School of Electronic Engineering and Computer Science

MSc in Big Data Science
Project Report 2017

Investigating heart rate variability: a machine learning approach

Christopher André Ottesen

August 2017

ABSTRACT

Advances in wearable technology and photoplethysmography make non-invasively measures of bio signals more accessible and accurate, opening up for a future where health care is better delivered at a lower cost. Heart rate variability, a measure of the difference between consecutive heart beats is a measure directly linked to the nerve system. Research successfully connects heart rate variability to measures for stress, cancer monitoring, diabetes, mellitus, sleep problems and difficulties regulating emotions. The research undertaken in this project uses automated machine learning techniques through TPOT, Auto-Sklearn and Grid Search to create a model that successfully detects mental stress and differentiates it from physical stress through the use of accelerometer data in wearable sensors. The Random Forest algorithm tuned by Auto-Sklearn achieved an F1 score of 0.8346 in detecting stress from heart rate variability features in the "Drive" database from Physionet. Further, the algorithms were tested on data collected with a Fitbit Charge 2 and a Polar H7 chest strap, where both Auto-Sklearn and TPOT created models that could detect the stressed states of the subject down to the minute. These findings further prove that stress is measurable with heart rate variability through non-intrusive wearable devices and machine learning.

ACKNOWLEDGEMENTS

I want to thank my supervisor Dr Tony Stockman for great support both morally and academically throughout the project. Further, Hesham Dar and Petur Einarsson have been excellent sparring partners and of great help throughout the project and Master's degree in general. Last but not least, I want to thank my family for great moral support and in particular my uncle M.D. Hans Rydningen for giving me feedback from a professional medical perspective.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

With the rising popularity of wearables and devices that constantly monitor several key measures of the body, the access to physiological data from individuals are getting more and more accessible and opening up for a future with healthcare delivery of higher quality at a lower cost. This project examines how machine learning techniques can be used to detect and measure stress based on data from wearable devices.

Stress is one of the most occurring health problems in Europe, and according to studies, one of four people that were absent from their work for more than a month was away due to stress triggered issues. (Jacqueline Wijsman, 2011) Stress is a major both economic and social problem in the society, and an automatic system that can detect stress in an early stage has significant proactive benefits for both the society and individuals.

Broad view of the general research area

Advances in technology are gradually allowing consumers to take control of their health with devices that collect health data on an individual level. This data enables people to take a more active role in their health as well as enhancing the decision powers of professional medical personnel. (NHS England, 2015) This data will in the close future be an important tool for giving patients the correct treatment, utilising internet of things devices can to both automatically or in the supervision of a professional identify persons need of proactive treatment or changes in their lifestyle, detect acute illness or help in monitoring of chronic diseases. There is huge potential in this technology and individuals, private organisations, educational organisations and governments all see the potential that lays in this technology. For example, Queen Mary University of London is one of multiple research institution to receive funding from EPSRC to investigate the use of wearable devices and Bayesian networks in long-term medical conditions. (Fenton, 2017)

Additionally, private organisations are competing to get a market share in the personal health sphere. Companies such as Google (Diamond, 2015), Samsung (Lorenzetti, 2016) , Fitbit and more are investing heavily in personal health with both software and hardware devices. (Stables, 2017). Where for example Apple has partnered up with many big firms and research institutions to open up their system to be used to empower their devices for use as medical, fitness and well-being devices. They have made it easy to make health based apps through their purpose-built APIs, and both Apples own devices and third-party hardware (Hodson, 2015).

THE CONTRIBUTION FROM THIS PROJECT

The research conducted in this project explores a machine learning approach on bio signals from wearable devices such as fitness trackers as a tool for proactive treatment for detection of short term

mental stress. This study focuses on data that is available through non-invasively wearable devices and separates itself from similar studies where the data foundation is based on more complex data such as clinical level ECG.

Chapter 2: Review of Literature

2.1 Similar research

The most directly similar research conducted is a final project done in the machine learning course at Stanford University by David Liu and Mark Ulrich. Where they tried to classify stressed state on the drive database from Physionet and followed the heart rate variability and ECG workflow given by Physionet to extract ECG and heart rate variability features from the dataset. With a linear SVM, they were able to get an F1 score of 0.7855, but they do not clearly enough state how data and parameter tuning is handled to recreate their results. (Liu & Ulrich, 2014)

Cardiogram is a mobile application paired with the research study mRythm where a team of medical researcher from the University of California San Francisco is applying Deep Learning on data collected on the Apple Watch to identify abnormal heart rhythms - atrial fibrillation (Cardiogram, 2017). Their aim is to detect various heart problems currently focusing on stroke where early results detect abnormal heart rhythms with a sensitivity of 98.04% (Singh, 2017).

2.2 Physiological measures of stress

Stress is an emotion the body goes through as a response to the particular situation, by releasing stress hormones. The response includes adrenaline and cortisol which sharpens the alertness and strength of the body. (Villarejo, et al., 2012) Stress affects several physical processes in the autonomic nervous

system (ANS) leading to increased muscle tension, change in concentration and changes in the heart rate and the heart rate variability. (Taelman, et al., 2008)

In general, stress can be split into three different categories, acute stress, episodic acute stress and chronic stress. (Bakker, et al., 2011) The acute stress factor is characterised by a short-term arousal where the body returns to its normal state after the stress factor has passed (Bakker, et al., 2011) and it is this type of stress that is examined in the scope of this project.

2.2.1 The heart

The fundamental purpose of the heart is to rhythmically contract and pump blood into the lungs for oxygenation (Goldberger, et al., 2013) and to pump oxygenated blood into the organs and deliver nutrients to different cells. (Bourghelle, 2015)

2.2.2 The hearts ECG patterns

Cardiac contraction is measured by measuring how electrical current is spread through the heart muscle. The heart muscle contracts by flowing electrical current into the muscle, produced by *pacemaker cells,* and specialised conduction tissue that repolarize and depolarize the heart muscle by transporting electric through the muscle (Goldberger, et al., 2013). This electrical activity is commonly measured with electrocardiogram – or ECG device and provides a time-voltage chart of heartbeats (Goldberger, et al., 2013).

2.2.2.1 The ECG WAVEFORM

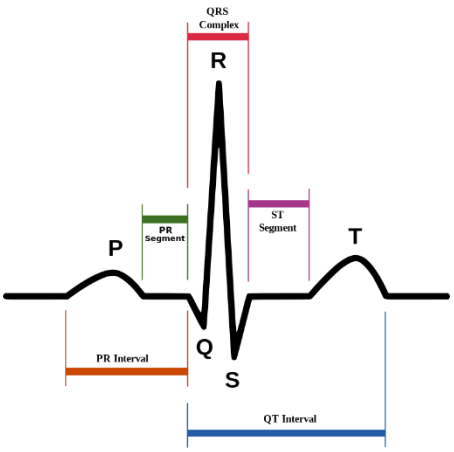


Figure 1 by (Atkielski, 2007) shows a labelled representation of the different building blocks of the ECG signal by. The figure shows labels for the P wave, QRS complex, ST segment and T wave.

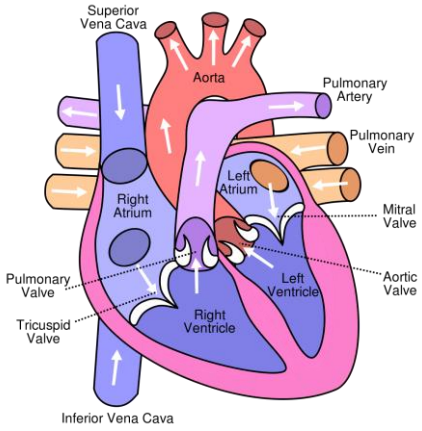


Figure 2 by (Wapcaplet, 2012) shows a diagram of the human heart where the ventricles and the atrium are labelled.

The ECG signal consists of different phases that represent how electricity travels through and stimulates the atria and ventricles. The ECG signal is built up in five phases which are *P wave, QRS, ST, T wave* and *U wave.* The two first phases are an activation phases where the muscles are depolarized; the three final stages are recovery stages where the muscles are repolarised and represent the muscle in resting stage. (Goldberger, et al., 2013). On the figure above the QRS complex is representing depolarization, and the ST segment, T wave, and U wave are representing repolarization of the ventricles.

2.2.2.2 RR INTERVAL AND HEART RATE VARIABILITY

RR interval is the interval between two QRS elevations, and it is the measure used to derive the Heart Rate Variability, by extracting the normal sinus to normal sinus intervals (NN) of the QRS interval.

This distance also represents the person's current heart rate. (Mietus & Goldberger, 2014) The NN interval is then the foundation of Heart Rate Variability which is a measure of the difference in time intervals between heartbeats that is considered normal. This is measured by examining the time in milliseconds between each peak. (Rollin McCraty & Shaffer, 2015).

The heart rate variability measure can be obtained based on consecutive RR samples and measures of heart rate variability is traditionally split into two different categories. The first type is time domain, where the most commonly used time domains for heart rate variability are listed in the table below. (Mietus & Goldberger, 2014)

Table 1 from (Mietus & Goldberger, 2014) shows commonly used time domain measures for heart rate variability

AVNN	**Average of all NN intervals**
SDNN	Standard deviation of all NN intervals
SDANN	Standard deviation of the averages of NN intervals in all 5-minute segments of a 24-hour recording
SDNNIDX	Mean of the standard deviations of NN intervals in all 5-minute segments of a 24-hour recording
rMSSD	Square root of the mean of the squares of differences between adjacent NN intervals
pNN50	Percentage of differences between adjacent NN intervals that are greater than 50 ms

From these five measures SDNN, pNN50 and RMSSD are known as the standard and most commonly used ones.

The second category is frequency domain measures where it is common to use the Lomb-Scagle periodogram, which is a variant of the Fourier Transform. From the periodogram, the following features are commonly extracted. (Mietus & Goldberger, 2014)

Table 2 from (Mietus & Goldberger, 2014) shows commonly used frequency measures for heart rate variability

TOTPWR	**Total spectral power of all NN intervals up to 0.04 Hz**
ULF	Total spectral power of all NN intervals up to 0.003 Hz
VLF	Total spectral power of all NN intervals between 0.003 and 0.04 Hz
LF	Total spectral power of all NN intervals between 0.04 and 0.15 Hz
HF	Total spectral power of all NN intervals between 0.15 and 0.4 Hz
LF/HF	Ratio of low to high-frequency power

The Lomb algorithm is especially suited for data with uneven samples, which is in the nature of RR intervals in contrast to the Fourier Transform which requires data to be sampled at equal time intervals. The generalised form of the Lomb-Scargle Periodogram equation is given by the following equation (VanderPlas, 2017)

$$P(f) = \frac{A^2}{2}\left(\sum_n g_n \ COS(2\pi f[t_n - \mathcal{T}])\right)^2 + \frac{B^2}{2}\left(\sum_n g_n \ sin(2\pi f[t_n - \mathcal{T}])\right)^2$$

Research shows that the different parts of the spectrogram show significant fluctuations when measuring stress (Hjortskov, et al., 2004). Hjortskov et al. conducted a study in a stressing work environment where they measured the effects of mental stress on heart rate variability. Findings show that both the HF and the LF bands were significantly lower during resting periods compared to work sessions and the ratio between the LF and HF was significantly reduced during resting periods (Hjortskov, et al., 2004).

