Predicting Cardiovascular Risk with Objective Physical Activity Measurements

van Dooren, B.J.T.A.

Award date:
2021

Department of Mathematics and Computer Science
Data Mining Research Group

Predicting Cardiovascular Risk with Objective Physical Activity Measurements

Master Thesis

B.J.T.A. van Dooren

Supervisors:

Prof. dr. Mykola Pechenizkiy (TU/e)
Rianne Schouten, MSc (TU/e)
Felipe Masculo, MSc (Philips)
Dr. Warner ten Kate (Philips)

Assessment Committee:

Prof. dr. Mykola Pechenizkiy (DM)
Felipe Masculo, MSc (Philips)
Dr. Warner ten Kate (Philips)
Prof. dr. Milan Petkovic (SEC)

Version 1.0

Eindhoven, July 27, 2021

Abstract

Cardiovascular disease (CVD) is the leading cause of death in the United States (US) and costs the US around 219 billion each year [93]. Hence, it is of significant interest for both the health industry and US public health to detect and treat people at risk for CVD [40]. One method to detect people at risk is with physical activity (PA). Regular exercise has been shown to associate with a lower CVD risk [11]. Wearable devices hosting accelerometers can track the intensity, frequency, and duration of a person's PA profile. In this study, we explore how one week of objective measured PA can be used to predict CVD risk. We use the National Health and Nutrition Examination Survey (NHANES) accelerometer data from the 2003-2006 cycles, with the Reynolds Risk Score (RRS) as the outcome variable to predict CVD risk.

The richness of objective PA, compared to self-reported PA, should help us to acquire a better understanding of a patient's physical health. This understanding enables us to identify those in need of care and minimize CVD risk in a more timely and precise manner. However, due to the size and noise of PA measurements, interpreting one week of objective PA is challenging for healthcare professionals. As a result, specialized methodologies must be applied to access and extract information from the objective PA data. Extracting information will be accomplished by expanding on the current baseline approach, the cut-point analysis, which uses pre-defined intensity thresholds to calculate the average time spent in various PA categories. Furthermore, time-series analysis and deep learning techniques will be used on the PA measurements.

When only PA data was utilized, we obtained good classification results using the baseline cut-point analysis features on four supervised learning techniques. Adding more features (e.g., bouts of activity and activity at specific times of the day) to this analysis allowed the classification, and regression results, to improve even further. Using more sophisticated techniques, such as MiniRocket and deep learning, improvements were obtained over the baseline technique. We conclude that MiniRocket provides consistent improvements over the baseline cut-point analysis technique when using only PA data. Furthermore, with the inclusion of non-modifiable risk factors (i.e., age, gender, and history of close relative having a heart attack before the age of 60) in our models, we discovered that PA features improve the predictive performance even when the non-modifiable risk factors are known. This improvement is especially noticeable in the age group 40-70. Lastly, we argue that the acquired predictive power of PA can be more valuable than well-established risk factors such as diabetes, total cholesterol, HDL-cholesterol, CRP, and smoking.

In conclusion, we show that CVD risk calculated with the RRS can be predicted using objective PA. The MiniRocket approach obtained the most additional predictive power compared to the baseline cut-point analysis when only PA data is used. The PA showed additional value alongside the non-modifiable risk factors. We even argue it contained more predictive power than well-known risk factors used to calculate the RRS in specific age groups.

Keywords: *physical activity, cut-point analysis, deep learning, MiniRocket, time series, cardiovascular risk*

Acknowledgments

In these challenging times during the COVID-19 pandemic, I am fortunate to have been given the opportunity to perform my thesis at Philips. As a leading health technology company, Philips aims to improve the lives of 2.5 billion people, a goal to which I hope to have contributed in some small way. I would like to thank Philips for the opportunity and valuable experience.

I would also like to thank my supervisor, Mykola Pechenizkiy, for his feedback and supervision throughout the thesis on a regular basis. Furthermore, I would like to give thanks to my supervisors at Philips, Felipe Masculo and Warner ten Kate for the excellent feedback, valuable ideas, and sharing of knowledge during the weekly meetings. Additionally, I would like to thank Rianne Schouten, who provided valuable feedback and suggestions on the project and the report. Finally, many thanks to my friends, colleagues, and family, whose encouragement and feedback helped me stay motivated.

Contents

List of Figures

List of Tables

Acronyms

ADA	AdaBoostRegressor
AI	Artificial Intelligence
BA-diff	Bland-Altman-diff
BOSS	Bag-of-SFA-Symbols
BS	Brier-Score
CNN	Convolutional Neural Network
CRP	C-Reactive Protein
CVD	CardioVascular Disease
DEC	Decision Tree
DTW	Dynamic Time Warping
FCN	Fully Convolution Network
FPR	False Positive Rate
FRS	Framingham Risk Score
GRU	Gated Recurrent Unit
HDL	High-Density Liporotein
HIVE-COTE	Hierarchical Vote Collective of Transformation-based Ensembles
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MiniRocket	MINImally RandOm Convolutional KErnel Transform
MVPA	Moderate-Vigorous Physical Activity
NHANES	National Health and Nutrition Examination Survey
NMF	Non-Modifiable Factors
OSA	Obstructive Sleep Apnea
PA	Physical Activity
PR	Precision-Recall
RFC	Random Forest Classifier
RFR	Random Forest Regression
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RRS	Reynolds Risk Score
SFA	Symbolic Fourier Approximation
SMOTE	Synthetic Minority Oversampling TEchnique
SOTA	State-Of-The-Art
SVC	Support Vector Classifier
TPR	True Positive Rate
TSC	Time Series Classification
US	United States

Chapter 1

Introduction

The medical sector is continuously pursuing the newest technological advances to make the best decisions for a patient based on all available scientific data. This phenomenon is also referred to as the concept of Evidence-Based Medicine [87]. This concept aims to integrate clinical experience and patient values with the best available research information. Due to recent technological advancements, we can now provide additional scientific knowledge to pursue the technological advances in the medical sector.

Physical Activity (PA) is one of those areas where with current advances in technology, we can track the intensity, frequency, and duration of a patient using accelerometers. These devices have allowed the Centers for Disease Control and Prevention to create an objectively measured PA dataset. This data is made available in the National Health and Nutrition Examination Survey (NHANES) [51], [53].

A particular interest in studying the benefits of PA is disease prevention. The ability to predict a person's health status can significantly reduce a person's risk of developing a disease. Preventing a disease is often preferred to treating it since it provides for a higher quality of life and lower healthcare costs. As a result, objective PA has piqued the interest of many researchers due to its high correlation to several health outcomes, allowing for preventive measurements.

CardioVascular Disease (CVD) is one of these health outcomes, which is the leading cause of death for people in the US [93]. CVD also costs the US around $ 219 billion each year [40] and is thus of major interest for both the healthcare industry and US public health to detect and treat people at risk for CVD in time. PA can aid in the diagnosis and prevention of CVD in patients at high risk for CVD. According to Benjamin et al. [11], this is attributed to PA enhancing cardiovascular function.

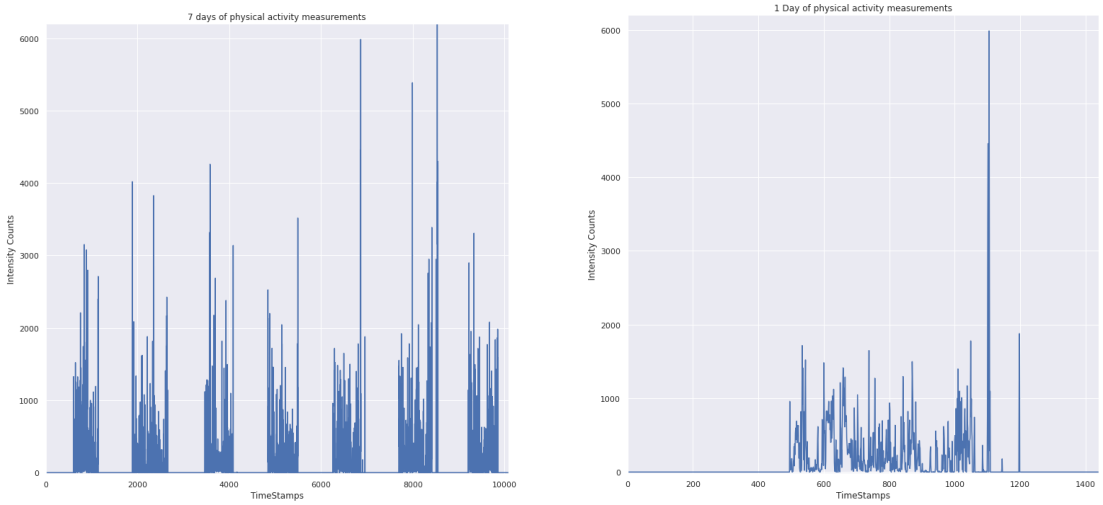
Despite the apparent correlation between PA and health benefits, recent findings from National Health and Nutrition Examination Survey 1999–2002 indicate that there has been no significant improvement in PA over the last decade in the US [91]. Seeing no improvements is concerning since there is clear evidence of PA's health advantages regarding CVD risk. Given the many benefits, many healthcare professionals also recommend certain PA practices based on the self-reported PA of patients. Advising PA practices is also supported by the initiative of 'exercise is medicine' [103].

Troiano et al. [114], however, showed that objective PA data adheres substantially lower to PA recommendations compared to self-report. Hence, there are noticeable differences between a person's self-reported PA, the believed level of PA, and the actual level of PA conducted. The objective PA thus allows us to better understand a patient's physical status compared to the self-reported status currently used. In addition to knowing if a patient adheres to the PA recommendations, objective measurements can show us different harmful patterns related to CVD risk. For instance, sitting for long periods without moving has been recognized as a health concern related to factors as hypertension, according to Lakerveld et al. [69], which is a risk factor for

CVD. Such a pattern could be better recognized and detected in an objective PA dataset compared to self-reported PA patterns. Hence, using objective PA measurements provides us with better information about patient's PA patterns, allowing healthcare professionals to detect people at high risk for CVD earlier if clinically employed.

Objective PA data consists of one intensity value per minute (i.e. the sum of activity, based on acceleration, for one minute) for seven consecutive days, resulting in 10080 intensity values for one participant. An example of one objective PA time series is shown in Figures 1.1a and 1.1b, for seven days and one day of PA, respectively. There are, however, a few difficulties when analyzing PA data. Even with knowledge of fundamental patterns in the data, obtaining useful information from the data is a tremendous challenge for healthcare professionals. Due to the data being recorded with a frequency of one minute, we obtain a very long, high-dimensional time series, which can be regarded as very 'spiky' due to the summation of the activity over one-minute intervals. We will refer to this as noise. Additionally, the data consists of periods of nonwear, where the device was presumably not worn, introducing significant areas of zero values or unreliable areas in the data. Hence, we must utilize specialized methods to extract information from this dataset. The current approach used, which will be discussed in Section 2.1, does not capture all of the information that can be obtained from this data. As mentioned by Ainsworth et al. [89], it uses a crude categorization of a patient's activity status and can also overlook the unique patterns of PA on an individual level.

Given the recent technological advancements, we believe that additional predictive power can be obtained from this dataset by extending the current approach or applying State-Of-The-Art (SOTA) techniques. These techniques include Artificial Intelligence (AI) as well as time series analysis approaches. SOTA techniques have been demonstrated to provide solutions in numerous domains, such as more accurately detecting cancer in the early stages or assisting people to stay healthy due to AI in the internet of medical things [99]. SOTA techniques could be the answer to dealing with this massive volume of high-dimensional data. This problem is what we are going to explore in this study.



(a) 7 days of physical activity data

(b) Physical activity zoomed in on day 5

Figure 1.1: Objective physical activity as a time series.

1.1 Problem Statement

In this section, we will explain the primary goal of this work. Before explaining the primary goal, we will go over the semantics required to formally explain the problem and understand the sub-goals of this work.

1.1.1 Problem Semantics

Cardiovascular Disease (CVD) Risk As CVD risk, we have chosen to use the Reynolds Risk Score (RRS) [104][105], which defines the probability, from 0 to 100%, of developing a heart attack, stroke, angioplasty, coronary artery bypass surgery, or death related to heart disease in the upcoming ten years. Why a risk score is used instead of a binary health outcome and why the RRS has been used compared to other risk scores will be explained in Section 2.4.

Physical Activity Measurements The objective physical activity measurements can be seen as a univariate time series. A univariate time series is an ordered set of values over time, in our case, the amount of activity a person conducts every minute, recorded by consecutive timestamps. A formal definition is presented in Definition 1.

