. Anevski. *A concise introduction to mathematical statistics*. Studentlitteratur AB, Lund, 2017.