2.2.2.3 HEART RATE VARIABILITY DIFFERENCES BETWEEN INDIVIDUALS

heart rate variability varies between people depending on several factors such as age, gender, fitness level and other factors such as both physical and mental health. (Moore, 2017). Data shows that the biggest factor for heart rate variability among healthy individuals are age and aerobic fitness level. People in the same age have similar levels of heart rate variability (Altini, 2016) and the level is greatly affected by aerobic fitness level, and individuals with a higher aerobic threshold has a healthier heart rate variability compared to its peers. In a study conducted by Melo et al. to analyze the sedentary versus active young and older males, the researchers found that the RMSSD of active individuals were higher than its non-active peers of similar demographics. (Melo, 2005)

Further heart rate variability are directly linked to health issues, where a lower heart rate variability is an indicator of compromised health. For example, low heart rate variability is connected with several disorders for diabetes, mellitus, sleep problems and difficulties regulating emotions. (Subhadra Evans, 2013)

2.2.3 DIFFERENT WAYS OF MEASURING ACTIVITY OF THE HEART

ECG (Electrocardiography) and PPG (photo plethysmography) are commonly used to measure heart rate variability. ECG measures changes in electrical signal directly produced by the heart and is commonly referred to as the gold standard, while the PPG sensors commonly found in wearable devices measure electrical signals based on light reflected from blood flow changes. (Plews, et al., 2017) Studies show that PPG is not as accurate as ECG, but it's satisfactory giving results for measuring heart rate variability. For example, a study carried out to compare PPG, and the Polar H7 chest strap concluded that both of them gave satisfactory results compared to ECG as a reference point (Plews, et al., 2017) . The researchers also found that the PPG sensors might be a better choice for athletes as its more practical and easy to use compared to for example the H7 that requires the user to put on a chest strap.

2.2.4 Measuring stress with heart rate variability

Several studies successfully connect heart rate variability measures to stress for example a study conducted by (Taelman, et al., 2008) describes how stressing factors affects physiological factors in the body and how the autonomic nervous system is activated, leading to change in both heart rate and heart rate variability. (Taelman, et al., 2008)Another study done by (Hjortskov, et al., 2004) comes to the same conclusion. They further conclude that heart rate variability is a more suited measure of short term stress than blood pressure as stress leads to a higher heart rate, and a following shorter time between each RR interval (CHUDY, 2017) .

2.2.5 Measuring stress with galvanic skin response

Galvanic Skin Response sensors (GSR) are commonly used to measure stress by measuring the resistance of the skin, and the more stressed a person is, the more the persons sweat and resistance

decreases (Villarejo, et al., 2012). Research conducted by (Villarejo, et al.) were able to correctly classify the persons stressed state with a success rate of almost 91% by just using GSR measurements, but they note that in general settings it's hard to differentiate between being stressed and just normal sweating from physical activity.

2.2.6 Wearable devices

Devices from manufacturers such as Apple and Fitbit comes with highly accurate optical heart rate sensor (Apple, 2017) (FitBit, n.d.) that deliver heart rate readings as beats per minute. Unfortunately, most of these devices do not report RR intervals as they tend to smooth the signal to more accurately deliver heart beats per minute (EliteHRV, n.d.). The lack of RR samples leads to the loss of potentially valuable information about for example the persons health as more heart rate variability shows that the heart is responding properly to variations in related physiological signals such as respiration. (Sisson, 2014)

Providentially, there are trackers on the market today that with a high degree of accuracy can deliver HRV readings. For example, the Band 2 from Microsoft is one of the wearables devices available that accurately measures RR peaks. In a study conducted by (CHUDY, 2017) the Band 2 was compared to the performance of an ECG device in measuring RR peaks by measuring 49 students taking a memory test in front of a computer, the results show that there was a respectable consistency between the two devices. Another device, the Zoom HRV made by LifeTrak claims to accurately measure RR peaks on the wrist (LifeTrak, n.d.), but for the time writing there is no available open research supporting their claim.

2.3Machine Learning

Machine Learning is a method where machines are trained to detect patterns in data (Varun Gulshan, et al., 2016) to provide automated approaches and tools for data analysis (Murphy, 2012).

2.3.1 Supervised learning and classification

The machine learning techniques used in this project are all supervised, this means that the data provides labeled targets for the algorithms. The overall aim of supervised learning is to find a function *f(x)* that predicts a target *y*. (Murphy, 2012)

2.3.2 Algorithms for supervised learning

This project uses a range of different classification algorithms and techniques to find the approach that gives the best results. Description for the algorithms and techniques used follow below.

2.3.2.1 Automated machine learning and evolutionary algorithms

Automated machine learning is a technique where a tool automatically tries of a range of different hyper parameters and models to find the best solution for the given problem. This approach builds further on regular hyper parameter tuning such as grid search and randomized search. Grid Search (Olson, et al., 2016) is a method that uses a brute force approach to find the best parameter for the problem to be solved. While randomized search is a method that tries a random combination of parameters (Bergstra & Bengio, 2012). Tuning hyper parameters can drastically improve the performance of a model and it's crucial to have as well tuned parameters as possible. (Feurer, et al., 2015). Research shows that automatic hyper parameter tuning in most cases outperform manual tuning due to its ability to test a wider range of parameters (Bergstra & Bengio, 2012).

Automatic parameter tuning has, in turn, led to evolutionary algorithms, which is a genetic approach to programming where a tool by itself tries to find the best model to fit the data by maximising the classification accuracy. The optimisation is done by automatically doing optimisation and applying transformations on given supervised learning data set. (Olson, et al., 2016). The idea behind genetic algorithms consist of creating an initial population of different models, then for each iteration, we do something called a fitness test. Which similarly to natural selection only keeps the best performing models to the next iteration. This approach allows for testing a big range of different models, and build on what the previous generation did well, thus being more efficient than a pure brute-force approach. (Shiffman, 2012)

2.3.2.2 Algorithms

2.3.2.2.1 Naïve Bayes

Naïve Bayes is a classification method based on the Bayes theorem where it derives the probability of the given feature vector to be associated with a label (Scikitlearn, 2016). The Naïve Bayes has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent which not always is the case. Regardless of the naïve assumption the algorithm has proved to be a good classifier especially on smaller datasets and its quite immune to overfitting due to its simplicity (Murphy, 2012) This project uses the Gaussian Naïve Bayes classifiers, a classifier that assumes the likelihood to be distributed according to a Gaussian distribution.

2.2.2.2.2 Decision Trees

Decision trees try to learn to predict an outcome by making small decisions inferred from the features. (Scikit-learn, 2017) Each node in the decision tree represents a decision, and as a feature goes through the nodes in the tree, it will in eventually end up in a *leaf* holding the end class. Decision trees are easy to understand, and they are easy to visualise, but a down side is that it often makes to complex models and its prone to over fit (Scikit-learn, 2017).

2.3.2.2.3 ENSEMBLES

Ensembles in the context of machine learning is a technique where a range of models is used to make a prediction. The ensemble is built up of two components, the models and decision rules that decide the combination of a unified predictor based on each steps results. (Hearty, 2016) This project uses the averaging and boosting methods. The averaging ensembles develop a range of models and use techniques for averaging to pick the best parts from each model into a unified model. (Hearty, 2016) The algorithms used in this project is the Extra Trees classifier and the Random Forest classifiers.

The random forest fits several decisions trees on subsamples of the data and uses averaging to create the final model. (Scikit-learn, 2017)

The Extra tree classifier is similar to the random forest and works by fitting various randomised, random forests trees on random sub-samples of the data.

Boosting ensembles builds models in sequences where each model tries to improve the score of the overall model. (Hearty, 2016) This project uses gradient boosting approaches through the Xgboost and the Adaboost modules. Gradient boosting techniques is similar to gradient

descent which iteratively minimises the error by moving towards a loss gradient. (Hearty, 2016)

2.3.2.2.4 K NEAREST NEIGHBOURS

The foundation of K nearest neighbour is to find the K number of nearest neighbours to a given sample, then classify the sample based on the class of the majority of the closest neighbours. (Scikit-learn, 2017)

2.3.2.2.5 LOGISTIC REGRESSION

Logistic regression is a binary classification methodology that learns the probability of a feature belonging to a certain class. The objective function of Logistic Regression is to minimise the logistic function, a variation of a sigmoid function. (Murphy, 2012)

2.3.2.2.6 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is an algorithm used for classification that uses linear mappings in a high-dimensional feature space to learn a nonlinear function. (Murphy, 2012)

2.3.2.2.7 MULTI-LAYER PERCEPTRON NEURAL NETWORK

Neural networks are biology-inspired machine learning approaches meant to mimic the brain. Multi-layer Perceptron (MLP) also known as feedforward neural network is a type of neural net consisting of several logistic regression models as hidden layers and a final logistic regression model as output layer (Murphy, 2012).The MLP classifiers are trained using backpropagation meaning weights are iteratively adjusted by updating weights backwards from the output to the input.

CHAPTER 3: IMPLEMENTATION AND METHODOLOGY

This chapter describes the approach taken and the dataset used.

3.1 Design Science Research

There are several different scientific methods, but aimed at an information technology project where the goal is to create a useful artefact, Design Science Research distinguishes itself from other natural science methods, as the goal is not only to understand and observe but also to create (Peffers, et al., 2007). The Design Science Research framework has previously been used in related projects, for example in the early nineties to create a data warehouse consisting of health data from numerous different sources in the United States health system (Peffers, et al., 2007)

The main principle of Design Science Research is that knowledge can be materialised, aiming at solving a problem that may be of interest to the general public (Johannesson & Perjons, 2014). Additionally, this is an iterative framework where the process and different iterations of the artefact will be as significant as the first version of the artefact (Abbasi, et al., 2016). This is because IT systems usually are delivered in iterations where features are gradually implemented and created, the same applies to this project where data and algorithms are iteratively improved.

3.2 Tools and techniques

The project is written in Python using libraries including Pandas, Numpy, SciKit-learn, TPOT,auto-sklearn and more. Additionally, Physionets terminal tools are used for data reading and transformation; these tools are run in a virtual Ubuntu 32bit installation due to a 32-bit requirement by some of the legacy tools. The Spyder IDE found in the data science platform Anaconda is used for software design. GitHub was used to backup and store different iterations of the code, following standard coding practice.

3.3 Data source

The dataset is from a project conducted at MIT by Healey as a part of her PhD thesis (Healey, 2000) and is available from Physionet (Goldberger, et al., 2008) . The data consist of body measurements conducted on various young people driving in stressing environments, e.g. rush hour, highways, red lights. The drivers initially start and end with a resting period in a garage where they sit for several minutes to get a base reading; then they have to drive out of a six-story garage which initially raises the stress of the drivers. The research measured physiological signals such as heart beats, ECG and EKG and galvanic skin response measures which allow for determining when the person was stressed. The data does not natively contain any heart rate variability measures, so these are created from the ECG data by using a beat-by-beat annotation tool to annotate RR intervals from the ECG samples. (Moody, 2016)

Chapter 4: Data Preparation

4.1 Dataset and tools

The dataset is in a physionet specific format divided into 18 *.dat* files and 18 *.hea* files with accompanying meta data. The data consists signals for ECG, EMG, GSR measures from the foot, GSR measures from the hand, HR and Respiration. All values are float values, with a sampling frequency of 15.5 samples per second. The WFDB command *rdsamp* from the native terminal installation of Physionets tools named WFDB is used to read the data (Moody, 2015), then they are merged and saved as .txt files with column names, the measurement unit and the time in seconds for each row – including the data samples. The header names are manually cleaned and then the data is stored in a Pandas dataframe. Each file contains a sampling time starting at zero and stopping at the

end of the sampling session. The time interval is incremented based on the last time interval of the previous file to transform the data into one continuous time-series. See appendix A.1 for detailed file content.

4.2 Getting RR intervals

The ECG signal is used to derive the RR intervals by first creating beat annotations files from the .dat files by using WQRS, a tool available through the Physionet toolbox (Zong & Moody, 2015).These map each heartbeat and gives it a timestamp (Physionet, 2016) along with a beat annotator to specify if the beat is normal or abnormal. The annotation files are then read with Physionets HRV toolkit to extract the RR intervals along with a timestamp for each interval. The RR intervals are saved as a space separated text file where the first column is the timestamp, the second column is the RR interval, and the third column annotates the beat as a normal or abnormal.

4.2 Initial data exploration and cleaning

4.2.1 Initial Data Summary

Doing simple summary statistics on the relevant data before the cleaning process shows the presence of outliers and missing data in all the individual signals. This is most likely due to sensor issues such as outages and poor contact. See appendix A.2 for the table with the indicators highlighted.