Definition 1. ***A univariate time series*** $X = \{x_1, x_2, ..., x_n\}$ *is an ordered set of values, often measured at successive moments in time separated by uniform time intervals. The length of time series X is equal to n.*

The univariate PA consists of 10080 timestamps with intensity values, where some timestamps are missing or can be seen as unreliable. The missing and unreliable data will be further explained in Section 3.1. In addition to the intensity values, the PA data also consists of steps, resulting in a multivariate time series with more than one observed value per timestamp. A formal definition is presented in Definition 2.

Definition 2. ***A multivariate time series*** $Y = \{X^1, X^2, ..., X^m\}$ *is a set of* m *univariate time series* $X^i \in Y$ *with typically the same timestamps. Each univariate time series* $X^i \in Y$ *typically has the same length* n*, resulting in* m *values per timestamp for* n *timestamps.*

Nonwear Nonwear refers to timestamps x_i, in our time series X, where the accelerometer was supposedly not worn. The participants were instructed to remove the accelerometer during sleep and water-related activities. We cannot differentiate between those activities and possible other periods where the device was removed, for example, due to discomfort. Therefore, we are unsure about the level of activity that was conducted during those timestamps. We refer to those timestamps as nonwear, which are seen as unreliable points in the data. More details about how these timestamps are identified will be explained in Section 3.1.3.

Time Series Classification (TSC) In this work, we try to predict a label Y_i, which can either be a participant's risk or a label consisting of them being low or high-risk. This task consists of our entire univariate PA dataset D, where $D = (X_1, Y_1), (X_2, Y_2), ..., (X_k, Y_n)$ and where X_i is representative of one univariate time series, with an accompanied label Y_i. Given D, we want to obtain a model Z that correctly predicts outcome Y_i, given X_i. In other words, $Z : X \longrightarrow Y$. Hence, TSC consist of training a model on a dataset, either with machine learning, deep learning, or any other technique, such that we can correctly predict a label given an unseen time series.

1.1.2 Research Goals

As previously mentioned, predicting a person's CVD risk is of great interest to healthcare professionals. Predicting CVD risk allows healthcare professionals to identify people who need to adjust their lifestyle or need treatment to reduce their CVD risk based on their PA status. Reducing CVD risk will not only lower global healthcare costs but will also improve the US global health. The main objective that this work tries to address is the following problem:

- ***Can we predict CVD risk based on one week of univariate objectively measured physical activity data?***

To measure how well we can predict a person's CVD risk, we will measure the performance in a classification and regression task. The classification task separates participants into low and high-risk groups based on a clinical threshold of 10 %. Justification for this threshold is given in Section 2.4. This clinical threshold allows healthcare professionals to immediately identify high-risk people while compromising the ability to determine their true risk. In the regression task, we will attempt to predict a person's RRS score, allowing healthcare professionals to handpick people at risk. This strategy could be more efficient because healthcare professionals will have more information about the actual CVD risk and may speed up when treatment is administered. After both tasks are completed, a comparison in performance can be made to determine whether it is better to apply a threshold first or after a continuous regression task.

Closely connected to our main problem is extracting the most predictive power from the PA data. Without a proper technique, the model will fail to make good predictions for CVD risk since it cannot extract information out of the data. Therefore, this work also tries to address the following main question:

- ***Which technique can obtain the most predictive power out of one week of objective physical activity data?***

The current method used on the PA data is frequently regarded as a crude categorization of the activity status [89]. By extending the current method and applying more sophisticated techniques, we hope to obtain more predictive power out of the PA data. The current method(s) used by researchers will be explained in Section 2.1. Being able to obtain more predictive power out of the PA is crucial for better prediction models and better selection of people in need of treatment. Besides these two main goals, we have several sub-goals which we want to investigate. After our literature study, these goals and interests came to light, where we discovered several areas of interest. These areas will be discussed in the paragraphs below.

Wear Time As previously explained in Section 1.1.1, nonwear areas are areas where we are uncertain about the activity conducted during a period. Wear flags will flag unreliable parts of the time series and these areas will then not be used in the cut-point analysis, which is explained in Section 2.1. Hence, we are interested in the effects of these areas on the performance of predicting CVD risk. We are also interested in a feature referred to as normalized features, allowing the features to become less correlated to wear time. Lastly, we are interested in the imputation of the nonwear areas, which should replace nonwear areas with the actual activity that was conducted. For this section, we aim at answering the following questions:

- *What is the effect of nonwear periods on our performance of predicting CVD risk?*
- *What is the effect of normalized features, compared to counts, on our performance of predicting CVD risk?*
- *What is the effect of imputation on nonwear areas?*

Value of PA After developing a prediction model, we need to evaluate the predictive power of PA. Is PA a good predictor of CVD risk and is it useful in conjunction with other variables? Age and gender are known to be associated with PA and CVD risk, and it is thus interesting to observe the predictive power of PA alongside those variables. The following questions will address these points:

- *What is the predictive power of objective PA independently of age and gender?*
- *What does the best model provide for healthcare professionals?*

Goals of Literature Review Based on our research questions and the challenges of the data, we will specify several goals which we want to answer from our literature review. First, to know how we can obtain the most predictive power out of the PA data, we need to investigate how researchers address the challenges of PA data and what is currently lacking from these approaches. This knowledge allows us to think more carefully on how to extend the current techniques used and what other SOTA techniques require to solve the PA challenges. Moreover, observing what techniques have been utilized on other or similar datasets allows us to determine which techniques could obtain the most predictive power out of the PA data. To address these points, we define the following literature goals:

- *What techniques have been used to process objective physical activity data, which will be reviewed in Section 2.1.*
- *What techniques have been used to process time-series data, and which technique(s) seems most promising for physical activity data. This will be reviewed in Section 2.2.*

Furthermore, we discuss why CVD is chosen as an outcome variable compared to other health outcomes and why CVD is the most interesting given PA data. Additionally, we give argumentation for using the RRS compared to other risk scores and why using a risk score is beneficial for answering our research questions. The following literature goals will address these points:

- *The decision for choosing CVD as health outcome will be reviewed in Section 2.3.*
- *Why a risk score is used and why the RRS specifically will be reviewed in Section 2.4.*

1.2 Outline

This thesis is divided into several chapters. It continues with Chapter 2, which consists of the background information and literature review. It will discuss the current techniques used on this dataset and other methods that we could use. Additionally, we explain why we use a risk score, why the RRS is used, and give more details about the RRS. In Chapter 3, we discuss the dataset in more detail and discuss how it should be processed. Additionally, we provide an analysis of the data and define the experimental setup. In Chapter 4, we describe and discuss the experiments and results on the classification task. In Chapter 5, we discuss the results of the models on the regression task. Furthermore, we relate the classification task to the regression task. In Chapter 6, we discuss an example how our proposed model could be used in a CVD risk management program. Lastly, in Chapter 7, we conclude this work by summarizing, describing the contributions, and providing suggestions for future work on this topic.

Chapter 2

Background and Literature Review

This section will explore the current techniques used on our dataset and explores several methods used on other time-series datasets, which could be applied to our PA dataset. Furthermore, we discuss different health outcomes and discuss why CVD is most interesting. Lastly, we discuss the RRS score and explain why the RRS is used as an outcome variable.

2.1 Current Physical Activity Processing Techniques

As the NHANES dataset has been available for a long time, many researchers have attempted to use this dataset to obtain interesting findings. Researchers have used both the objective and subjective PA data with different methods. Due to many different researchers working on this dataset, different ideas and methods to utilize the data have been proposed. One of these techniques, which is most often used, is called a cut-point analysis.

Cut-Point Analysis A cut-point analysis processes the objectively measured PA with the usage of intensity thresholds, as displayed in Table 2.1. These thresholds are obtained from calibration studies, that relate accelerometer counts to energy expenditure [14], [73], [125]. These intensity thresholds represent different types of activities. A sedentary activity represents an activity where the energy expenditure is low, hence sitting or lying down. Light activities involve an activity that only needs a small amount of energy expenditure, such as cooking or ironing. Lifestyle activities include strolling, doing household activities, and biking languidly. Moderate activities involve activities such as brisk walking, off-road biking, and weight training. Lastly, vigorous activities are activities we generally can only partake in for a limited time, such as running and aerobics. The latter two intensity levels are often combined into one activity type, given that moderate activity is often seen as the most important intensity, together with vigorous activity. Hence, the latter two activity types are often referred to as Moderate-Vigorous Physical Activity (MVPA).

Activity Type	**Intensity Counts**
Sedentary	$0-100$
Light	$100-759$
Lifestyle	$760-2019$
Moderate	$2020-5998$
Vigorous	≥ 5999

Table 2.1: Physical activity intensity thresholds per activity type.

These thresholds split the PA time series into five different areas, each representative of an

activity type. Each activity type is representative of an energy expenditure level. This split can be observed in Figure 2.1, which displays the threshold applied on one day of PA data. Hence, the cut-point analysis splits the time series into five different areas and counts the number of points in each activity group. These counts give us a general idea of a person's PA profile.

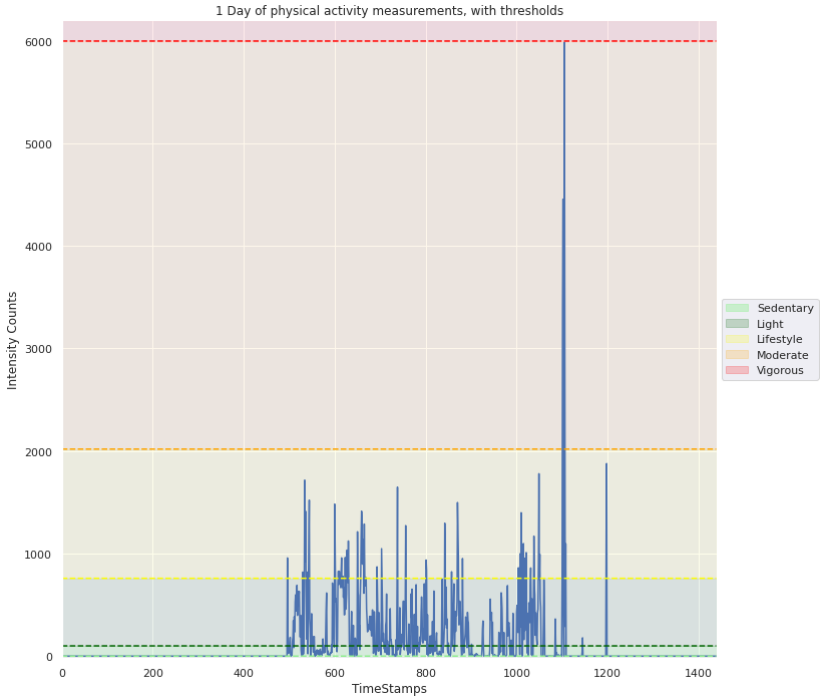


Figure 2.1: Intensity areas for one day of physical activity.

Only counting the number of minutes in each intensity group is a crude way of describing the PA patterns of a person [89]. Additional features could be of importance, leading to more information of a person's PA status. However, providing these features is more complex and is often not done. Some researchers, however, have applied more advanced features, one of which is moderate bouts of at least ten minutes. This feature is based on the PA guidelines, which suggest that bouts of ten minutes are beneficial for our health [97]. However, at this point, as argued by Ayabe et al. [7], no evidence is available to support what the best length for a bout is.

Additionally, it is argued that with the usage of accelerometer data, interruptions in the bouts can be seen as a reasonable assumption, as suggested by Masse et al. [88]. It is reasoned that if we are jogging or doing an intense activity for several periods of five minutes and rest for one minute in between those periods, no bouts would be recorded. Therefore, we also find it more reasonable to allow for allowance intervals of one minute in bouts. Additionally, the PA guidelines suggested that any bouts of activity should be beneficial for our health; hence, allowing for interruptions in bouts seems more than reasonable [97].

We also observed Niemela et al. [94] using features in the cut-point analysis that differentiate the activity per week and weekend. Furthermore, we have observed features consisting of the maximum sedentary or moderate activity that was conducted on a day, as used by Boiarskaia et al. [13]. No clear indication is given why this feature is used; however, we argue that it could display how evenly the activity is spread out over the week. Suppose a person has only conducted moderate activity on one particular day. In that case, the features about the average moderate activity and maximum activity on a day should allow us to observe how evenly spread-out the activity was over the week. As also suggested by the physical activity guidelines [97], spreading the activity over at least three days could lead to additional health benefits.

Mean Activity and Adherence to PA Recommendations Another technique used to utilize the PA data is simply calculating the mean intensity value. Furthermore, researchers have also used an estimate of adherence to PA recommendations related to different health outcomes [114], [81]. The physical activity and sedentary guidelines recommend engaging in at least 150 minutes of moderate activity or 75 minutes of vigorous activity per week (or equivalent combination), as well as sitting less, and for shorter periods [16], [97]. We know from other analyses that these guidelines are often not reached; specifically, only 8 % of the participants in the NHANES data reach these requirements, as denoted by Troiano et al. [114].