4.2.2 Inspecting the galvanic skin response values

The initial data summary indicates outliers and missing data in the GSR values as there is a big variation between the median and the mean for both the foot and the hand signal.

The figure below shows foot and hand GSR measurements plotted against each other as a time series, where the x axis represents the index of the dataframe. The plot evidently indicates that the two GSR measurements follow each other's peaks, but there is strong visual evidence for noisy data with outliers and missing values. The GSR plot shows that the GSR readings from the hand consist of big periods without readings, while the readings from the foot were more consistent with very few visual missing periods.

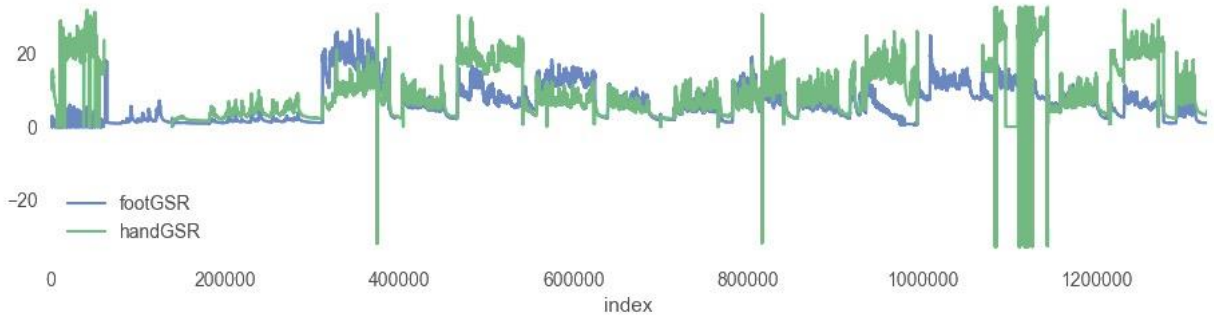


Figure 3 shows the GSR values for both foot and hand measurements plotted

4.2.3 RR STATISTICS

The RR intervals were plotted to see the distribution of the data, where it is clear that the distribution is according to a Pareto distribution and the majority of the data is within one a standard deviation from the mean, but some extreme outlier values are giving the distribution a heavy tail.

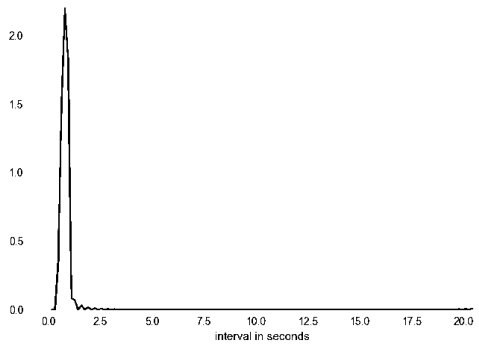


Figure 4 shows the distribution of the RR data

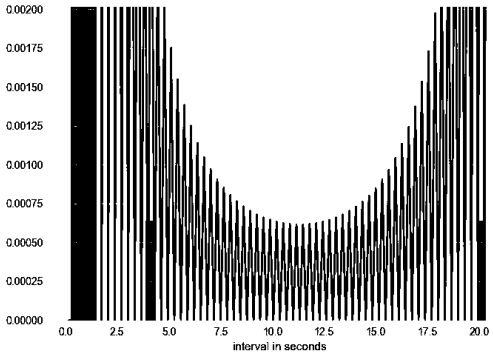


Figure 5 shows the lower values of the distribution

Additionally, there are values below the normal range of RR peaks from 0.6 to 1.2 seconds (Christensen, 2014)and it is safe to assume that these values are present due to sensor issues, these values also are pulling the mean slightly away from the median. See appendix A.3 for a time series visually showing the presence of outliers in the RR data.

4.3 Cleaning the data

4.3.1 Cleaning galvanic skin response readings

Because the two GSR readings follow each other quite closely, the readings for the foot will be used to label the stressed state. A median filter is applied to clean the signal, this type of filter can remove noise while keeping the peaks opposed to for example a moving average filter where peaks are not preserved leading to the loss of information (Bakker, et al., 2011). The median filter works by going over each sample at the time, to determine if the sample is representative according to its neighbouring samples. When a sample deviates too much compared to the samples within the sliding window gets sorted from smallest to largest, and the middle value is used to replace the deviating value. (R. Fisher,

2003) The implementation used in this project uses the median filter available in the signal processing library SciPy. The SciPy median filter takes two inputs, an array of samples and a kernel size. The kernel size is the number of samples to be used in the sliding window. (SciPy, 2014)

In Figure 3 from the previous chapter, it was visible unwanted sensor artefacts. Additionally, the GSR values have different ranges based on what reading they are from caused by the different subject's level of sweat (Healey, 2000). This means that the outlier visible in the first part of the data is interpreted as a normal value in the dataset as a whole. Therefore, the different readings are handled separately to make sure each GSR value is labelled and cleaned correctly. The outliers are removed with a median filter after visually inspecting several kernel sizes, a kernel size of 13 was small enough to give good results and eliminate the outliers while keeping the spikes. See appendix A.4 for a time series of the GSR data after applying the median filter.

4.3.2 Cleaning RR intervals

The RR intervals consisted of some significant outliers and artefacts dropping down to closer than zero than normal RR intervals. The data was cleaned by replacing all values below 0.5 and above 1.5 with the median to ensure no values are outside the normal RR interval range. After that, a median filter with a kernel of five is applied to smooth the signal after the replacement process. See appendix A.5 for statistics and a time series of the RR interval data.

4.3.3 Cleaning heart rate data

The heart rate data is cleaned by first checking for infinite values and replacing them with NaN; then the NaNs are replaced with the mean HR. In the end, a median filter with a kernel of 13 is applied to remove unnatural heart rates further.

4.4 Combining RR peaks with ECG data

The next step is to join the RR peaks datasets and the ECG data set, this is done by a separate function that read the RR files, and the ECG files pairwise and does an inner join keeping only RR peaks and associated data. The *time* column is used to match the RR peaks with the right data.

The total length of the DataFrame after the join is 107592, which is one less than the original RR DataFrame as there was a missing row in the 7th part of the ECG data leaving one RR peak without an ECG partner. See appendix A.6 for a visual representation of the combination process.

4.4.1 Time domain measures

The goal of this project is to use heart rate variability measures to predict stress, the heart rate variability measures are a set of different computations in both time domain and frequency domain as described in the literature review. To calculate the various heart rate variability metrics, the data set is run through a middleware function that connects the C implementation of Physionets HRV toolkit to Python. Python then creates a file containing the RR peaks and the time; then a system call is triggered by Python to activate the *get_hrv* function in Physionets tool which reads the file and returns the heart rate variability metrics to Python.

The table below describes the entire dataset with some of the common descriptive statistics for NN intervals as outlined in the literature review. The SDANN and SDNNIDX are calculated by measuring on five minutes bins on a 24hour dataset, the dataset, in this case, is approximately 22.5 hours and not full 24 hours, but the same approach is used as the dataset is very close to a full 24 hours measurement. When comparing the results from the time domain measures before and after

the median filter is applied, we can see that there are significant changes to the values. Particularly with the RMSSD which was unnaturally high before it the cleaning due to the outliers.

Table 3 shows time domain measures with and without a median filter

STATISTIC	NO FILTER	MEDIAN FILTER
NN/RR	0.9999	0.9999
AVNN	0.7438	0.7831
SDNN	0.9390	0.1383
SDANN	0.1501	0.1225
SDNNIDX	0.2204	0.0597
RMSSD	1.3017	0.0316
PNN50	0.6302	0.1799

4.4.2 Frequency domain measures

The next step is to see how the values look in the frequency domain; here the Lomb Periodogram is used to apply frequency measures (Mietus & Goldberger, 2014). The same approach as with the time domain measures is used to implement the frequency domain. The table below shows the frequency measures, where it is evident that the median filter removed unwanted samples as particularly visible with the total power that has shrunken significantly.

Table 4 shows frequency domain measures with and without a median filter

STATISTIC	NO FILTER	FILTER
TOTPWR	0.5607	0.0206
ULF	0.0312	0.0170
VLF	0.0615	0.0018
LF	0.1461	0.0011
HF	0.3217	0.0005
LF/HF	0.4542	1.9348

Before median filter

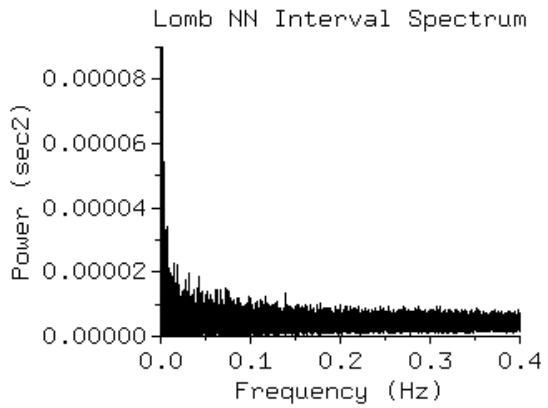


Figure 6 spectrogram of the RR data without a median filter

With median filter

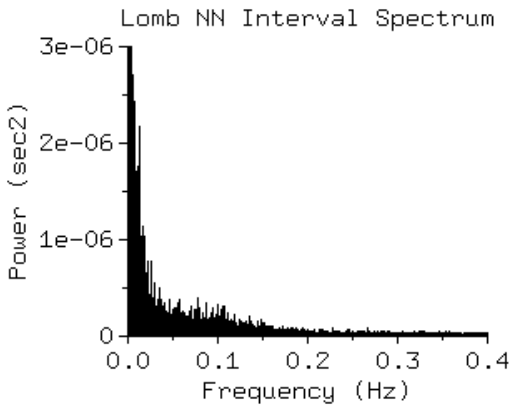


Figure 7 spectrogram of the RR data with a median filter applied

4.5 STRESS VALUES

The task to determine the stressed state of a person is framed as a binary classification problem, where the goal is to assess the stressed state as stressed (true) or not stressed (false). The context of this data is people sitting still in cars driving in stressful/stress inducing environments; this means the data does not contain peaks created due to extraneous reasons e.g. physical activity.

4.5.1 Labelling RR peaks as stress

To assess if the driver was stressed or not stressed, the median of the GSR values is taken as the cut-off point, and any value above the median value is labelled as stress, and any value below the median value is labelled as not stressed. The figure below shows in red the portions of the GSR signal marked as stress.

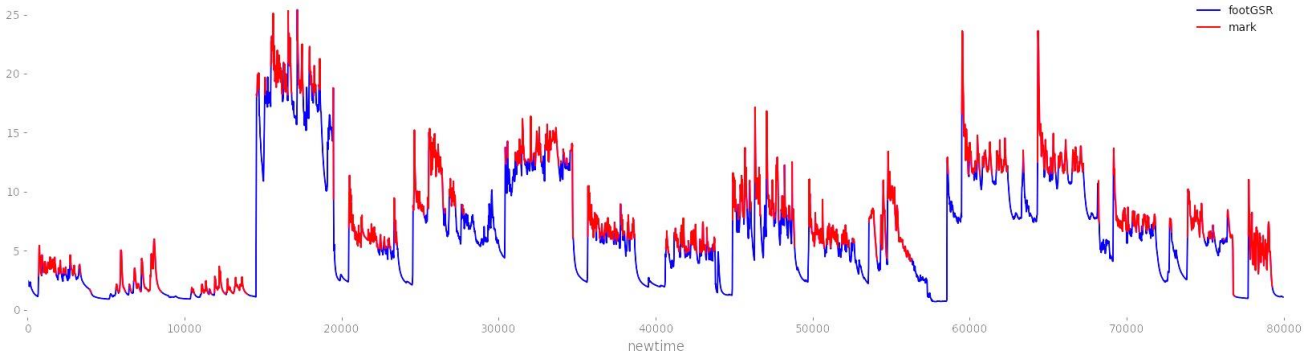


Figure 8 shows the portions of the GSR signal marked as stressed in red.