Deep Learning Pyrkov et al. [100] also developed an age prediction model that could predict a patient's biological age. The intention was to observe if this helped to predict when the model expected someone to die. It used a follow-up study of ten years after the PA measurements to determine who died. This follow-up study only consists of all-cause deaths data. The study proposed to use a Convolutional Neural Network (CNN), a deep learning technique, used on the raw PA measurements. For a detailed explanation of what deep learning entails, we refer to Chapter 2 of the book of Josh Patterson et al. [95]. Pyrkov et al. showed that this method outperformed a principal component analysis score and a regression model in predicting a participant's biological age. However, it remains unclear how advanced the features used are compared to cut-point analysis and thus how much gain the CNN model gained compared to the cut-point analysis.

2.2 TSC Techniques

Provided that there are several ways to solve TSC, we will investigate whether we should use machine learning, deep learning, or specialized algorithms to solve TSC for PA data in this section. As argued by Keogh et al. [63], an efficient and effective representation of a time series is key to the successful discovery of time-related patterns. In recent years, many TSC approaches have been proposed where the majority of these approaches are evaluated on the UCR archive [29]. This archive primarily consists of time series datasets that are short, noiseless, and without intervals of zero values. These datasets have been heavily prepossessed and do not accurately represent the structure of a real-world dataset. As a result, when analyzing TSC approaches, we must keep in mind that data from the UCR archive must be evaluated with caution. The remainder of this section is presented as follows. First, we examine how specialized TSC algorithms handle time series data, followed by a discussion of deep learning models. Lastly, we summarize the results and specify which method(s) will be used in this work.

2.2.1 Algorithms for TSC

Numerous algorithms are successful in TSC, which have obtained good accuracies on the UCR archive [29], [78]. Many of these algorithms use different strategies to solve TSC. However, there is one algorithm that has been considered by many as SOTA [9], [39], [78], [108]. This algorithm is the Hierarchical Vote Collective of Transformation-based Ensembles (HIVE-COTE) method, as presented by Bagnall et al. [9]. As stated by Shifaz et al. [108], the HIVE-COTE is a meta ensemble algorithm for TSC, which is considered as SOTA for non-machine learning models. HIVE-COTE incorporates several classifiers, each of which extracts information about a particular feature from a time series. This feature can be either in the time domain, frequency domain, or summary of intervals within the time series. By applying all these classifiers, HIVE-COTE is capable of dealing with noisy data [78]. The usage of different classifiers and the detection of a wide variety of features could be essential for our dataset. However, HIVE-COTE is unsuitable for many applications due to its high training time [39]. HIVE-COTE has a training time complexity of $O(T^2 * N^4)$, with a dataset of T time series, where each time series has a length N [39], [108]. Hence, we will need to focus on different techniques given the length of our time series, which causes an exponential training time.

2.2.2 Shape-Based Methods

Shape-based methods utilize a distance function, which computes the similarity between two or more time series. A time series is then classified based on its nearest neighbor in the training data. We will discuss the most noteworthy methods from this class and how applicable they are to the PA data.

Euclidean Distance The Euclidean distance is a well-known method, which has proven to be an effective and straightforward method for TSC [121]. This method compares two time series by calculating the distance between them. This approach, although proven to be successful in many different applications, seems unsuitable given our data. As argued by Fulcher et al. [41], this metric should mainly be used for highly comparative time series. Euclidean distance could put too much emphasis on the zero regions and is very sensitive to minor distortions, as argued by Keogh et al. [64]. Given the zero areas and noise in PA data, it would make more sense to compare it in a time-warped manner, which compares time series more flexibly. This method is known as the Dynamic Time Warping method.

Dynamic Time Warping Dynamic Time Warping (DTW) has been a well-known concept for many years, having been introduced by Berndt et al. [12] in 1994. Although it is an older method, it is still capable of outperforming numerous TSC methods [77]. Lines et al. [77] compared several distance measures and showed that no distance measure outperforms DTW. However, given our dataset, DTW may struggle to discover comparable alignments in two sequences because a feature in one sequence could be lower, higher, or in a different location than another sequence, as argued by Keogh et al. [65]. Additionally, DTW does not address the noise in the dataset, which may lead to an overemphasis on the spikes in the data if correctly aligned. For these reasons, even though DTW is regarded as one of the most SOTA distance metrics, distance metrics seem to lack what we require in our solution. From this evaluation, we can conclude that shape-based methods are not the solutions to our problem. Therefore, we will look into structure-based methods in the following sub-section.

2.2.3 Structure-Based Methods

Structure-based methods make use of features in the time series. These methods try to detect features in the data that could be representative of a class. There are various alternations available of these methods, which we will discuss in the following subsections.

Shapelets Shapelets are sub-sequences of a time series, which are maximally representative of a class, as discussed by Ye et al. [124]. The idea of shapelets is that it tries to detect the most representative features of a class. Instead of comparing time series against each other, we search for shapelets in each time series representative of their class. As a result, shapelets are more noise resistant and could completely ignore the zero values since they search for class-related features. An example of different features of two leaves is displayed in Figure 2.2, where the shapelets detect that urtica leaves have more of an angle to the stem compared to verbena leaves.

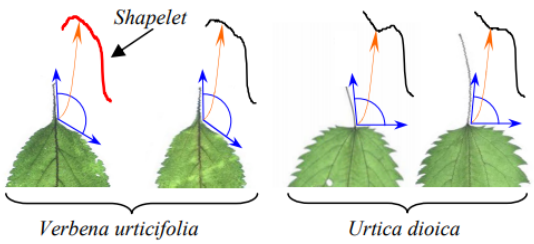


Figure 2.2: Shapelets for Verbena and Urtica leaves. Image taken from [124].

The main difficulty is findings these shapelets, as it proves to be very computationally expensive, as argued by Raktanmanon et al. [101]. It is hard to know which shapelets to use, especially in a naive way. We would have to test each shapelet on many time series, which is very expensive. Additionally, shapelets focus on one feature of the class, ignoring the big picture of the time series. A different take on shapelets is called motifs, which we will discuss in the upcoming section.

Motifs Compared to shapelets, motifs try to find a pattern that repeats in the time series several times, as argued by Mueen et al. [92]. These repeating patterns indicate the most important patterns in the data. A motif is the most similar pattern that can be found in a time series, which is repeated the most amount of times [92]. This method will thus totally disregard a global trend in the data. It would, however, be crucial not to take the zero intervals as a repeating feature. Furthermore, a repeating pattern seems complicated to observe at first sight, given the spikiness of the data. However, the concept has proven to be very successful and it can handle noisy data. Although only using motifs seems not applicable in our case, a specific feature-based algorithm has shown good results, which we will discuss in the upcoming paragraph.

Bag-of-SFA-Symbols A Bag-of-SFA-Symbols (BOSS) model represents a time series as a set of substructures with Symbolic Fourier Approximation (SFA) words. SFA words are similar to the features of a dataset, where one can also compare SFA words to a set of shapelets. According to Schafer et al. [106], it has a level of noise reduction by setting an appropriate SFA alphabet size. Choosing the correct SFA alphabet and window size can, however, be a difficult task. Lastly, it provides in-variance to phase shifts, offsets, amplitudes, and occlusions. These claims seem to make it an applicable model for PA data.

An example of the BOSS model is depicted in Figure 2.3. As can be observed, we obtain a histogram of an extensive list of different features of our time series. Although it seems like an applicable model, it is known for its scalability concerns, limiting its applicability given our large time series [90]. We also observe that the scalable version, cBOSS, performs worse than the upcoming methods [32]. As a result, while it may be intriguing to test, we believe the upcoming methods have more potential than the BOSS model. Given that all previously mentioned methods seem inappropriate or have remaining challenges that limit their applicability, we will look into if deep learning methods could solve our problem, as Pyrkov et al. [100] already demonstrated the potential of these techniques on PA data.

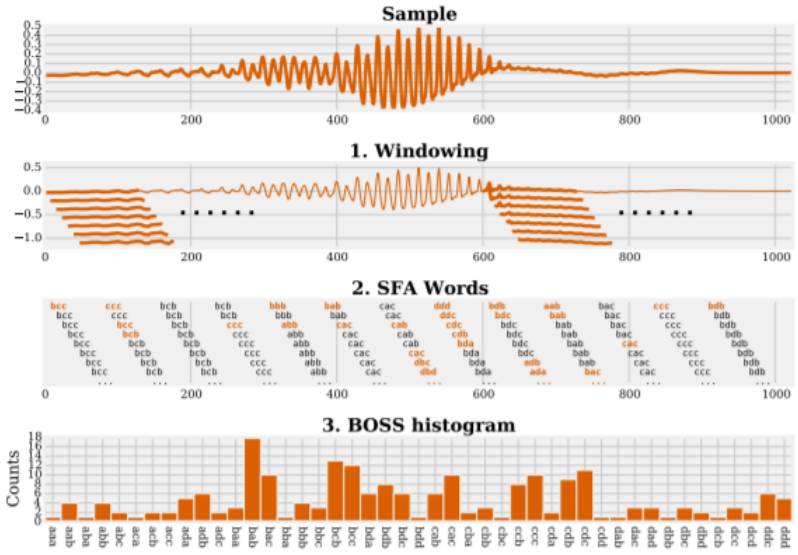


Figure 2.3: Workflow of BOSS model. Image taken from [106].

2.2.4 Deep Learning for TSC

As we have seen in the previous sections, many recent TSC techniques are too computationally expensive or inapplicable for our dataset. Therefore, we consider deep learning models, as they are becoming an emerging field in TSC, as shown by Fawaz et al. [38]. We focus on deep learning models, since with high dimensionality and very high feature correlation, simple machine learning methods generally do not work for TSC due to its simplicity, as argued by Keogh et al. [64].

Deep learning models have shown remarkable performance on a variety of time series analysis tasks, as argue by Wen et al. [122]. For instance, deep learning models have been used in image recognition tasks and achieved human-level performance [68]. Besides image recognition, deep learning models have also proven to obtain good accuracies and results in word-embedding problems and document classification, as argued by Mikolov et al. [71]. These models can learn deep hierarchical patterns in different datasets, which could be important for our dataset. Given the rise and promises of deep learning methods, we are to believe those models could surpass the performance of HIVE-cote, as specific deep learning models are already very close to the performance of HIVE-COTE, as shown by Fawaz et al. [38]. Given that deep learning could be the future for TSC, we will further investigate which models are most promising in TSC for our data. Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) are commonly used in TSC, as argued by Guan et al. [48]. These models are considered SOTA methods due to their ability to find deep features in the data. These in-depth features may help the model ignore the zero intervals and noise and only focus on the essential elements of the PA data. Therefore, we believe that they could be the solution to our problem. We will discuss these two methods in the upcoming chapters.

2.2.5 Convolutional Neural Networks

Convolutional Neural Networks (CNN) can extract features from signals and achieve promising results in image classification, as previously argued by Krizhevsky et al. [68]. We will provide a few variants of CNN that could potentially provide significant improvements compared to 'simple' CNN models. A problem with CNN models is that for deeper layers, the quality of the model degrades. This is due to the vanishing gradient problem, as explained by Hochreiter et al. [56]. The deep layers in deep learning architectures are capable of gaining in-depth features, but instead, in CNN it can degrade the quality of the model, as argued by Wen et al. [122]. However, we are interested in deep layers for our problem as these layers show the most potential in capturing in-depth features previously unknown, capable of ignoring the noise, and intervals of zero values.

As argued by Wang et al. [120], a current challenge in deep learning models is their learning abilities and more shallow models should be considered, as they could be more efficient. A shallow model, a model with few layers, could prevent the vanishing gradient problem. Instead of using shallow models, which are typically unable to capture in-depth features, we examine residual networks, which use residual connections to solve the vanishing gradient problem. These residual connections allow the depth of the model to increase without causing the exploding gradient problem. The decision to review residual networks is also based on the success of these networks in time series, as shown by Fawaz et al. [39]. The concept of residuals connections and a commonly used residual network will be discussed in the next paragraph.

ResNet Fawaz et al. [38] performed an exhaustive comparison between several popular deep learning models for TSC, such as an Fully Convolution Network (FCN), encoder, t-LeNet, MLP, MCNNN, and ResNet. Fawaz et al. discovered that for TSC, ResNet outperforms all other proposed models on most occasions. The FCN was the only model that could contest ResNet in some cases. The ResNet model used by Fawaz et al. is the proposed ResNet network from Wang et al. [121]. Wang et al. also argued that ResNet could improve the interpretability of the model and learn highly complex patterns from the data due to the shortcut connection. Moreover, models such as FCN tend to overfit more often, especially on larger datasets. Therefore, we classify the

ResNet model of Wang et al. [121] as a good candidate network for our problem. In addition, the FCN model, used by Fawaz et al. [38], also seems a good candidate, given that it can contest the ResNet model on specific datasets.

The ResNet model proposed by Wang et al. [121] makes use of residual blocks. In deep learning networks, a layer feeds its output into the next layer directly. This flow is also present in residual blocks; however, residual blocks also use a shortcut connection. An example of such a residual block is depicted in Figure 2.4.

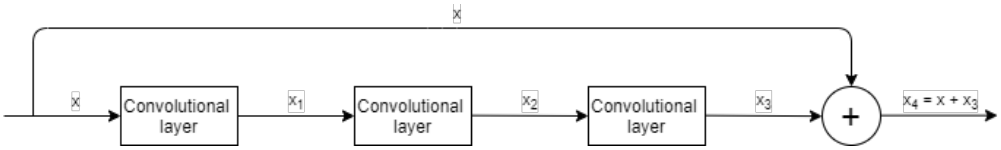


Figure 2.4: Shortcut connection in residual block.

In the residual block, the output "x" goes through all three convolutional layers and then comes out of the last convolutional layer as x_3. Then together with the shortcut connection, it gets summed up as x_4. This method allows gradients to flow through the network directly instead of flowing through all the layers. This method has shown to deliver SOTA performance in the 2015 ILSVRC challenge, as stated by Targ et al. [112]. The complete network proposed by Wang et al. in [121] is depicted in Figure 2.5. Han et al. [49] showed that the ResNet model could capture deep structures in long time series due to their deep convolutional network structure. These deep layers could help to detect patterns in our long PA time series and ignore the noise and intervals of zero values. However, the major downside of such a model is that it has a complex architecture and could make it more difficult to understand how the model makes its decisions.

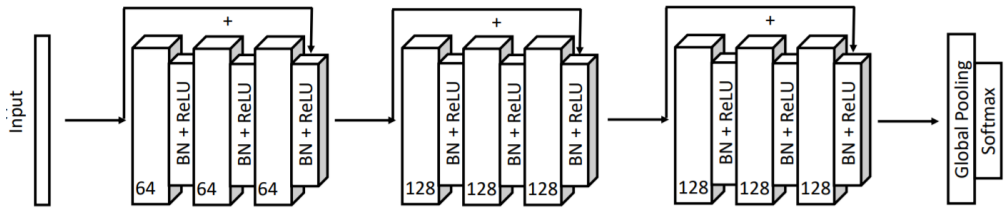


Figure 2.5: ResNet network. Image taken from [121].

InceptionTime In addition to ResNet, Fawaz et al. [39] proposed the InceptionTime model just recently at the time of writing. This model is inspired by finding the equivalent for the 'AlexNet' model but then applicable for TSC. It is also inspired by the ResNet network, as it also makes use of residual connections. It has demonstrated that it does not only achieve SOTA accuracy, but it is also more scalable and quicker than existing SOTA models such as HIVE-COTE [39]. The InceptionTime model uses multiple inception layers, also called inception modules, which allows the network to extract features from both short and long time series. An inception module is illustrated in Figure 2.6.

This model can capture both short and long patterns due to the different kernel sizes (the amount of timestamps a neuron sees at once) of the inception module, allowing for various receptive fields. This inception module could be of great importance in our dataset, as it could enable the model to capture smaller features, such as bouts, and large features, such as sitting for long periods. The inception module consists of the following layers:

- One-dimensional convolutional layers, which obtain the output of the bottleneck layer. The kernel sizes are 10, 20, and 40, with stride one, for the three convolutional layers.
- A bottleneck layer which reduced the dimensionality of the input. This layer helps to reduce the computation time and speeds up training. It has kernel size one and stride one.
- A max pooling layer of pooling size three and stride one.
- Concatenation layer, where all the convolutional layers and the bottleneck come together.

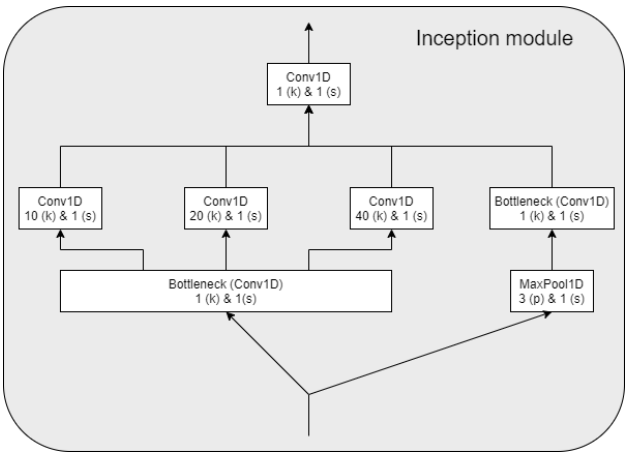


Figure 2.6: Inception module with kernel size (k), stride (s), and pooling size (p).

The complete network makes use of two residual blocks, with each three inception modules. These blocks work as residual blocks, as explained in Section 2.2.5. The entire Inception network is displayed in Figure 2.7.

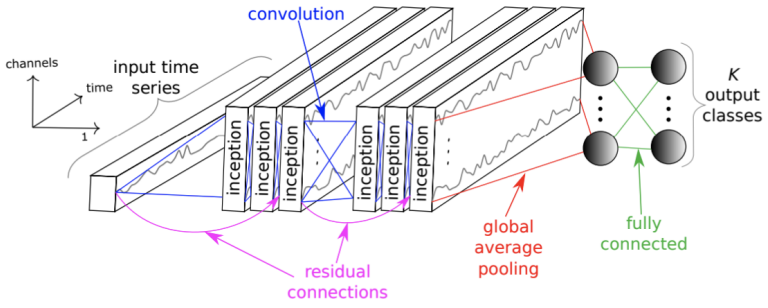


Figure 2.7: InceptionTime network. Image taken from [72].

As argued by Fawaz et al. [39], this model further builds on ResNet, as it tries to utilize residual connections. Additionally, it uses multiple convolutional layers with different kernel sizes to enable various receptive fields. Fawaz et al. [39] also made a comparison between the performance of ResNet and InceptionTime, as ResNet seems most promising from his previous comparison in [38]. This comparison showed that InceptionTime has a ratio of 53/7/25 against ResNet, with 53 wins, 7 ties, and 25 losses. Based on these findings, we will attempt to use both models. It seems that ResNet is better than InceptionTime in some cases; however, in most cases, InceptionTime outperforms the ResNet model.

MiniRocket As previously discussed, non-deep learning TSC models that obtain SOTA accuracies generally have high computational complexity, such as HIVE-COTE. This complexity makes those models unusable for our dataset. Also, specific algorithms focus on either shapes or structures. However, Dempster et al. [31] proposed a new method that builds further on the success of convolutional neural networks for TSC, which can focus on both types of features, namely shapes and structures. This model uses a method that transforms the time series using 'random' convolutional kernels and using the transformed features to train a linear classifier. This strategy is called Rocket.

At the time of writing, Dempster et al. [32] also proposed the MINImally RandOm Convolutional KErnel Transform (MiniRocket) approach, which improves on the computational complexity of the original Rocket method. As discussed by Dempster et al. [32], it also provides a slight improvement upon the performance compared to the original method. It reduces the number of kernels used and can be regarded as almost deterministic. Only how the bias is selected from the kernels can be considered as minimally random. From the mean rank performance from MiniRocket, compared to InceptionTime, it is significantly better. Additionally, we believe that the dilation kernels and non-dilated kernels can results in interesting features, given that we have seven consecutive days of PA data. In the next paragraph, we go into more detail about the functioning of the MiniRocket approach.

MiniRocket Workings We will give a short explanation of how the MiniRocket method works. However, we also advise reading the paper of Dempster et al. [32] which goes into even further detail of the working of this method. During the explanation, we advise referring to Figure 2.8, as this should help visualize the inner workings of this method.

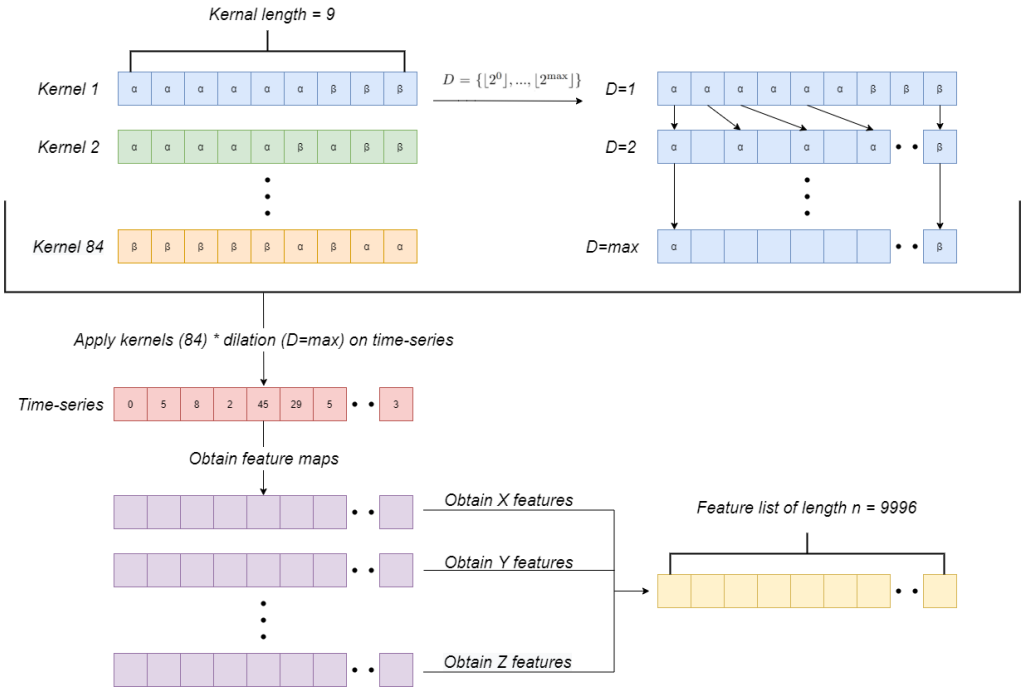


Figure 2.8: MiniRocket structure for feature extraction.

In MiniRocket, we specify the number of features we want to obtain from a time series. However, the author argues that 10000 features, 9996 specifically, obtains optimal performance. More features will not improve the performance but only increase the complexity. As previously explained, MiniRocket was inspired by convolutional networks, which uses kernels to find interesting patterns in the data. MiniRocket also uses kernels, however, not with multiple layers as is done in

a convolutional network. MiniRocket uses 84 deterministic kernels, all of length 9, which can also be seen in Figure 2.8. Each cell in a kernel can have either a weight of -1 or 2, which is predefined. The scale of these weights is unimportant because the bias is drawn from the convolution output and thus matches the input scale. After these kernels, we apply dilation, which 'spreads' a kernel over the input. For example, a dilation of two causes the model's weights to be convolved with every second element of a time series. This dilation of two also causes the kernel to double in length. This can also be seen in Figure 2.8. The number of dilations that can be applied will be totally dependant on the length of the time series. Dilations are interesting, given that in this manner, we can find features based on frequency.

Now that the kernels are defined, we will apply them to our time series to obtain a set of feature maps, as displayed in Figure 2.8. From each feature map, we will obtain a different amount of features. We want to obtain a total of 9996 features; however, the number of feature maps will be less than the number of features we want to obtain. Hence, the method scales the number of features it obtains based on the dilation of that specific feature map. Smaller dilations will obtain exponentially more features than larger dilations. This scaling ensures that the 9996 features are obtained. Each feature is calculated by the proportion of positive values of that feature map - a bias value. This bias value is drawn from quantiles of the convolution output for a randomly selected training example. For a low dilation value we will need multiple features per feature map. Multiple bias values are then drawn from the convolution output to obtain multiple features. This bias causes the method not to be deterministic due to being drawn from a random training example. The feature will always be between 0 and 1. Ultimately, we receive a large set of features that explain the time series based on structure and shapes. As the author advises, a ridge classifier is used to train a model and make predictions.

2.2.6 Recurrent Neural Networks

Similar to a CNN, a Recurrent Neural Network (RNN) has also shown great potential regarding TSC, as argued by Guan et al. [48]. RNNs are a class of networks similar to CNN; however, in an RNN, we want to know our previous inputs before evaluating our next input. Knowing previous inputs is very important in sequential data, which is data where the order matters. One of the most promising models used is the Long Short-Term Memory (LSTM) network.