Following this method around half of the data is labelled as stressed, a labelling one can reason around being quite accurate as the drivers were not stressed at all times, e.g. when standing still in the garage. The driving route consists of a lap that takes up to 1.5 hours to complete and starts off in a garage where the subjects had to close their eyes and relax for 15 minutes to create a non-stressed base measure. Then they headed off onto a narrow ramp declining six floors down, before they headed straight to a busy main street where drivers encountered stop and go traffic, pedestrians, cyclists and other stressing factors meant to create high-stress levels in the subjects. Then they headed away from

the city and over a bridge and onto a highway with a toll on the way, this was meant to trigger medium stressed levels. After that, the drivers took the same path back to the motorway and through the busy city and back up the six-floor garage path where the subjects again rested for 15 minutes. (Healey, 2000)

The total number of values labelled as stressed is 58039, and the number of values labelled as not stressed is 49553. What's evident here is that the data is somewhat imbalanced, and important factor to consider when modelling the algorithms and choosing evaluation metrics.

CHAPTER 5 FEATURE ENGINEERING AND DATA PREPARATION

Feature engineering involves transforming and combining features in an attempt to produce better representations of the dataset for modelling.

5.1 HEART RATE VARIABILITY FEATURES

The most important features are the heart rate variability measurements from the RR peaks. These are obtained by sending the dataset through a function that splits it up into pieces of 20-second windows and adds ten samples from the previous window and five sample from the next window. This results in a window size of about 30 seconds and avoids a hard cut-off by using overlapping windows. This approach makes it possible to include events at the end or the start of the windows (Matthey, 2009) so that more RR peaks are counted in when calculating heart rate variability features, at the same time the resulting dataset will have more samples than if the window was e.g. 30 seconds without overlapping. Each window is directly sent through the function that gets heart rate variability measures so that the measures are applied to the RR intervals of that window. When the heart rate variability measures are obtained, they are cleaned and combined with the average of the original

features from the current window. See appendix A.7 for an overview of how the new dataframe appears.

5.1.1 Cleaning heart rate variability features and labels

The newly created heart rate variability features contain some rows with infinite and NaN values, the approach to fix this was first to use numpy to replace all infinite values with NaNs, then replace the NaN values with the mean of the given column. Additionally, as the mean was used to merge the values, the labelling is now no longer limited to 1 and 0. This is fixed by replacing all values above 0.5 with 1 and all values bellow 0.5 with 0.

5.2 Feature selection

The full feature set was sent through a random forest model to measure the feature importance, to determine the usefulness of each feature. The function to determine relative feature usefulness is based on the feature selection model from Sklearn and is using an Extra Tree Classifier to rank the features (Scikit-learn, 2017). The Extra Tree classifier takes the mean importance over the number of trees which in this instance is set to 1000 trees.

Before the modelling stage, some features are removed as some of them directly contain the potential for data leakage which can lead to overconfidence in the performance of the model. Data leakage is the notion of information about the target variable being included in the features used for prediction (Kaggle, 2013). Some of the features are potential holders of information about the target, and these are not to be included in the feature set fed to the model. The column containing the stressed stage is left behind, the same applies to the GSR values as they are the basis for the stressed state and a strong potential and obvious avenue for data leakage. The data leakage is evident when inspecting the results

from the feature importance determinator, where the GSR values from foot measurements are by far the most important feature, this is also the feature that lays the baseline for labelling the data set. See appendix A.8 for an overview of the importance of features.

Further, the marker column is mostly a NaN value throughout the data set and is discarded as it provides no useful information. Additionally, the time columns are unwanted and not used in the prediction. Furthermore, since this project aims to use Wearable devices as a basis, the ECG and EMG data is left out as this is information not easily obtained from wearable devices. The ULF band is also discarded as a majority of the values are zero, which is probably due to the short time interval. Also, the VLF band is discarded as research shows that the VLF band proves to be an unreliable measure in readings under 5 minutes (Hjortskov, et al., 2004).

This leaves the dataset with 10 features, three from the original data and the rest being heart rate variability measures. See appendix A.9 for a list of features selected.

The newly created dataset is run through the feature importance model where it is evident that most features are relevant in predicting the outcome, but with the RR interval, respiration and HR being the more important ones and the high and low-frequency measures scoring relatively low. Note that respiration typically, and including in this dataset, is collected with a strap around the chest, but the respiration feature is kept as research show that respiration is possible to derive from waveform analysis of PPG data (Meredith, et al., 2012)

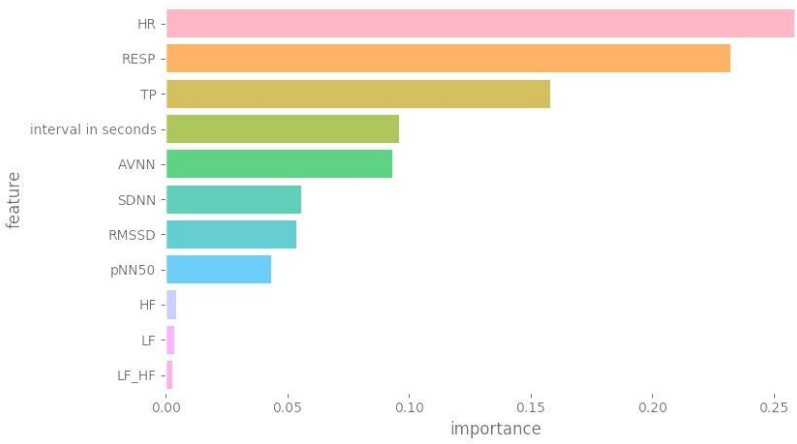


Figure 9 shows the feature importance of the data set. The heart rate is the most important feature, and the high and low-frequency band is the most insignificant parameters

5.2.1 Creating more features with a polynomial feature expansion

More features were created using Polynomial feature expansion through the Polynomial pre-processing library in scikit learn (Scikit-learn, 2017).This creates a new matrix of Polynomial combinations of the existing features, resulting in a new dataset with 66 features using a polynomial degree of two, the new features can uncover new relations between features that are not independently evident. The polynomial degree was decided by testing different degree values by using cross validation in the benchmarking model described in 5.4 Benchmarking as it can lead to overfitting if a too high polynomial order is used. The testing performance of polynomial degrees started to decrease from four, and the model began to overfit on the training data, and at this point, there were too many features for a realistic model.

5.3 Evaluation Metrics

The next step is to choose a suitable evaluation metric to be used as the basis for model comparison, the metric chosen is the F1 measure, chosen after a careful consideration of the modelling problem as well as the nature of the dataset. The F1 measure is more accurate on an imbalanced dataset than for example accuracy which uses a more naïve approach to evaluate the scores where the algorithm can label all data as the dominant class resulting in a high score, but not a useful model. (Jeni, et al., 2013)

The F1 measure is computed by taking the harmonic mean of precision and recall. Recall is given as $TP/N+ = p(\hat{y} = 1|y = 1)$ and is a measure of how many of the *true positives* that were found, while precision is given as $TP/N\hat{+} = p(y = 1|\hat{y} = 1)$ and is a measure of how many of our detection that is positive. The F1 measure is given in the equation below, where R is Recall, and P is precision. (Murphy, 2012)

$$F1 \triangleq \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{R + P}$$

5.4 Benchmarking

A simple linear Support Vector Machine (SVM) was created as a baseline to compare further work and results. The features given to the model was the raw RR intervals and the corresponding heart rate after the initial cleaning. The benchmarking function splits up the data in a training and testing set where it trains on the training data and tests the model on the test data with five-fold cross-validation. The function then scores the model with the F1 score. The initial benchmark achieved an f1 measure of 0.5609.

5.4.1 Benchmarking new features

With the new feature set, the benchmarking model was again tested to see if there is a performance gain, both with and without the polynomial features. Without the polynomial features, the new feature set achieved an F1 measure of 0.5315, with the Polynomial features the resulting score is 0.6175. Overall the f1 score had an increase over the raw RR intervals, up from 0.56 to 0.61 when using the Polynomial features, but with the non-polynomial features, the prediction score dropped compared to using raw RR intervals and heart rate.

Chapter 6: Machine learning modelling

Following in the analysis and feature crafting of the dataset is to build a classification model that can classify the stressed state as a binary problem.

6.1 Data Preparation

The correct procedure for modelling involved splitting the dataset into two different subsets; training and test. This is done to minimise the risk of the model overtraining on the available data and making sure the model is as generalizable as possible. By splitting the dataset in such way and having separate training and testing sets allow the model to be evaluated on data it has not seen earlier, this enables a healthier performance measure to be achieved. The splitting was done using Sklearns *train_test_split* function which splits the data into random subsets of training and testing sets. (Scikit-learn, 2017)

6.2 Model selection and parameter tuning

The modelling approach taken is through automatic machine learning where both grid search and genetic programming is utilized. Taking this method allows for a wider algorithmic test as opposed to a manual approach where models and hyperparameters must be customised manually. Genetic

programming aspect is done through the Python library TPOT, which is an open-source genetic programming library well recognised for its performance and ability to construct pipelines with high accuracy. (Olson & Moore, 2016) Apart from just tuning the algorithm itself, TPOT also does some level of feature preprocessing and feature selection where all this is combined with a pipeline utilising genetic programming. See Appendix A.10 for an illustration on how the TPOT pipeline is composed.

Auto-Sklearn is the second library used and in contrast to TPOT that tries to optimise the whole Pipeline, Auto-Sklearn focuses on tuning hyper parameters based on Bayesian Optimization and ensembles (Feurer, et al., 2015) In addition to the two libraries mentioned above, a Neural Net optimised with Grid search is also tested. The neural net is a Multi-layer perceptron classifier found in the Sklearn library (Scikit-learn, 2017). Grid search on the neural net is done through *GridSearchCV*, also a part of the Sklearn library (Scikit-learn, 2017). The general notion is that with data with good feature representation a simple model should be sufficient to give a good result, but in cases where this does not apply, a complex model such as a neural net has the potential to compensate for the lack of good feature representation.

Auto-Sklearn and TPOT are customised so that it optimises for one algorithm at the time, then the score for each algorithm is returned after a five-fold cross validation. This allows for comparing the classification score for each algorithm, as opposed to only the highest performing algorithm which the libraries return as default. The algorithms used from each library are given in Appendix A.11.

6.2.1 Algorithm parameters

The TPOT library is set to train one algorithm at the time with a population size of 100 and number of generation set to 400. The Grid search of the Neural Net and the algorithms in Auto-Sklearn tries

different parameters for the different hyperparameters such as learning rate alpha, activation layer, and hidden layer sizes.

6.3 Model scores

This section compares the individual F1 scores from each algorithm across TPOT, Auto Sklearn and the Neural Net. Random forest in Auto-Sklearn is the highest scoring algorithm with a score of 0.8346. Overall the algorithms score above 0.6, and the best scoring models score at the upper end of 0.7 and the lower end of 0.8.

There isn't much research to compare results to, except the research by (Liu & Ulrich, 2014) where they got a F1 score of 0.7855, and these scores show that their findings are possible to achieve. The best scoring TPOT algorithm uses the Xgboost model with a score of 0.7815, slightly lower than Auto-sklearn. These results show that the decision tree algorithms do an exceptional job in both TPOT and Auto-sklearn. The Neural Net implementation used "tanh" activation layer, alpha value of 0. 0001and a hidden layer size of 300,300,4 and scored 0.7241, effectively lower than the shallow algorithms.

Algorithm	Auto-Sklearn	TPOT	Grid Search NN
Random Forest	0.8346	0.7735	n/a
K nearest neighbours	0.8289	0.7470	n/a
Gradient Boosting	0.8279	0.7794	n/a
Extra Trees	0.8086	0.7691	n/a
Adaboost	0.8053	n/a	n/a
Decision Tree	0.7934	0.7595	n/a
XGboost	n/a	0.7815	n/a
MLP	n/a	n/a	0.7241
Linear SVM	0.7468	0.7117	n/a
Logistic Regression	n/a	0.6364	n/a
Gaussian Naive Bayes	0.6073	0.6294	n/a

Table 5 shows F1 scores from each algorithm where the random Forest algorithm has the highest score with 0.8346

This section gathers data from wearable sensors from a Polar H7 HRM chest strap and a Fitbit Charge 2 wrist-worn fitness tracker and applies the model to detect stress on never before seen data.