LSTM As argued by Guan et al. [48] and Wen et al. [122], LSTM is one of the primary techniques used in human activity recognition and have been destined to be used to analyze sequential data. LSTM models use several gates, including multiple hidden units, which all facilitate error back-propagation for very complex features. For a more broad explanation of LSTM models, we refer to the explanation of Karim et al. [62]. Karim gives an explanation of the usage and concepts of LSTM models in regards to time series classification.

As argued by Karim et al. [62], LSTM models are an improvement over general RNN, which suffer from the vanishing gradient problem. LSTM generally learn temporal dependencies in sequences over a short period; however, they have difficulties finding long-term dependencies in long sequences. Given that LSTM focuses on short sequences, it could be that these models focus too much on local patterns instead of the bigger picture. Additionally, although it has been SOTA in human activity recognition, it is essential to recognize that such a dataset is recorded with a higher frequency than PA data. PA data is summed over one minute, making it impossible to deduct the exact activity that was conducted; hence, there is a clear difference in the use case. However, we do not rule out LSTM models, given their strong performance on sequential and sensor data. Therefore, we will implement an LSTM model to observe the performance. Additionally, very similar to LSTM cells are Gated Recurrent Unit (GRU) cells; which have shown to be better in some instances, as argued by Gruber et al. [46]. GRU cells use different gates compared to the LSTM cells, which allows them to need less computing power and remember the data differently. Therefore, we will also implement a GRU model, alongside the LSTM model, given the similarities and success of these models.

2.2.7 Most Applicable TSC Techniques

Based on our literature review, we observe a variety of approaches to tackle the PA data. We reasoned that shape-based methods, also referred to as distance-based methods, are not the most suitable approach to tackle our problem. Techniques such as Euclidean distance or DTW will focus too much on the noise and intervals of zeros values in PA data. Given this knowledge, we looked at structure-based methods, or also referred to as feature-based methods. We evaluated the concept of shapelets, as introduced by Ye et al. [124]. Shapelets may emphasize small features; together with the complexity, shapelets appear to be not the solution to our problem. Having evaluated non-machine learning-based methods, we conduct that some show potential, such as BOSS. However, we reasoned that deep learning methods, especially CNNs, seem to have more potential. CNN models are previously used in a variety of TSC tasks and have been shown to succeed [38]. Therefore, we will evaluate a simple CNN, an FCN, ResNet, and InceptionTime model as our CNN models. These models have shown to be very competitive against each other and perform very well on the UCR archive. Additionally, we will experiment with the MiniRocket approach, given its success on various datasets compared to popular techniques such as ResNet and InceptionTime. Lastly, although an RNN seems less applicable to our dataset, we will experiment with an LSTM and GRU model to observe how well these models work on our dataset. The specifics of the architectures used is defined in Appendix F.2.

2.3 Health Outcomes

This section summarizes how PA data has been used in relation to several health outcomes and why CVD is the most promising health outcome. A more complete and detailed analysis is available in Appendix Section C. Additionally, we denote several areas where improvements could be made on the cut-point analysis based on findings from other studies.

PA has been used in correlation to sleep apnea, where it was recognized that the time of exercise is connected to sleep apnea, hypertension, and CVD [36]. Hence incorporating the time of day exercise is conducted can be a valuable feature in a model, which is currently overlooked and not used. However, we conclude that sleep apnea classifies poorly as a health outcome due to the low prevalence in the data and often going unrecognized [33], [81]. Moreover, for respiratory, we observe a lack of variables available in the NHANES dataset. Although researchers tried to link PA to severe asthma in other studies [28], this seems less applicable for the NHANES dataset. Moreover, mobility limitations have also been investigated with the cut-point analysis [82]. However, the relationship between mobility and PA is unclear as it can be an artifact of mobility issues. Hence, predicting mobility issues appears difficult given the bias correlating between mobility issues and PA. Furthermore, no significant results have been concluded when relation PA to either diabetes [83], [37] or cancer [113]. Given that no relationships seem to be present, we argue that another health outcome may be better to choose. Lastly, several researchers [13], [36], [83] used the cut-point analysis to relate PA to CVD. It showed that there was a clear relationship between PA and CVD. Moreover, we also observe that with CVD, we can use a risk score, which provides additional benefits compared to binary health outcomes. The other described health outcomes have missing variables for their respective risk scores. Hence, we conclude that CVD seems to be the most promising health outcome and can be used in combination with a risk score, which we will describe in more detail in the next section.

2.4 Risk Score

We opt to use a risk score compared to a binary health outcome due to a lower bias. When someone is identified with a particular illness, it is likely a healthcare professional prescribes a change in lifestyle to prevent the illness from worsening, or even as a cure for the illness long term [103]. This lifestyle change often includes advice to stop smoking, a healthy diet, or advising participants to

exercise more often [98]. Due to these recommendations, patients are more encouraged to change their lifestyle than people who did not get these recommendations. Hence, it can create a bias in the dataset, that before the illness occurred, a participant was very sedentary; however, after an illness occurred, due to advice from a healthcare professional, they became more active. Given that we cannot account for this bias, we reason that a risk score, a score that defines the percentage chance of something occurring, can be more stable in relating PA patterns to the risk of a health outcome. A risk score is a good screening tool and often easy to use in predicting and preventing a disease from occurring [23]. Risk scores have been used in many other illnesses besides CVD, including the health outcomes previously discussed, such as sleep apnea [23], type 2 diabetes [76], cancer [107], and asthma [1]. However, for the other illnesses besides CVD, we observe that the relationship between PA and health outcomes is not always evident and that variables are missing to calculate the risk. Hence, we conclude that a risk score provides numerous benefits over binary health outcomes and reduces the dataset's bias.

Risk Score to Use One of the most frequently used cardiac risk prediction models in clinical practice in the US is the Framingham Risk Score (FRS) [123] and RRS [104], [105], as argued by Klingenberg et al. [67]. The FRS, as stated by Asija Zaciragic [126], started in the 1980s; however, in recent years, it has shown its limitations due to new knowledge obtained about the classical risk factors. Additionally, there was substantial evidence on new risk factors that could lead to a better prediction model for cardiovascular risk prediction, as was shown by Ridker et al. [104], [105]. Multiple other models have been proposed to be used in everyday clinical practice for cardiovascular risk assessment. The models include ATP 3 [47] (produced by Framingham), ASSIGN, QRISK, QRISK2, SCORE, and RRS. It was, however, shown that care had to be taken by using a risk score that was validated and produced on a dataset from a different region [27]. The ASSIGN, QRISK, QRISK2, and SCORE have been made with a population of European populations. Given that the NHANES dataset contains information about participants in the United States, these risk scores are thus better to not use.

Cook et al. [26] performed a comparison between the ATP 3, Framingham CVD, and RRS for Global CVD risk prediction in multi ethnic women. It showed that the Framingham and ATP 3 model overestimated the risk for CVD compared to the RRS. The RRS was better calibrated compared to the other two risk scores. The RRS also showed significantly improved fit compared to either of the Framingham models. Additionally, DeFilippis et al. [30] tested the calibration and discrimination factor of the different risk scores and concluded that especially for males, the RRS performed best. RRS overestimated the risk score by a small margin, while the other risk scores overestimated significantly more. For women, the RRS underestimated women slightly, while the other risk scores overestimated risk. The other risk score overestimated percentage-wise more than the RRS underestimated. We, therefore, argue that the RRS seems to perform best.

The Reynolds Risk Score was validated and developed using data from 24,558 initially healthy American women [104]. After ten years, these women were followed up to observe if they developed a heart attack, stroke, angioplasty (balloon surgery to open an artery), coronary artery bypass surgery, or death related to heart disease. The Reynolds Risk Score for men was similarly developed using data from 10,724 initially healthy non-diabetic American men who were followed up after ten years for the development of a heart attack, stroke, angioplasty, bypass surgery, or death related to heart disease [105]. Ridker et al. [104], [105], who developed these risk scores, argued that high sensitivity C-Reactive Protein (CRP) and family history of having a heart attack independently associate with future cardiovascular events and have not been incorporated in current risk predictions models. Therefore, Ridker et al. investigated the usage of those variables and compared the test characteristics of global model fit, discrimination, calibration, and reclassification in prediction models for incident cardiovascular events. Compared to the traditional model, including CRP and family history resulted in a better model fit, a superior Bayes information criterion, and a larger C-index for both men and women.

Reynolds Risk Score The Reynolds Risk Score (RRS) tries to predict the risk, from a scale of 0 to 100, of one of the following events happening in the next ten years: heart attack, stroke, angioplasty, bypass surgery, or death related to heart disease. As previously stated, it is a data-driven risk score and thus uses a complex formula, as displayed in Equations 2.1 and 2.2, for men [105] and women [104], respectively.

$$\begin{aligned} &RRS_{men} = [1 - 0.8990^{(exp[B - 33.097])}] * 100 \text{ where} \\ &B = 4.385 * ln(age) + 2.607 * ln(\text{systolic blood pressure (mmHg)}) + 0.102 * ln(\text{CRP (mg/L)}) + \\ &0.963 * ln(\text{total cholesterol (mg/dL)}) - 0.772 * ln(\text{HDL-cholesterol (mg/dL)}) + 0.405\,(\text{if smoker}) + \\ &0.541\,(\text{if family history of premature myocardial infarction}) \end{aligned} \tag{2.1}$$

$$\begin{aligned} &RRS_{woman} = [1 - 0.98634^{(exp[B - 22.325])}] * 100 \text{ where} \\ &B = 0.0799 * age + 3.137 * ln(\text{systolic blood pressure (mmHg)}) + 0.180 * ln(\text{CRP (mg/L)}) + \\ &1.382 * ln(\text{total cholesterol (mg/dL)}) - 1.172 * ln(\text{HDL-cholesterol (mg/dL)}) + 0.818\,(\text{if smoker}) + \\ &0.438\,(\text{if family history of premature myocardial infarction}) + 0.134 * (\text{hemoglobin } A_{1c}\,(\%))\,(\text{if diabetic}) \end{aligned} \tag{2.2}$$

From the equations, we observe that the High-Density Liporotein (HDL) cholesterol is the only positive effect on the risk score and that diabetes is only accounted for in females. Moreover, due to the variables used in the studies by Ridker [104], [105], we have to make sure we do not include very large or unrealistic variables. It is also commonly known that certain variables already give a good indication of indicating being high risk. We have set several limits for specific biomarkers, which can be seen in Table 2.2. These limits are made based on the study of Ridker.

Bio-markers	**Min**	**Max**
CRP	0	5
Systolic blood pressure	105	200
Total cholesterol	140	400
HDL-cholesterol	30	150

Table 2.2: Value range for biomarkers in Reynolds Risk Score.

Classification Task The risk score predicts a probability score, between 0 and 100 %, that calculated the risk of having a CVD risk in the upcoming ten years. These probabilities are often grouped in categories, for clinical utility, based on U.S. clinical treatment guidelines [35], [80], [44], as argued by Cook et al [25]. These categories are made as follows, $<$ 5 %, which is seen as very low risk, where the risk is greater than the benefits when applying statin therapy or lifestyle changes. Statin therapy is a cholesterol-lowering medicine, which is seen as a high-risk factor for CVD. Applying a drug can have negative side effects and often the health benefits are not worth it for this risk group. 5 till 10 % is seen as low risk and minimal benefit of statin therapy compared to risk and cost of therapy in preventing CVD events in the next ten years. 10 till 20 % is seen as moderate risk, where a discussion of lifestyle modification and initiation of statin therapy is advised. Above 20 % is seen as a strong recommendation for statin therapy. These classification thresholds are also used across different risk scores, such as ATP 3 [80]. We have chosen to use a 10 % threshold for or classification task as this is commonly seen as the threshold for adjusting a patient's lifestyle or starting to apply statin therapy. Furthermore, treatment of hypertension is often combined with CVD risk, with a 10 % threshold, to determine what kind of approach will be taken [6], which is also supported by the American Heart Association [19]. Hence, after defining if someone has more than 10 % risk, it can be informative to check their blood pressure and cholesterol levels to observe if statin therapy, hypertension treatment, or lifestyle changes should be advised.

Chapter 3

Data Description, Analysis and Setup

This chapter will discuss the data in more detail and explain how missing data and nonwear periods are handled. Then we discuss the dataset's inclusion and exclusion criteria, and lastly, we analyze the dataset, define our experiment setup, and explain our metrics in more detail.

3.1 Dataset Description

The dataset discussed in this work is made publicly available by the Center for Disease Control and Prevention. The dataset is referred to as the National Health and Nutrition Examination Survey (NHANES) [51], [53]. The NHANES dataset uses a multistage probability sampling design, which allowed it to recruit a representative sample of the total civilian non-institutionalized population [3]. This strategy made it possible to collect a broad collection of data, subdivided into two-year cycles and categorized into six categories namely; demographics, dietary, examination, laboratory, questionnaire, and limited access information. NHANES is one of the largest and most diverse in terms of accessibility of data and one of the first studies to make a dataset public containing PA data measured by accelerometers [74]. The NHANES dataset also includes biomarkers and health outcomes in other files, which allows us to establish a connection between PA and health outcomes. The following sections will describe the data in more detail and go over the dataset's inclusion and exclusion criteria.

3.1.1 Accelerometer Data

The NHANES has made available PA data for four-year cycles, namely 2003-2004, 2005-2006, 2011-2012, and 2013-2014. The NHANES examined participants aged six and above who received a accelerometer to be worn for seven consecutive days. Subjects who used wheelchairs or had other impairments that prevented them from walking or wearing the device were excluded. It is important to note that the cycles 2003-2004 and 2005-2006 can be combined into one dataset. The same is also possible for the cycles 2011-2012 and 2013-2014. It is, however, not possible to combine one of the first two cycles with one of the last two cycles. The last two cycles make use of a different measurement unit, which can not be converted to the measuring unit of the first two cycles. It is also impossible to convert the measuring unit of the first two cycles into the measuring unit of the last two cycles. Additionally, there are vital differences in device placement and rules, which makes a comparison not possible. We have chosen to make use of the first two cycles and disregard the last two cycles. We explain the structure of the last two cycles and give our reasoning for opting to use for the first two cycles compared to the last two cycles in the Appendix B. We suggest reading up to and including Section 3.1.2 and then read Appendix B for more details on the last two cycles.

Cycles of 2003 until 2006 The participants were asked to wear an actigraph (Model 7164; Actigraph, LLC; Ft. Walton Beach, FL) on their right hip on an elasticized belt for a total of seven days [51]. An actigraph is also often referred to as an accelerometer. The participants were asked to take off the accelerometer when swimming or bathing, as the accelerometer was not waterproof. The participants also had to take off the accelerometer when going to sleep. However, as can be later conducted from the data, this is not always the case. Participants sometimes forgot to take off their accelerometer or even forgot to put it on. The accelerometer can only measure uniaxial acceleration in the vertical direction, making it difficult to detect upper body movements correctly. The measurements are recorded with a frequency of 10 hertz and then summed up for each minute [60]. Hence, we cannot see direct activity patterns, such as walking or jogging, because the activity is summed up for each minute. After the accelerometers were returned to the contractor, the data was downloaded. The device was then checked to determine if it was still within the manufacturer's calibration specifications, using an actigraph calibrator. If the device was not in calibration, the timestamps that were not calibrated for that participant would be flagged with a calibration flag. Besides the calibration flag, there is also a reliability flag, indicating how reliable that minute of the data is. These flags are created by several data scientists, who have cleaned and inspected the reliability of the data.

2003-2004 Cycle A total of 7176 participants were partaking in the first cycle, namely 2003-2004. Of the 7176 participants, 346 who returned their accelerometer were not in calibration on return and are excluded from the analysis. Additionally, 25 participants on top of the 346 had unreliable data based on the reliability flags and were also removed from the analysis.

2005-2006 Cycle A total of 7455 participants were partaking in the second cycle, namely 2005-2006. Of the 7455 participants, 369 who returned their accelerometer were not in calibration on return and are excluded from the analysis. Additionally, 223 participants on top of the 369 had unreliable data based on the reliability flags and are also removed from the analysis.

Variables for 2003-2006 Cycles In Table 3.1, a list of variables is given which are present in both cycles of the dataset, unless stated otherwise. Moreover, a visual interpretation of the structure of the data, including the headers from Table 3.1, is displayed in Figure 3.1.

Header	Description
SEQN	A unique subject identifier
PAXSTAT	Data reliability flag
PAXCAL	Device calibration flag
PAXDAY	Day of the week
PAXN	Sequential observation number
PAXHOUR	Hour of the day
PAXMINUT	Minute of the hour
PAXINTEN	Intensity value
PAXSTEP*	Device step count

Table 3.1: Data headers, where * is not available for the 2003-2004 cycle.

3.1.2 Categories of Data

Besides the PA data, there is a large variety of other information available in the NHANES dataset. This information can be found in their respective categorized areas, which are available by simple questions or laboratory results. All questions or laboratory results are only available at one point in time, meaning no follow-up results are present. This structure is important because this means that we cannot observe the PA's effect over time. Hence, the data is not in a longitudinal format. The

Unique Identifier	Reliability Flag	Calibration Flag	Time Identifier				Intensity Value	Step Counter
SEQN	PAXSTAT	PAXCAL	PAXDAY	PAXHOUR	PAXMINUT	PAXN	PAXINTEN	PAXSTEP*
31130	1	1	1	0	0	1	0	2
31130	1	1	1	0	1	2	1259	0
31130	1	1	1	0	2	3	612	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31130	1	2	7	23	59	1440	52	1
31131	1	2	2	0	0	1	25	0
31131	1	2	2	0	2	3	50	0

Figure 3.1: Dataset structure, where * is not available for 2003-2004 cycle.

questions and laboratory results are scattered across different files, and hence, have to be carefully merged. For example, information about the systolic blood pressure level is available in the blood pressure file, while information about smoking is available in the smoking file. Additionally, many questions or laboratory results are unanswered or unavailable for certain participants. This can be due to a skip pattern, which we will explain in the next section, or a participant who refused to answer or partake in a test. Therefore, it is essential to make sure participants in our analysis have all the required data available, given that many questions are unanswered.

The laboratory results range from a wide variety of different tests that the participants took. These tests sometimes have been conducted multiple times because previous ones have failed or are taken for more reliable results. Therefore, it is essential to check if multiple readings are available for a laboratory test, such that the results are combined. If this is not done correctly, for example, only the first reading for systolic blood pressure is taken, we would remove a large set of participants from our analysis simply because we only look at the first reading. We often observe that if a laboratory result is not present for participants, the likelihood of other laboratory results not being available is relatively high. It is also important to denote that we often see that laboratory results are more often unavailable compared to questionnaire results.

The disadvantage of having information in different files is that it must be combined based on their unique sequence identifier. There are also many rules we have to follow to make sure files are combined correctly. These rules and correctly combining files results in a steep learning curve to find and combine the required data for an analysis. We hope to alleviate this curve by making several functions available for Philips and providing important details about the dataset. These functions are available in Appendix A.

3.1.3 Processing the Data

In this subsection, we discuss how the data is processed and discuss this analysis's inclusion & exclusion criteria. Additionally, we discuss how missing and nonwear data is handled.

Skip Patterns The NHANES dataset contains many skip patterns depending on the question asked and the answer that was given. If these patterns are not recognized correctly, only a small portion of the population will be included in the analysis instead of the entire study population. For instance, if we were to analyze current cigarette smoking status among adults aged 18 years and older, we would have to use the variable SMQ040: 'Do you now smoke cigarettes?' That seems simple enough; however, it is crucial to recognize that a question asked before this question introduces a skip pattern. Question SMQ020 asks: 'Have you smoked at least 100 cigarettes in your entire life'. Only participants who answered 'yes' to this question were asked question SMQ040. Thus, participants who answered 'no' to question SMQ020 have not answered question SMQ040 and are thus missing in the dataset. However, it is important to recognize that they can be classified as non-smokers, given the answer on the previous question. Therefore, we have to be careful when processing the data; otherwise, valuable participants are lost in our analysis.

Moreover, we could also introduce a bias, given that the percentage of smokers would be wrongly represented. If only question SMQ040 was taken into account, we already only consider people who have smoked 100 cigarettes and thus obtain a dataset with a very biased smoker percentage. For an even more detailed explanation, including the significance, we advise following the data tutorial introduced by NHANES [3]. This concept is also often not addressed in other analyses or research papers, which makes it unclear to the reader if they have carefully handled the data or are only taken a small part of the data and possibly created a bias in their analysis.

Inclusion & Exclusion Criteria

The NHANES is a study designed to assess adult's and children's health and nutritional status in the United States. This design thus means that this dataset is limited to participants from the United States. Furthermore, NHANES recommends combining data into at least four years due to low prevalence. Hence, 2003-2006 cycles are combined into one dataset. Moreover, the dataset includes participants between 6 and 85 years of age. However, in this analysis, we will only use participants between 30 and 85 of age. Ridker et al. [104], [105] studies were limited to adults over the age of 45; however, we assume that extrapolating the risk scores to people up to 30 is reasonable. We observed that other CVD risk scores, such as the Framingham Risk Score (FRS) [123], also apply to people from 30 onwards. By comparing the risk under the age of 45 between the RRS and FRS, we observe no significant differences, and thus reason that it is reasonable to extrapolate the RRS to people up to 30. Moreover, participants below the age of 45 are generally at lower risk and are thus mainly used for additional data. We also believe that additional data could benefit our models, given the small sample size of the data, especially in deep learning techniques, which need large quantities of data to train their parameters.

Required Labels To fill in the formulas, as presented in Section 2.4, we have to obtain the correct information from the NHANES dataset. Therefore, to make explicitly clear which variables we have used to calculate the RRS, we have specified these variables in Table 3.2. These variables are spread amongst different files, which can be individually downloaded from the NHANES website.

Data header	Definition
RIDAGEYR	Age
RIAGENDR	Gender
BPXSY(1,2,3,4)	Systolic blood pressure (mm Hg)
LBXCRP	C-reactive protein(mg/dL)
LBXTC	Total cholesterol (mg/dL)
SMQ040*	Do you currently smoke?
DIQ010	Doctor told you have diabetes
LBXGH	Glycohemoglobin(%)
LBDHDD (LBXHDD in 2003)	Direct HDL-Cholesterol (mg/dL)
MCQ300A (MCQ250G in 2003)	Close relative before the age of 50 had a heart attack?

Table 3.2: Labels from NHANES dataset to calculate the RRS, where * are skip patterns.

Question DIQ010 resulted in multiple answers, namely diabetes, no diabetes, and pre-diabetes. Therefore, we have chosen to classify pre-diabetes as diabetes. This decision is based on the average values for participants with pre-diabetes compared to people with and without diabetes. This can also be viewed in Table E.6. The characteristics of pre-diabetes are much closer towards people with diabetes than those without diabetes. Hence, it seems more logical to classify pre-diabetes as diabetes. Additionally, given that we measure diabetes' state with glycohemoglobin, we believe that including pre-diabetes as diabetes is better than as non-diabetic.

For the data header, BPXSY(1,2,3,4), different systolic blood pressure readings will be used, as it can happen that a measurement has failed. Measurements were taken at various points

throughout the day; however, not all participants had readings at every moment. Hence, although a participant may have only one valid measurement (e.g., BPXSY1 only), we still include them in the analysis. If we only took one specific reading, this would reduce our final dataset by 650 participants. If people have multiple readings, we take the average of those readings to get a better average value. While observing the readings, we observe that the readings are generally equal to the other readings; hence this approach should introduce no issues.

Moreover, one of the questions asked for the RRS is not available in the same state as in the NHANES. More specifically, in the RRS, the family history of a close relative having a heart attack has been specific to being below the age of 60. However, in the NHANES dataset, it is only available for less than 50 years old. Although this can be seen as an essential variable, we will use this variable as a predictor variable for that specified variable. Given other analysis, we observe that the percentage of people with having a close relative having a heart attack before the age of 50 is very close to the percentage that we are obtaining, namely 12 % [55]. Hence, given that the percentage is similar and is close to 60, we reason using this as a predictor variable is a valid choice.

Additionally, the studies by Ridker et al. [104], [105] were conducted on healthy participants. Healthy participants were defined as not having a previous heart failure, a related heart disease, heart attack, stroke, or having any form of cancer. Given that these factors increase the risk of having a CVD event, which is not captured by the risk score, it is important to exclude these participants from the dataset. Regular checkups are more advised for these people regardless of being at high risk for CVD. To make sure these participants are being excluded from the data, we have to know which variables are required to identify these participants, which is displayed in Table 3.3. We observe that 545 participants are affected by one of the first five factors (CVD factors). Nine hundred ninety-eight participants are affected by one of the diseases described in Table 3.3 and thus have to be excluded from the data. Additionally, we also exclude pregnant females, given that during pregnancy, they are at increased risk and generally have different PA patterns [43]. This exclusion was done with question RHQ131, which is also a skip pattern.

Participants who did not answer either one of the questions presented in Table 3.2 and 3.3, will be excluded from the analysis. If they did not answer one of those questions, we could not determine if they can be included in our dataset. Additionally, if a participant answered 'don't know' or 'refused' to answer a question or laboratory test, they are also excluded from the analysis.