7.1 Creating a dataset

The dataset was generated by building a test based on a study by (Javad, et al., 2016) where the researchers provoked physical, cognitive and emotional stress. They provoked physical stress by a walk, followed by a light jog on a treadmill. Cognitive stress was provoked by counting backwards by sevens from 2485, following a test where subjects had to tell the name of a colour written in a different colour than inscribed. Emotional stress was provoked by showing the subjects a five-minute clip from a horror movie. They then concluded that these tests successfully were able to reproduce the different stress types. (Javad, et al., 2016)

Following similar test in this project where the test starts with a 5-minute relaxation laying down, then a 15-minute walk starting with a six-floor descent down stairs, then a short walk along a busy road and again ending in the same place climbing six floors. The physical movement of the test is supposed to lower the RR intervals and raise the heart rate creating physical stress so that the model can differentiate between physical stress and mental stress by using the accelerometer available in the Fitbit device. The walk is followed by a rest period of 5 minutes laying down, to make the heart recover from the physical load. After that, an episode from the violent TV series Vikings was played to provoke emotional stress where the theory is that only violent scenes will impact the RR intervals and cause short-terms stress. The stress test is measured on a healthy 22-year-old male, which is similar demographics to the training data created by (Healey, 2000).

7.2 Model and hardware

The RR interval data was gathered with the Elite HRV iPhone app connected to a Polar H7 HRM chest strap as the Fitbit device do not give access to individual RR peaks, just minute to minute heart rate readings. The data from the two devices are merged, where accelerometer data in the form of steps and minute to minute heart beat is used from the Fitbit. The step data allows tagging RR peaks as a movement to avoid miss-classifications as a result of physical activity. The data is then run through the function that derives heart rate variability features, and a DataFrame containing the features for heart rate, RR peaks, and the associated heart rate variability features is created, this Dataframe is then run through the polynomial feature expansion. After that, the Neural Net, Auto-Sklearn and TPOT is re-trained on the original Dataframe but with the same features as available with the new dataset. This leaves out the respiration feature used earlier.

The models are then used to predict and label the stressed state on the new unlabeled data set. The predicted stressed labels are combined with the new data set, and the steps column is added to mark parts of the RR data that was recorded during movement. Next, the prediction is plotted to visualise the prediction and compare the results of the three modelling approaches.

7.3 Annotating stress

The subject had never before been exposed to the particular episode played but is familiar with the TV show. When the episode is over the subject continued to lay down and relax for a few minutes to calm down. The subject was told to mentally log scenes that could cause stress and after the session write down parts of the clip where stress might have occurred. The subjective notes are compared with the areas of the reading annotated as stress and compared to the scene occurring during the period of the classified stressed moments.

7.4 Test results

The subject noted that he only felt stressed during a scene around 10 minutes out in the clip. This scene lasts around five minutes where Viking warriors are entering a battlefield and include several moments where beloved characters in the series are killed and hurt. Moreover, the subject notes that there weren't any particular stressing moments after the battle scene, but there were some minor scenes that were a bit tense throughout the video that felt medium stressed. These scenes are when a mystical and bleeding person approaches the hometown of the Vikings and when one of the main characters grieve heavily over his friends' death. The reading starts with 5 minutes rest to get a base reading; then the subject starts the physical walk. The subject returned and rested for five minutes after the walk. The rest follows a short period of movement where the subject started the video clip. The video clip lasts for about 45 minutes which is followed by slight movement to turn off the video player then followed by a few minutes of laying down resting to normalise the RR intervals, but these last minutes includes some physical moment that might disturb the readings.

7.4.1 Results from TPOT

TPOT created a KNN Pipeline achieving an F1 accuracy of 0.8022, see Appendix A.12 for full pipeline detail. The figure below shows the RR intervals as a time series in blue, where movement is marked with green and stress is characterised in red. The video is started around the 30 minutes mark and at about the 40 minutes mark the violent scene mentioned by the subject starts and lasts for about five minutes. The segment of the scene has visible deviating RR intervals and is correctly marked as stress by the algorithm. After that, the rest of the episode is without significant stressing moments as reported by the subject and the algorithm has not marked anything as stressed beyond the scene communicated by the subject. When the clip was over the subject clicked the pause button, which

correctly marked as movement and then rested for a few minutes before ending the measuring session. These results correlate with the subjective feeling of the subject, and it correlates down to the minute of the provocative scene occurring on the screen.

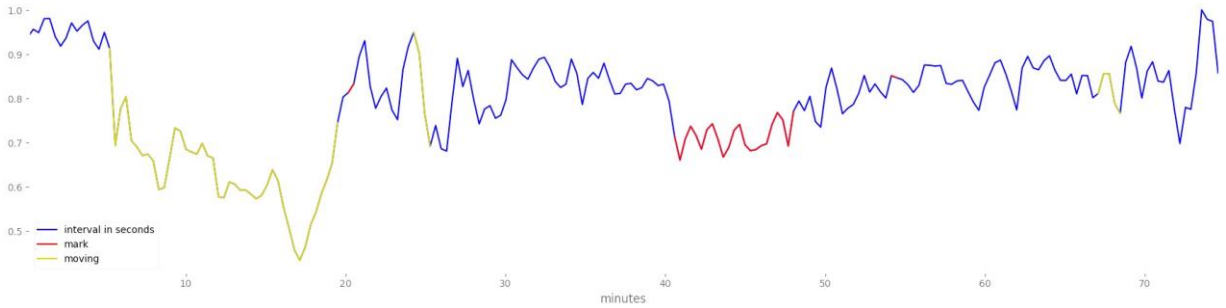
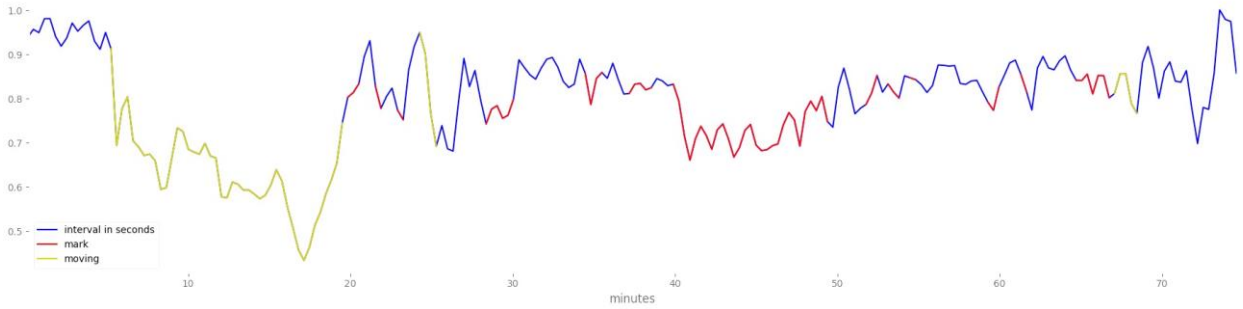


Figure 10 shows a time series of RR intervals where green represents movement and red represent stress; correctly classifying the different segments of the reading.

7.4.2 Results from Auto-Sklearn

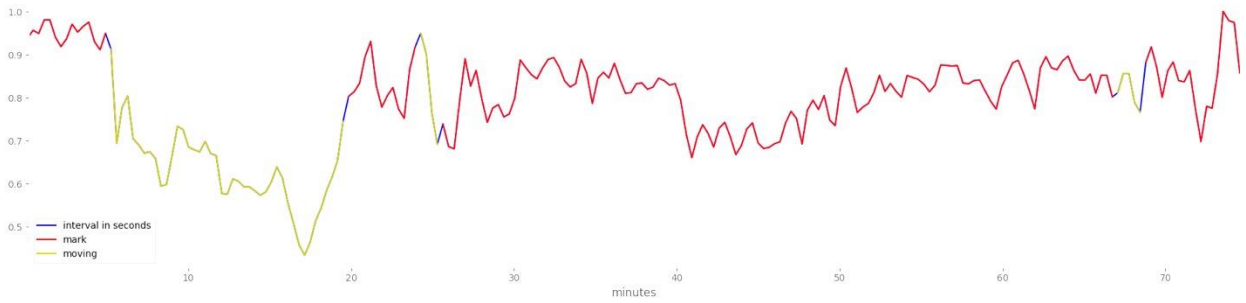
Auto-Sklearn created a Random-forest based model that achieved an F1 accuracy of 0.7838; this is slightly behind TPOT, but not a significant difference. Similar to TPOT the stressing sequence around the 40 minutes mark is correctly classified. Additionally, auto-sklearn have some more classifications as stress that shouldn't have been classified as stress but rather a movement. Straight after the walk while relaxing, some parts in the recovery stage is classified as stressed, this can be due to inactivity where the RR intervals are recovering after the physical activity. Additionally, some dips in the RR intervals at around 28 minutes are marked as stress. This segment was due to movement setting up the video, but the Fitbit did not recognise this as "steps" and thus not marked as a movement. The next segment marked as stress is at around the 37 minutes mark, this part was not reported as stress

by the subject. When manually checking the actions in the video, it corresponds with a scene where the Vikings are getting ready for battle, and the person to later be killed as mentioned in the TPOT results section are in severe pain after losing his arm. The next part classified as stress is during a dip in the RR intervals that are shortly followed by the main sequence of the battle. Following a few minutes after the fight, at the 52 minutes mark, the RR intervals are again marked as stress. This corresponds with the feeling reported by the subject where a bleeding man is entering the Viking village. After that, a stressed mark at 59 minutes corresponds to a scene where one of the main characters grieve heavily over the loss of his friend – also as reported by the subject.



7.5.3 Results from the Neural Net

The Neural Net obtained an F1 accuracy of 0.7241, with best parameters of activation layer tanh, alpha value 0.0001 and hidden layer sizes of 300, 300, 4. When visually inspecting the results the classification done by the Neural net is not correct as it has marked the majority of the reading as stressed. This misclassification might be due to overfitting to the training data, despite doing cross validation.



CHAPTER 8: CONCLUSIONS

This project presents a machine learning approach for detecting stress factors in heart rate variability measures on data available from wearable systems. The results from model tuning and cross validation give a synthetic F1 score of 0.8346 with the Random forest algorithm which in itself is not a perfect classification but correlates well with the results obtained by (Liu & Ulrich, 2014). These results both proves that heart rate variability is a reasonable metric for detecting stress and that automatic machine learning libraries do a good job in finding parameters. When testing the models on data collected from wearable sensors and taking advantage of detecting movement with the accelerometer, the TPOT and Auto-Sklearn library was able to correctly classify the subject as stressed down to the minute by using heart rate, heart rate variability and raw RR intervals as features. This demonstrates that the current consumer technology paired with machine learning techniques is capable of a self-monitoring system that can detect abnormalities in the nerve system as successfully classify stress. The project was developed with the Design Science Research framework in mind, thus attempted to contribute with new knowledge within the areas worked. This project has successfully demonstrated and created a foundation for others to build on for using wearable technology, machine learning, and ETL processes to better deliver health care of higher quality at a lower cost.

One of the major difficulties with this project is the lack of data. The model was trained on a relatively small sample set with about 22 hours' worth of samples and applied to a simple self-test, despite this, both of the tests show good results and demonstrates the plausibility of such as system, but a bigger sample set would provide a more confident model. In addition, the use of automated machine learning algorithms provided the possibility to test a wide variety of parameters, but this also meant less control over parameter tuning and a better approach would be to find the base models with automatic machine learning and further customise them manually to end up with one pipeline or model. Another aspect to keep in mind is that such a system is meant to run on a mobile system, and the model chosen has to be small enough to fit on wearable device.

Chapter 8.1: Future Work

8.1.1 More extensive data set gathered with wrist worn devices

The dataset presented in this project is reasonable small, and it is gathered using more extensive equipment than simple wearable devices, and a more advanced study would have to be done with more data collected from wrist worn or similar devices. Additionally, it would be useful to have features about the person movement such as accelerometer data or simple step data to annotate physical activity in the training data.

8.1.2 Examine different biological processes for stress

This project mainly focuses on heart rate variability, but it would be useful to know what biological processes feasible to measure with wearable devices could be a good predictor for stress. Both different heart rate variability features and other signals and transformations of for example GSR.

8.1.3 Wearable system for cancer monitoring

Because heart rate variability itself is a relatively general measure of the nerve system, there are many potential applications of using a wearable system that monitor the body and could detect abnormalities in the nerve system. For example, have heart rate variability been directly associated with lower fatigue in patients with breast cancer (Crosswell, et al., 2014) . Another study showed that cancer patient with an SDNN of less than 70 milliseconds was less likely to survive treatment than patients with higher SDNN. (Guo, et al., 2015)

CHAPTER 9: REFERENCES

References

Abbasi, A., Sarker, S. & Chiang, R. H. L., 2016. Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems,* 17(2), pp. 1-32.

Altini, M., 2016. *Heart rate & heart rate variability population values: update using data from 3000 users.* [Online]
Available at: http://www.hrv4training.com/blog/heart-rate-heart-rate-variability-population-values-update-using-data-from-3000-users
[Accessed 15 07 2017].

Apple, 2017. *Apple Watch Series 2 - Technical Specifications.* [Online]
Available at: https://support.apple.com/kb/SP746?locale=en_GB
[Accessed 27 06 2017].

Atkielski, A., 2007. *File:SinusRhythmLabels.png.* [Online]
Available at: https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.svg
[Accessed 24 06 2017].

Bakker, J., Pechenizkiy, M. & Sidorova, N., 2011. *What's your current stress level? Detection of stress patterns from GSR sensor data.* Eindhoven, Eindhoven University of Technology, pp. 573-580.

Bergstra, J. & Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research,* Volume 13, pp. 281-305.

Bourghelle, F., 2015. *Development of an automatic drowsiness monitoring system using the electrocardiogram.* Liège(Wallonia): University of Liège.

Cardiogram, 2017. *Cardiogram.* [Online]
Available at: https://cardiogr.am/
[Accessed 28 06 2017].

Christensen, B., 2014. *Normal Electrocardiography (ECG) Intervals.* [Online]
Available at: http://emedicine.medscape.com/article/2172196-overview
[Accessed 05 07 2017].

CHUDY, N. S., 2017. *TESTING OF WRIST-WORN-FITNESS-TRACKING DEVICES DURING COGNITIVE STRESS: A VALIDATION STUDY.* Orlando(Florida): University of Central Florida L.

Crosswell, Lockwood, Ganz & Bower, 2014. Low heart rate variability and cancer-related fatigue in breast cancer survivors.. *Psychoneuroendocrinology,* 07, Issue 45, pp. 58-66.

Diamond, D., 2015. *Google in Health.* [Online]

Available at: https://www.forbes.com/sites/dandiamond/2015/08/11/google-is-now-alphabet-and-it-could-spell-big-things-for-healthcare/#6eab86476c1d

[Accessed 28 06 2017].

EliteHRV, n.d. *Compatible Devices.* [Online]

Available at: https://elitehrv.com/compatible-devices

Fenton, P. N., 2017. *Probability and Risk.* [Online]

Available at: http://probabilityandlaw.blogspot.co.uk/2016/11/queen-mary-in-new-2-million-project.html

[Accessed 28 06 2017].

Feurer, M. et al., 2015. *Efficient and Robust Automated Machine Learning.* s.l., NIPS.

FitBit, n.d. *Fitbit.* [Online]

Available at: https://www.fitbit.com/uk/purepulse

[Accessed 27 06 2017].

Goldberger, A., Goldberger, Z. D. & Shvilkin, A., 2013. *Goldberger's Clinical Electrocardiography A Simplified Approach.* 8th Edition ed. Philadelphia: Elsevier Saunders.

Goldberger, A. et al., 2008. *Stress Recognition in Automobile Drivers.* [Online]

Available at: https://physionet.org/physiobank/database/drivedb/

[Accessed 4 02 2017].

Guo, et al., 2015. Prognostic Value of Heart Rate Variability in Patients With Cancer.. *Clin Neurophysiol.* , 12.pp. 516-520.

Healey, J., 2000. *Wearable and Automotive Systems for Affect Recognition from Physiology,* Massachusetts: Massachusetts Institute of Technology.

Hearty, J., 2016. *Advanced Machine Learning with Python.* 1st edition ed. Birmingham: Packt Publishing Ltd..

Hjortskov, N. et al., 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *Eur J Appl Physiol* , Volume 92, pp. 94-89.

Hodson, H., 2015. *Apple ResearchKit and Watch will boost health research.*. [Online] Available at: https://www. newscientist. com/article/dn27123-apple-researchkit-and-watch-will-boost-health-research/
[Accessed 28 06 2017].

Jacqueline Wijsman, B. G. H. L. H. H. a. J. P., 2011. *Towards Mental Stress Detection Using Wearable Physiological Sensors.* Bostom, IEEE, pp. 1798-1801.

Javad, B., Cogan, D., Pouyan, M. B. & Nourani, M., 2016. *A Non-EEG Biosignals Dataset for Assessment and Visualization of Neurological Status*. Dallas, TX, ieeeexplore.

Jeni, L. A., Cohn, J. F. & Torre, F. D. L., 2013. *Facing Imbalanced Data Recommendations for the Use of Performance Metrics.* s.l., s.n., pp. 245-251.

Johannesson, P. & Perjons, E., 2014. *An Introduction to Design Science.* 2014 edition ed. s.l.:Springer.

Kaggle, 2013. *Leakage.* [Online]
Available at: https://www.kaggle.com/wiki/Leakage
[Accessed 31 07 2017].

LifeTrak, n.d. *ZoomHRV product page.* [Online]
Available at: https://lifetrakusa.com/product/lifetrak-zoomhrv/
[Accessed 28 06 2017].

Liu, D. & Ulrich, M., 2014. *Listen to Your Heart: Stress Prediction Using Consumer Heart Rate Sensors.* [Online]
Available at: http://cs229.stanford.edu/proj2013/LiuUlrich-ListenToYourHeart-StressPredictionUsingConsumerHeartRateSensors.pdf
[Accessed 02 07 2017].

Lorenzetti, L., 2016. *N HEALTH IS TRANSFORMING THE HEALTH CARE INDUSTRY.* [Online]
Available at: http://fortune.com/ibm-watson-health-business-strategy/
[Accessed 28 06 2017].

Matthey, J., 2009. *Prosig.* [Online]
Available at: http://blog.prosig.com/2009/07/20/data-windows-what-why-and-when/
[Accessed 19 07 2017].

Melo, R. S. M. S. E. Q. R. M. M. R. M. .. .. .. C. A., 2005. Effects of age and physical activity on the autonomic control of heart rate in healthy men. *Braz J Med Biol Res Brazilian Journal of Medical and Biological Research,* 38(9).

Meredith, D. J. et al., 2012. Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. *Journal of Medical Engineering & Technology,* pp. 60-66.

Mietus, J. & Goldberger, A., 2014. *Basic Time and Frequency Domain Measures.* [Online]
Available at: https://physionet.org/tutorials/hrv-toolkit/
[Accessed 10 6 2017].

Moody, G. B., 2015. *WFDB Applications Guide.* [Online]
Available at: https://physionet.org/physiotools/wag/
[Accessed 02 06 2017].

Moody, G. B., 2016. *RR Intervals, Heart Rate, and HRV Howto.* [Online]
Available at: https://www.physionet.org/tutorials/hrv/
[Accessed 12 7 2017].

Moore, J., 2017. *Normative Elite HRV Scores by Age and Gender.* [Online]
Available at: https://elitehrv.com/normal-heart-rate-variability-age-gender
[Accessed 28 06 2017].

Murphy, K. P., 2012. *Machine Learning A Probabilistic Perspective.* Cambridge: The MIT Press.

NHS England, 2015. *The 'Internet of Things' is revolutionising healthcare – Jeroen Tas.* [Online]
Available at: https://www.england.nhs.uk/expo/2015/07/20/the-internet-of-things-is-

revolutionising-healthcare-jeroen-tas/

[Accessed 01 07 2017].

Olson, R. S. a. B. N. a. U. R. J. a. M. J. H., 2016. *TPOT GitHub.* [Online]

Available at: https://github.com/rhiever/tpot

[Accessed 29 06 2017].

Olson, R. S., Bartley, N., Urbanowicz, R. J. & Moore, J. H., 2016. *Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science.* New York, ACM, pp. 485-492.

Olson, R. S. & Moore, J. H., 2016. *TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning.* s.l., ICML 2016 AutoML Workshop.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S., 2007. A Design Science Research Methodology. *Journal of Management Information Systems*, 24(3), pp. 45-78.

Physionet, 2016. *PhysioBank Annotations.* [Online]

Available at: https://www.physionet.org/physiobank/annotations.shtml

[Accessed 14 6 2017].

Plews, D., Altini, M., Scott, B. & Laursen, P., 2017. Comparison of Heart Rate Variability Recording With Smart Phone Photoplethysmographic, Polar H7 Chest Strap and Electrocardiogram Method. *International Journal of Sports Physiology and Performance.*

R. Fisher, S. P. A. W. a. E. W., 2003. *Median Filter.* [Online]

Available at: https://homepages.inf.ed.ac.uk/rbf/HIPR2/median.htm

[Accessed 12 07 2017].

Rollin McCraty, P. & Shaffer, F., 2015. Heart Rate Variability: New Perspectives on Physiological Mechanisms, Assessment of Self-regulatory Capacity, and Health risk Read More: http://www.gahmj.com/doi/full/10.7453/gahmj.2014.073. *GLOBAL ADVANCES IN HEALTH AND MEDICINE,* 1, 4(1), pp. 46-61.

Scikit-learn, 2017. *scikit-learn user guide.* s.l.:scikit-learn .

SciPy, 2014. [Online]
Available at: https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.signal.medfilt.html
[Accessed 12 06 2017].

Shiffman, D., 2012. *The Nature of Code.* 5 ed. s.l.:Daniel Shiffman.

Singh, A., 2017. *Cardiogram blog.* [Online]
Available at: https://blog.cardiogr.am/applying-artificial-intelligence-in-medicine-our-early-results-78bfe7605d32
[Accessed 28 06 2017].

Sisson, M., 2014. *Mark's Daily Apple.* [Online]
Available at: http://www.marksdailyapple.com/have-you-checked-your-heart-rate-variability-lately/
[Accessed 27 07 2017].

Stables, J., 2017. *Wareable.* [Online]
Available at: https://www.wareable.com/smartwatches/best-smartwatch-2017
[Accessed 28 06 2017].

Subhadra Evans, L. C. S. J. C. T. K. C. L. L. K. Z. a. B. D. N., 2013. Heart rate variability as a biomarker for autonomic nervous system response differences between children with chronic pain and healthy control children. *Journal of Pain Research,* 11 06.pp. 450-457.

Taelman, J., Vandeput, S., Spaepen, A. & Huffel, S. V., 2008. *Influence of Mental Stress on Heart Rate and Heart Rate Variability.* s.l., Springer, pp. 1366-1369.

VanderPlas, J. T., 2017. *Understanding the Lomb-Scargle Periodogram,* s.l.: arXiv.org.

Varun Gulshan, P. et al., 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *American Medical Association.,* 13 12, 316(22), pp. 2402-2410.

Villarejo, M., Zapirain, B. G. & Zorrilla, A. M., 2012. A Stress Sensor Based on Galvanic Skin Response (GSR) Controlled by ZigBee. *Sensors,* 05, Volume 12, pp. 6075-6101.

Wapcaplet, Y., 2012. *File:Diagram of the human heart.svg.* [Online]
Available at: https://commons.wikimedia.org/wiki/File:Diagram_of_the_human_heart.svg
[Accessed 24 06 2017].

Zong, W. & Moody, G. B., 2015. *Physionet.* [Online]
Available at: https://www.physionet.org/physiotools/wag/wqrs-1.htm
[Accessed 14 06 2017].