Data header	**Definition**
MCQ160B	Ever told you had congestive heart failure ?
MCQ160C	Ever told you had coronary heart disease ?
MCQ160D	Ever told you had angina/angina pectoris ?
MCQ160E	Ever told you had heart attack ?
MCQ160F	Ever told you had a stroke ?
MCQ220	Ever told you had cancer or malignancy ?

Table 3.3: Labels & indicators for non-healthy participants based on RRS.

Missing PA Data Missing PA values generally occur later in the day and appear in larger groups, which continue until the end of that day. These missing values are randomly present on different days, and there seems to be no indication of a particular day with the most missing values. After our first inclusion & exclusion criteria, a total of 20 out of the 46161 days have missing values. Out of the 20 missing days, only three days would meet the requirement of being considered a valid day if the missing data would be classified as nonwear. As we will discuss in the next paragraph, we only include days if they obtain a certain amount of data; otherwise, they are classified as invalid. These three days were present in females aged 70 and older. After observing these participants, we observed that they have an average of 412 null values per day. We observed

that only one female would meet the requirements to be included in the analysis if their missing values are interpreted as nonwear. Without knowing what happened during the missing values, we have decided to remove that one female participant, as this could be an outlier in our data analysis. In this study, no participants with missing data have thus been included. Having few participants with missing values is an artifact of the nonwear time requirements, which we will explain in the next paragraph.

Nonwear Time One of the exclusion criteria is wear time, which denotes the time a participant has worn the accelerometer. Certain participants only wear the device for a couple of days and then take it off. The accelerometer used to measure the PA also had to be taken off during periods it could get wet or when the participants were going to sleep. We cannot differentiate between the two, and there are likely also periods where participants took their accelerometer off due to discomfort. Therefore, to make the analysis more ridged, we remove these unreliable periods in the time series for the cut-point analysis. We will exclude a participant if they have less than four days of valid wear time, given that this is the most widely known and accepted technique to use [115]. A day of valid wear time is classified as a day where the participant has worn the device for at least ten hours. If this criterion is not met, we classify the day as having too little information and will not be included.

Nonwear time can be defined in multiple ways, as can be identified by the paper of Tudor-Locke et al. [115]. Different papers have taken different approaches to solve the issue of nonwear time [115]. The most common method, and introduced in the paper of Troiano et al. [114], defines nonwear as 60 minutes of consecutive zeros with allowance for 1-2 min of consecutive counts between 0 and 100. This definition can be interpreted in different manners, such as at most two minutes of activity below the threshold in an interval of 60 minutes. It can also be interpreted as at most two consecutive activity counts in an interval of 60 minutes, thus allowing multiple activity counts, also referred to as an allowance interval, as long as it does not exceed the threshold, and maximum consecutive number [22]. This definition is interpreted differently in various papers, as various papers obtain different results by citing the same paper. Also, researchers deviate from this algorithm and use a wide variety of other interpretations, such as 30 consecutive zero counts or allowance for more or less than two consecutive minutes of counts. All these configurations are described by Tudor-Locke et al. [115]. Therefore, it could not be possible to make a good comparison between papers, as the results obtained are affected by the wear time algorithm. We have decided to implement the most used method, as denoted in Tudor-Locke et al. [115] and use allowance intervals of at most two minutes. Hence, as long as the interval does not have three consecutively counts greater than an intensity value of 0, none of them are greater than an intensity value of 100, and the total length is at least 60 minutes, the area is considered nonwear. The pseudo-code for our wear time algorithm is presented in Algorithm 1.

Algorithm 1 Wear time flags algorithm

```
procedure WEAR_TIME_FLAGS(max_count = 2, max_value = 100, interval_min = 60)
    Add missing minutes to dataset and set those to nonwear
    Aggregate individual days to 1 total time-series, per participant
    for time series in dataset do
        Calculate bouts of consecutive 0's.
        Combine bouts if they are not apart for more than max_count, and the values between
        the bouts, does not exceed max_value
        Check which bouts are ≥ then interval_min, classify those bouts as nonwear
    Create wear-flag file, consisting of the minutes that are wear and nonwear
```

Besides the wear time algorithm, it is also often unclear what happens with the nonwear areas. It is discussed that a day needs at least a certain amount of time to be determined as valid, and if a participant has at least four days, in general, the participant is included in the analysis. It is,

however, not clear if the days that are not valid for participants with at least four days of valid data are also taken into account in the analysis. In our analysis, we remove the days that do not pass the requirement, as it is most logical to remove the days that were seen as invalid. In addition to the days, it is also often unclear if periods considered nonwear are considered for calculating the cut-point features. In our analysis, we have decided to not include the nonwear data, as this data can be seen as irrelevant for our goals. However, to observe the effects, in Section 4.2, we have tested the impact of including or excluding nonwear time data from the time series analysis.

3.2 Experiment Setup and Evaluation

In this section, we will describe the setup(s) used to conduct the experiments. If a particular section deviates, this will be explicitly noted in that section. This allows other researchers to re-do the experiments and obtain the same results obtained in this work.

Environment We used Tensorflow-GPU 2, together with Python 3.8.5 and Scikit-learn 0.24.1, to implement and obtain our results.

Random State We employ random states to ensure that results remain consistent after running an experiment. A random state allows a model, or other methods, to pick a specific seed, which it will re-use if re-run. This state allows each experiment to use the same seed and results in the same process to be executed. We have chosen seed 200, and it is used whenever possible.

Model Parameters We made use of GridSearch of Scikit-learn to optimize the model parameters using our validation data. The GridSearch settings used for our machine learning models and MiniRocket are displayed in Appendix Figure F.1.

Deep Leaning We let our models train for at least 100 epochs with a batch size of 64. A larger batch size generally causes issues due to memory limitations. Some models, such as InceptionTime, take around three minutes per epoch to train; hence using 100 epochs generally resulted in acceptable running times, together with converging losses. Additionally, we use ReduceLROnPlateau from TensorFlow, with a factor of 0.75, which reduces the learning rate by a factor 0.75 (LR*0.75). We also apply a patience of 10, which indicates that we will reduce the learning rate if we did not see any improvement for the last ten epochs. We set the min learning rate to 0.00001 while monitoring the val_loss for classification and val_mean_squared_error for the regression task. We apply this technique to prevent our learning rate from being too high and overshooting the minimum. We hope that with this callback, our model will not overshoot the minimum loss. Choosing a low learning rate at the start could also be troublesome, given that it may take too long to converge. We have chosen the Adam optimizer, with default settings, for both tasks. This optimizer was also used to train the ResNet and InceptionTime network by Fawaz et al. [39], and is often seen as a good optimizer due to an adaptive learning rate. Lastly, we store the best model with ModelCheckpoint for an epoch if our monitor value was lower than any epoch before. We also save the model from the last epoch; however, we always use the best-epoch model because the last model is always slightly worse.

3.2.1 Classification

Machine Learning Algorithms We have evaluated a large set of machine learning algorithms for this project. Executing each machine learning model for each experiment would cause much overhead and make the final results less interpretable. Therefore, we have decided to stick with four machine learning algorithms that performed well on our range of tests. We also believe that these models have different characteristics and can be used for different purposes, such as feature importance and coefficients. We will make use of a Logistic Regression (LR), a Support Vector Classifier (SVC), Decision Tree (DEC), and a Random Forest Classifier (RFC) which are

implemented with Scikit-learn. For more details about these algorithms, we refer to the Scikit-learn website [4].

K-fold Cross Validation Given the size of the dataset, we believe it is best to use K-fold cross-validation. Cross-validation is a method that splits our data into different splits and allows us to obtain a less biased or less optimistic estimate of the model skill than a simple train-test split. Our test data contains 670 participants, which has an imbalanced ratio of four low-risk participants for one high-risk participant. We will be using five-fold cross-validation, as this will allow us to obtain a reasonable amount of testing data, validation data, and training data. We use the train and validation data to check the best parameters for our models and then check the performance on our test set. It is important to denote that, after testing if non-healthy participants can be included in our training data, the distribution will change to a ratio of 2:1, in favor of low-risk participants. When more advanced methods are tested, such as deep learning and MiniRocket, we will be using a one-fold test because doing a five-fold test is too costly.

Classification Metrics

Given that we are dealing with a class imbalance, we should not only look at the accuracy, given that this might give us a wrong impression of the performance. For example, an accuracy of 80 % would sound reasonable; however, if the imbalance ratio is 4:1, this could mean it only predicts one class correctly. Therefore, we have opted to use the receiver operating characteristic (ROC) and the precision-recall (PR) curves. In addition to these metrics, we use the stratified brier-score to classify how well-calibrated our model is, as this is often clinically preferred.

Receiver Operating Characteristic (ROC) The ROC curve is a graph that displays the True Positive Rate (TPR), also known as sensitivity or recall, against the False Positive Rate (FPR), also known as 1-specificity, with different classification thresholds. Usually, a 50 % threshold is used to label something as either label a or b; however, this could be troublesome in an imbalanced dataset. The ROC curve's power stems from the different thresholds to classify something either low or high risk. Using a threshold of 20 % could lead to a suitable classifier for both high and low risk, while using a 50 % threshold could only result in the prediction of one specific class. These different thresholds are then used to plot the TPR against the FPR at different thresholds. An example of the ROC-curve is shown in Figure 3.2, which displays a few different classifiers indicated by the different colors. The purple line represents the best classifier since we can predict all classes perfectly at 0 FPR. We aim to make a classifier that has the highest possible TPR with the lowest FPR. The threshold to classify someone as high risk goes down the more we move to the right in the curve. This shift means that the lower our threshold is, the more cases we will start labeling as positives, in our case, high risk.

As previously mentioned, the ROC curve can help us define different thresholds to observe what a reasonable TPR rate against a FPR rate is. It is important to note that specificity is equal to 1-FPR. Hence, this allows us to define our threshold for classifying high or low-risk cases by selecting our optimal TPR and FPR rate. The ROC curve does not have any bias toward models that perform well on the minority class at the expense of the majority class. This property is quite attractive when dealing with imbalanced data [50]. Which threshold we want to use is case dependant because sometimes a high TPR is preferred, where we care less about a high FPR, while in other cases, we also care about the FPR, and thus want to take a lower TPR to account for a lower FPR. Lastly, to make a quick comparison between classifiers, we will be using the area under the curve, which indicates how well the classifier performed.

Precision-Recall (PR) The PR curve also uses various thresholds, making it suitable for an imbalanced classification task. PR curves are also frequently used in clinical settings, given that they focus on the minority class, which is where the PR curves are more informative than the ROC curve. The PR curve makes a trade-off between the precision and recall of the classifier. Precision

defines the fraction of how often the minority class is predicted correctly among all selected positive classes. In contrast, recall defines the fraction of positive classes correctly identified among all positive classes. An example of the PR-curve is displayed in Figure 3.3. We observe that compared to the ROC-curve, it works slightly differently; however, the intuition is similar. The best classifier is indicated by the purple line, where our precision is always one, even if our recall increases. We want to obtain the maximum possible precision and recall.

The PR-curve is thus used because the ROC-curve can give an optimistic view of the performance of a classifier. The ROC-curve does not take into account the fractions of positive and negatives compared to the PR curve. Therefore, in a clinical setting, both curves are often used to depict a complete picture of the predictive performance of a model. The PR curve, in our case, is also easier to use. If we want to select a certain amount of participants for treatment, we can do this based on precision, which is more informative in a clinical setting than the TPR. In the PR curve, we primarily inspect the differences at high precision values due to being more relevant when selecting patients for treatment.

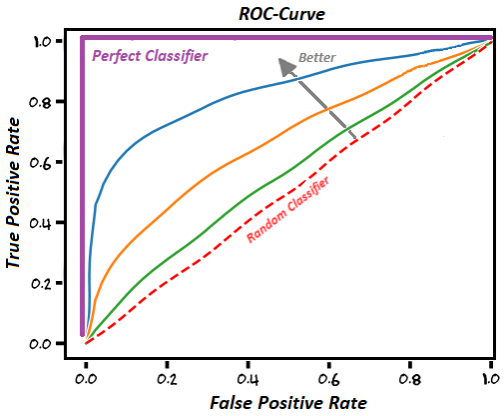


Figure 3.2: ROC curve example.

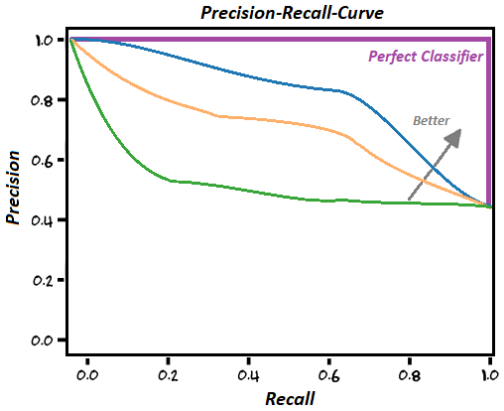


Figure 3.3: PR curve example.