APPENDIX

A.1. DATASET CONTENT

Key	Type	Size	Value
comments	list	0	[]
fs	float	1	15.5
signame	list	6	['ECG', 'EMG', 'foot GSR', 'hand GSR', 'HR', 'RESP']
units	list	6	['mV', 'mV', 'mV', 'mV', 'bpm', 'mV']

Figure 11 shows the content of the header file belonging to the first .dat file

-0.03	-0.0072	2.5	11.1	84	10.9
-0.026	-0.0033	2.51	11.1	84	11
-0.016	0.0019	2.51	11.1	84	11
-0.02	-0.003	2.52	11.1	84	11

Figure 12 shows content from the first .dat file

A.2. INITIAL DATA SUMMARY STATISTICS

	ECG	EMG	HR	RESP	footGSR	handGSR
count	1326089	999320	1078806	1326089	1326089	1173197
mean	-0.037272426	0.700132714	79.48785103	29.2603306	6.391841445	9.264014885
std	0.627476727	1.237060827	21.50102533	19.48633076	4.78416437	7.283106464
min	-10	-32.76	0	-19.669	0	-32.763
25%	-0.09	0.186	67	23.1	2.329	4.057
50%	-0.037	0.331	76	35.19	5.646	6.964
75%	0.05	0.85	88	42.33	8.62	13.524

max	9.952	32.61	391	67.7	26.758	32.762

Table 6 shows descriptive statistics for the dataframe with the original data

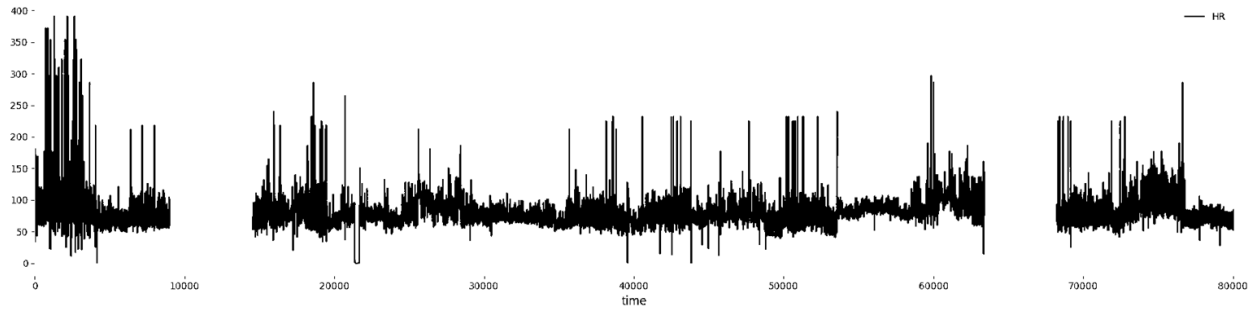


Figure 13 shows a time series of raw HR data before any cleaning has been done and indicates outliers and missing samples

A.3. RR STATISTICS FIGURES

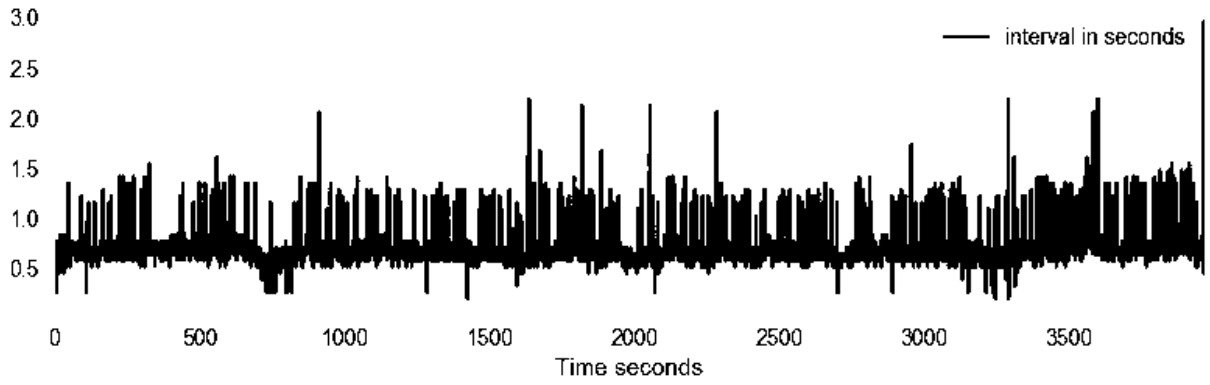


Figure 14 shows the raw RR intervals from the first driving session plotted as a time series

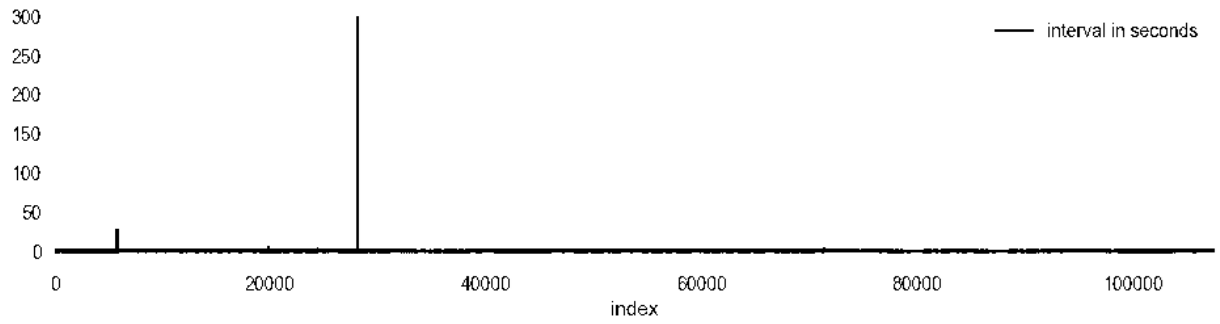


Figure 15 shows the raw RR intervals of the whole dataset plotted as a time series where an evident outlier is visible as an RR interval of almost 300 seconds

A.4. GALVANIC SKIN RESPONSE DATA WITH A MEDIAN FILTER

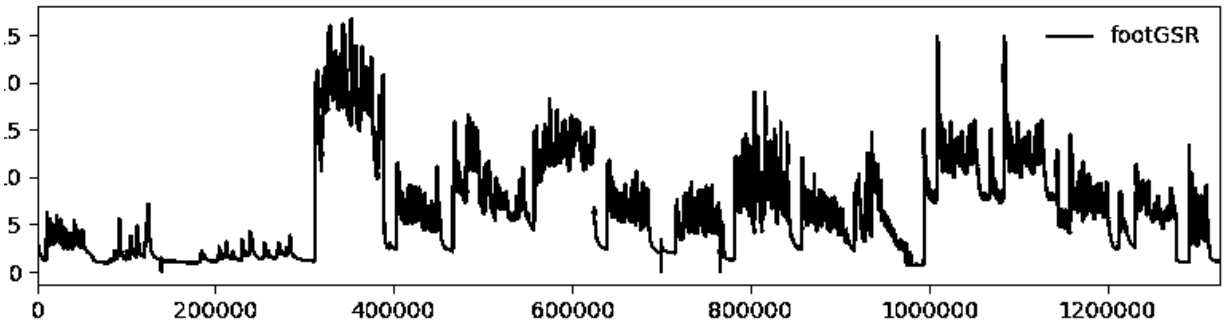


Figure 16 shows the GSR values from the foot after applying a median filter with a 13-sample sliding window

A.5. CLEANING THE RR INTERVALS

Table 7 shows the max and the min before and after applying a median filter

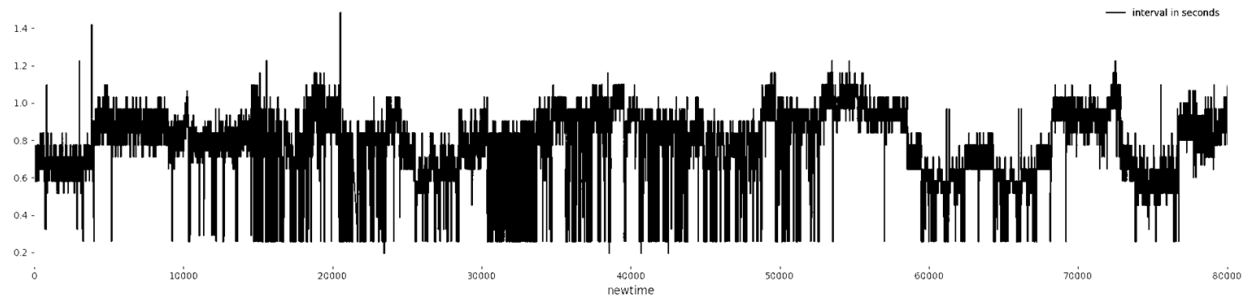


Figure 17 shows the RR intervals with a median filter alone

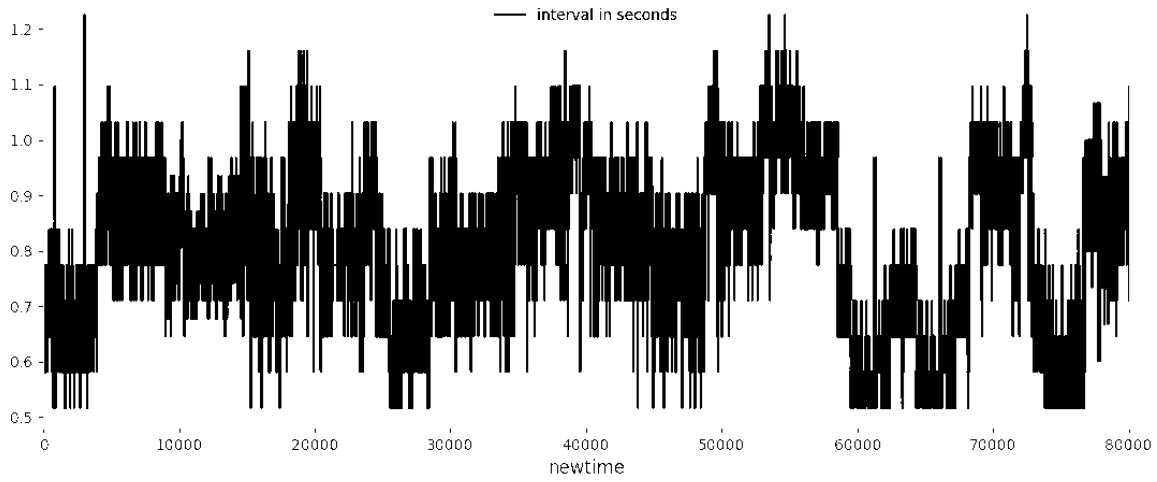


Figure 18 shows the RR intervals with a median filter applied after chopping off too low and too high values

	Max	Min
Before median filter	298.774	0.129
After median filter	1.484	0.194

Table 8 shows the max and min values in the dataset before and after doing the cleaning

A.6. GRAPHICAL PRESENTATION OF DATA COMBINATION

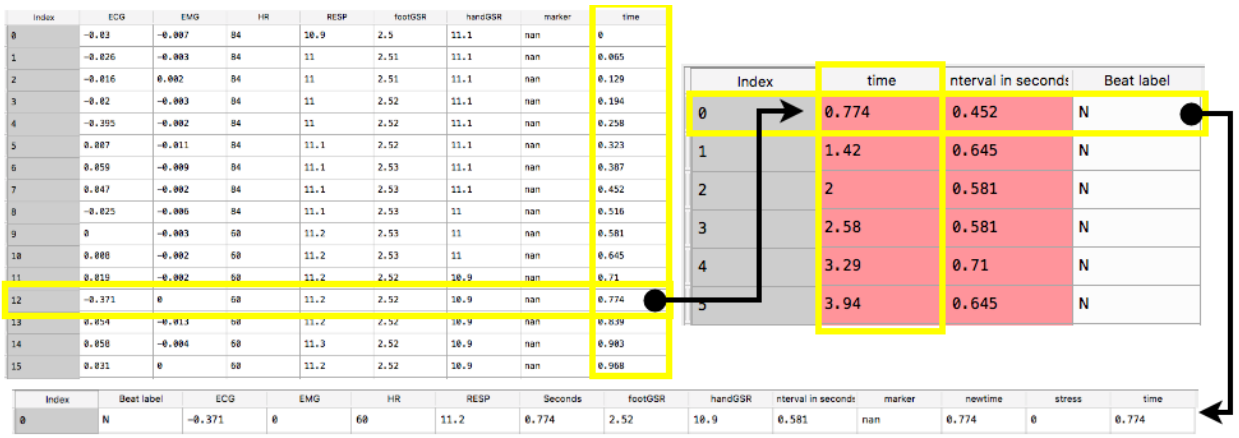


Figure 19 shows how the join process is taking place where each row in the ECG data is matched with an RR peak row based on the time column from both DataFrames.

Index	Beat label	ECG	EMG	HR	RESP	Seconds	footGSR	handGSR	nterval in seconds	marker	newtime	stress	time
0	N	-0.371	0	60	11.2	0.774	2.52	10.9	0.581	nan	0.774	0	0.774
1	N	0.004	-0.008	44	11.1	1.42	2.49	10.8	0.581	nan	1.42	0	1.42
2	N	-0.171	-0.013	44	10.9	2	2.46	10.6	0.645	nan	2	0	2
3	N	-0.243	0.002	44	10.5	2.58	2.43	10.6	0.645	nan	2.58	0	2.58
4	N	-0.15	-0.012	44	11	3.29	2.4	10.6	0.645	nan	3.29	0	3.29
5	N	-0.324	-0.001	44	11.5	3.94	2.39	10.6	0.645	nan	3.94	0	3.94
6	N	0.044	0.002	88	11.4	4.19	2.38	10.6	0.645	nan	4.19	0	4.19
7	N	-0.06	-0.012	88	11.1	4.77	2.37	10.6	0.645	nan	4.77	0	4.77
8	N	-0.14	-0.001	88	11.1	5.42	2.36	10.5	0.645	nan	5.42	0	5.42
9	N	-0.048	-0.008	88	11.3	6.13	2.34	10.6	0.645	nan	6.13	0	6.13
10	N	0.033	-0.004	33	11.1	6.84	2.34	10.9	0.71	nan	6.84	0	6.84

Figure 20 shows how the 10 first rows of data looks after the combination of the original data and the RR peaks using an inner join.

A.7. OVERVIEW OF HEART RATE VARIABILITY FEATURES

NNRR	AVNN	SDNN	RMSSD	pNN50	TP	ULF	VLF	LF	HF	LF_HF
0.974	0.617	0.0356	0.0152	0.0556	0.00124	0	0.000696	0.000407	0.000135	3
0.978	0.648	0.0135	0.0139	0.0455	0.000144	0	9.19e-06	5.97e-05	7.52e-05	0.794
0.979	0.645	2.24e-08	0	0	nan	0	nan	nan	nan	nan
0.979	0.645	2.24e-08	0	0	nan	0	nan	nan	nan	nan
0.979	0.645	2.24e-08	0	0	nan	0	nan	nan	nan	nan
0.979	0.645	2.24e-08	0	0	nan	0	nan	nan	nan	nan
0.979	0.645	2.24e-08	0	0	nan	0	nan	nan	nan	nan
0.978	0.655	0.0238	0.0098	0.0227	0.000548	0	0.000223	0.000292	3.32e-05	8.79
0.977	0.699	0.0243	0.01	0.0238	0.000538	0	0.000237	0.000271	2.97e-05	9.12
0.977	0.71	nan	0	0	nan	0	nan	nan	nan	nan
0.977	0.71	nan	0	0	nan	0	nan	nan	nan	nan
0.977	0.71	nan	0	0	nan	0	nan	nan	nan	nan

Figure 21 shows the heart rate variability measurements after extracting the heart rate variability features from the RR intervals.

Index	ECG	EMG	HR	RESP	Seconds	footGSR	handGSR	nterval in seconds	marker	newtime	stress	time
0	-0.00197	-0.00474	77.8	10.8	12.5	2.42	10.9	0.615	nan	12.5	0	12.5
0	0.00293	-0.00446	102	10.8	30.5	2.42	11.3	0.648	nan	30.5	0	30.5
0	0.00674	-0.00343	105	10.6	52.5	2.23	11.4	0.646	nan	52.5	0	52.5
0	-0.00404	-0.00253	87.7	10.6	74.4	2.17	11.5	0.645	nan	74.4	0	74.4
0	0.0127	-0.00443	88.8	10.7	96.2	2.02	11.1	0.645	nan	96.2	0	96.2
0	0.0278	-0.00496	91.4	10.7	118	2.15	12.2	0.645	nan	118	0	118
0	-0.0109	-0.00468	87.3	10.8	139	2.28	13.1	0.645	nan	139	0	139
0	-0.00522	-0.00413	79.5	11.1	160	2.31	13.8	0.655	nan	160	0	160
0	-0.0194	-0.00373	88	10.9	181	2.27	13.6	0.698	nan	181	0	181
0	-0.0654	-0.00241	89.9	11.2	202	2.09	12.4	0.71	nan	202	0	202
0	-0.000977	-0.00305	86.9	11.3	222	1.96	11.3	0.71	nan	222	0	222
0	-0.00373	-0.00407	85.3	11	243	1.86	10.3	0.71	nan	243	0	243

Figure 22 Shows the mean of original features after extracting the heart rate variability features from the RR intervals

A.8. FEATURE IMPORTANCE

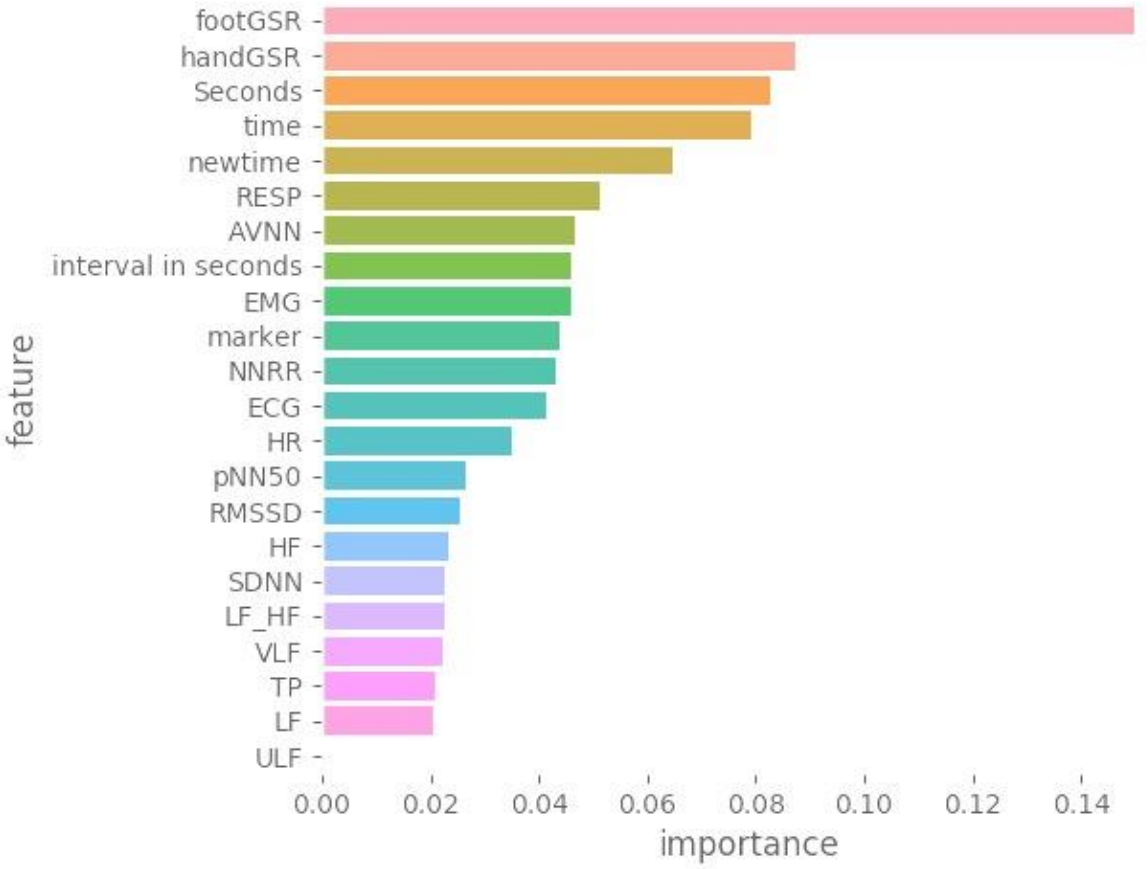


Figure 23 shows the feature importance of all the features in the dataset where it's strong evidence for data leakage as expected with the GSR values

A.9. FEATURES AFTER MANUAL FEATURE SELECTION

Table 9 shows the features used after feature selection

Features after manual feature removal
['HR', 'RESP','AVNN', 'interval in seconds', 'SDNN', 'RMSSD', 'pNN50', 'TP', 'LF', 'HF', 'LF_HF']

A.10. TPOT PIPELINE

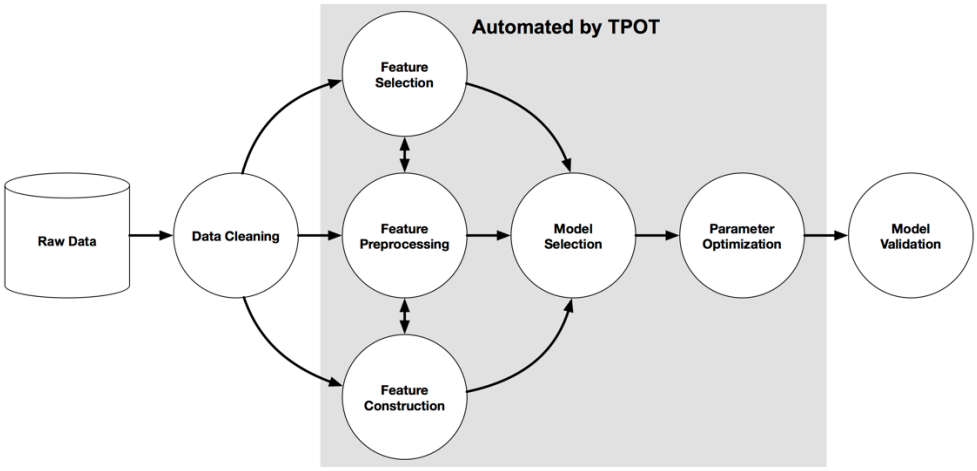


Figure 24 represents how TPOT optimises a machine learning problem with feature selection, feature preprocessing, feature construction, model selection and parameter optimisation. The figure is by (Olson, 2016)

A.11. MACHINE LEARNING ALGORITHMS

Table 10 shows the algorithms used from TPOT and Auto-Sklearn

Auto-Sklearn	TPOT
gaussian_nb	sklearn.naive_bayes.GaussianNB
decision_tree	sklearn.tree.DecisionTreeClassifier
extra_trees	sklearn.ensemble.ExtraTreesClassifier
random_forest	sklearn.ensemble.RandomForestClassifier
gradient_boosting	sklearn.ensemble.GradientBoostingClassifier
k_nearest_neighbors	sklearn.neighbors.KNeighborsClassifier

adaboost	sklearn.svm.LinearSVC sklearn.linear_model.LogisticRegression xgboost.XGBClassifier

A.12. Model parameters from test on wearable data

a. TPOT results

KNeighborsClassifier(FeatureAgglomeration(SelectPercentile(input_matrix, SelectPercentile__percentile=34), FeatureAgglomeration__affinity=manhattan, FeatureAgglomeration__linkage=complete), KNeighborsClassifier__n_neighbors=47, KNeighborsClassifier__p=1, KNeighborsClassifier__weights=distance)

0.802216066482