Brier-Score The Brier-Score (BS) measures how well a model is calibrated. A calibrated model is clinically preferred as it indicates the model is confident in its predictions. Normally, a calibration curve is constructed to show how calibrated our model is. We observed, however, when calculating the calibration curves in different settings that our test set is too small to show how well our model is calibrated. Therefore, we have opted to make use of the BS, more specifically, the stratified BS, given that this is often better for imbalanced classification tasks [119]. The BS is a scoring function that measures the accuracy of the probabilistic predictions. This score shows how confident the model is in making its predictions. The stratified BS, is defined in Equations 3.1 and 3.2, for the negative and positive class respectively, where $y_i \in \{0, 1\}$ and $x_i \in [0, 1]$.

$$BS_{neg} = \frac{\sum_{y_i=0}(y_i - P\{y_i|x_i\})^2}{N_{neg}} \tag{3.1}$$

$$BS_{pos} = \frac{\sum_{y_i=1}(y_i - P\{y_i|x_i\})^2}{N_{pos}} \tag{3.2}$$

Another advantage of using a simple metric, which we want to minimize in both cases, although we will be mainly using the ROC and PR curves, is that we do not have to investigate three graphs per experiment, making interpreting the results more challenging. However, we have to keep in mind that the bier-score is not as informative as a calibration curve. Lastly, we have opted not to use any other metrics, such as the F1-score, kappa coefficients, G-mean, given that they are generally hard to interpret and give little additional value to our current metrics.

3.2.2 Regression

Machine Learning Algorithms We have chosen to test a Random Forest Regression (RFR), AdaBoostRegressor (ADA) and Ridge model as our machine learning techniques for the regression task. Like our classification models, we argue that using too many models causes too much overhead, and these models use different structures and should interpret the features in different manners. More details about these techniques is available on the Scikit-learn website [4].

One-Fold Validation For the regression task, we will use a one-fold test. We assume that the experiments on the classification task about normalized features, wear time, and imputation already gave us clear conclusions and thus do not have to be repeated. Hence, the regression chapter will mainly focus on the performance of the different models and how well the models can predict the RRS. Given these points, using a five-fold test is not obtainable due to the more sophisticated approaches being used, which is very costly.

Regression Metrics

BA-diff To evaluate the performance of our regression task, we have chosen to use our own metric, given that other metrics do not adequately visualize the performance. With this metric, we can inspect where the model is over predicting and under predicting the RRS. The Bland Altman plot inspired this metric which we slightly altered to fit our problem better. We will refer to this metric as the Bland-Altman-diff (BA-diff). The BA-diff shows on the y-axis the true risk of our test set and on the x-axis the difference between the predicted risk - true risk, as displayed in Figure 3.4. Hence, the y-axis shows the difference between the true and predicted risk. A negative score/risk indicates areas where the model under-predicted the risk, and positive differences, where the model over-predicted the risk. Zero would mean a perfect prediction of that test sample. Hence, we observe at the lower true risk values that the model is over predicting the risk, while at higher true risk values, 40 on-wards, the model is under-predicting the risk in Figure 3.4.

Figure 3.4: BA-diff example for predicting the RRS.

In the BA-diff, we also visualize a mean line, which shows that the model predicted on average 0.51 risk to high in example Figure 3.4. In addition to the mean line, we also visualize two standard deviations (SD) lines. These should indicate where most of our predictions are. The further these lines are apart from each other, the more the model deviates in its predictions. However, depending on the use case, someone could prefer the negative or positive SD to be closer to zero, such that the model under-predicts or over-predicts less, respectively.

R2 The R2 score, also known as coefficients of determination, shows the proportion of variance between the predicted and observed variables. When the predicted variable fits the observed variable perfectly, we have a fit of one; hence the closer this metric is to one, the better the model fits the data. This metric allows us to known how well the data fits the model.

Mean Absolute Error (MAE) We also use the MAE, which displays the mean error of our models. We also use the top 10 % MAE, to observe the MAE for the top 10 % highest risk participants. The MAE shows the average mistake in risk percentage the model makes.

3.3 In Depth Analysis

This section will conduct a more in-depth analysis of the data and shows interesting results found from this in-depth analysis. The analysis is divided into three sections, intensity features analysis, risk factors analysis, and an activity comparison.

3.3.1 Intensity Feature Analysis

From Table 3.4, we observe the difference in intensity, also referred to as PA levels, per gender. We observe from Table 3.4 that there is an apparent decline in the amount of activity conducted at higher intensity groups for both genders. Additionally, a more interesting thing to observe is that females, on average, have less PA at higher intensities, starting from lifestyle activity. We even observe that females have twice as little moderate and vigorous activity compared to males on average. This difference is important because it may imply that females require less PA to obtain the benefits from PA. We observe from that it is known that women move less on average compared to males, especially at moderate activity levels, as argued by Azevedo et al. [8].

	Both (N = 3348)					**Male (N = 1596)**					**Female (N = 1752)**				
	Mean	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**
c_sedentary	475.64	120.64	474.48	67.0	1051.71	476.39	129.74	473.34	67.0	1051.71	474.95	111.71	476.07	110.14	988.86
c_light	267.41	66.43	265.62	30.8	530.6	258.34	66.82	258.5	30.8	508.6	275.66	64.99	274.23	63.57	530.6
c_lifestyle	89.64	50.42	81.67	1.14	393.2	102.17	53.54	95.14	1.14	393.2	78.22	44.43	71.0	2.5	316.0
c_moderate	21.78	22.2	15.67	0.0	208.5	29.04	25.58	23.07	0.0	208.5	15.17	15.96	10.07	0.0	180.83
c_vigorous	0.75	3.08	0.0	0.0	40.5	1.04	3.63	0.0	0.0	40.5	0.49	2.43	0.0	0.0	37.75

Table 3.4: PA intensity features per gender.

From this data analysis, we also observe that there is a clear difference in terms of activity between the low and high-risk participants, as is depicted in Table 3.5. We observe that high-risk groups generally have more sedentary time and less activity in other activity groups, especially starting from lifestyle activity.

	Both low (N = 2639)					**Both high (N = 709)**				
	Mean	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**
c_sedentary	463.13	119.41	460.33	67.0	988.86	522.18	113.59	515.29	170.25	1051.71
c_light	273.31	64.8	271.0	63.57	530.6	245.41	67.82	242.5	30.8	494.29
c_lifestyle	97.58	49.4	88.33	2.5	393.2	60.09	42.6	50.8	1.14	238.83
c_moderate	24.82	22.99	19.2	0.0	208.5	10.46	14.08	4.75	0.0	103.6
c_vigorous	0.91	3.42	0.0	0.0	40.5	0.14	0.86	0.0	0.0	14.57

Table 3.5: PA intensity features for low and high-risk.

Given that gender could play a big role, we also have display the difference in low and high-risk per gender, as depicted in Tables 3.6 and 3.7 for males and females, respectively. Besides the average being different, we observe that the max values are higher for both males and females for intenser activities when in the low-risk group compared to the high-risk group. This difference indicates that people in the high groups never conduct more activity than low-risk groups for each respective category. Hence, we conclude that there are clear differences between the low and high-risk groups and clear differences between males and females present.

	Male low (N = 1171)					**Male high (N = 425)**				
	Mean	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**
c_sedentary	458.26	129.57	452.86	67.0	983.71	526.36	116.38	519.86	170.25	1051.71
c_light	265.02	66.52	263.86	99.0	508.6	239.93	64.11	238.0	30.8	494.29
c_lifestyle	113.84	51.81	106.29	4.29	393.2	70.02	44.28	62.33	1.14	238.83
c_moderate	34.71	26.21	29.25	0.25	208.5	13.41	15.16	8.25	0.0	93.86
c_vigorous	1.35	4.17	0.14	0.0	40.5	0.17	0.82	0.0	0.0	10.57

Table 3.6: PA intensity features for low and high-risk males only.

	Female low (N = 1468)					**Female high (N = 284)**				
	Mean	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**
c_sedentary	467.02	110.49	465.62	110.14	988.86	515.91	108.98	513.4	223.14	894.5
c_light	279.93	62.61	277.66	63.57	530.6	253.62	72.25	249.48	92.0	463.4
c_lifestyle	84.61	43.2	77.54	2.5	316.0	45.23	35.06	35.51	3.14	187.86
c_moderate	16.94	16.18	12.33	0.0	180.83	6.04	10.87	2.57	0.0	103.6
c_vigorous	0.56	2.62	0.0	0.0	37.75	0.08	0.91	0.0	0.0	14.57

Table 3.7: PA intensity features for low and high-risk women only.

3.3.2 Risk Factors Analysis

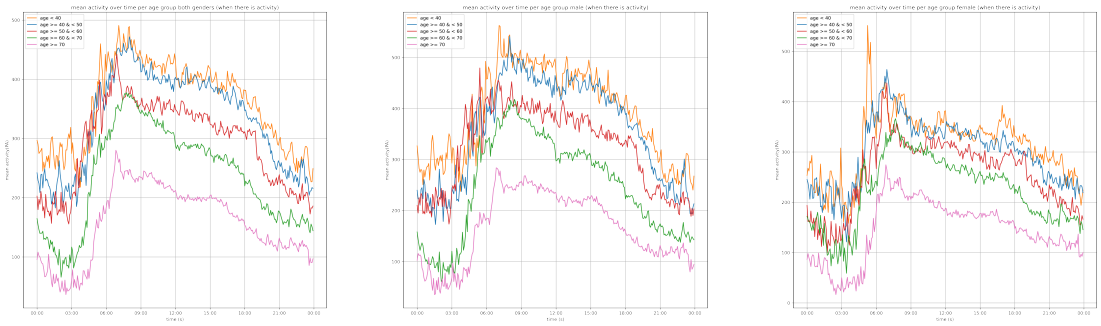
In this section, we compare the risk factors that are correlated with CVD and also used in the RRS, between males and females, as seen in Table 3.8. We observe no significant differences between males and females; we only notice that females have higher HDL-cholesterol values, which is explained by the effect of menopause. Post-menopausal women tend to have higher HDL-cholesterol values than younger women, as argued by Razay et al. [102]. It is, however, difficult to conclude if the HDL-cholesterol variables are representative of the average in the US, given that often the HDL-cholesterol variables are not disclosed and are very influential by age and the country [5]. From some papers, we would argue the values can be seen as invalid [104], [10]; however, some papers obtain similar HDL-cholesterol values, such as in the study of Vodak et al. [118]. We conclude that, although the HDL-cholesterol values, in our opinion, are on the higher side, they could still be possible. We do, however, have our doubts if they are perhaps a bit too high. Given that HDL-cholesterol is seen as good cholesterol, which is beneficial to our body, we would like to have a high cholesterol value. We observe from the data that males are at higher risk than females, which is also known based on literature. Thus, it could be that due to females having higher cholesterol values, that they are at lower risk at certain ages.

	Both (N = 3348)					**Male (N = 1596)**					**Female (N = 1752)**				
	Mean	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**	**Mean**	**SD**	**Median**	**Min**	**Max**
Age	52.67	14.74	51.0	30.0	85.0	51.39	14.56	49.0	30.0	85.0	53.84	14.8	52.0	30.0	85.0
Systolic blood pressure	126.64	19.91	123.33	80.0	256.0	126.11	16.73	123.33	87.33	216.67	127.13	22.42	123.33	80.0	256.0
CRP	0.42	0.74	0.21	0.01	17.5	0.34	0.76	0.17	0.01	17.5	0.49	0.71	0.25	0.01	9.31
Total cholesterol	205.86	39.14	204.0	85.0	394.0	204.15	37.97	201.0	98.0	360.0	207.42	40.11	205.0	85.0	394.0
Glycohemoglobin	5.54	0.79	5.4	4.1	13.4	5.46	0.6	5.4	4.1	12.8	5.62	0.92	5.4	4.3	13.4
HDL-cholesterol	55.35	16.21	53.0	22.0	154.0	49.58	14.17	47.0	22.0	116.0	60.6	16.17	58.0	26.0	154.0
Diabetes	0.06	0.23	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.11	0.31	0.0	0.0	1.0
Smoking	0.2	0.4	0.0	0.0	1.0	0.25	0.43	0.0	0.0	1.0	0.16	0.36	0.0	0.0	1.0
Family Stroke	0.12	0.33	0.0	0.0	1.0	0.11	0.31	0.0	0.0	1.0	0.14	0.34	0.0	0.0	1.0

Table 3.8: Risk factors for the RRS per gender.

3.3.3 Activity Comparison

In Figure 3.5a, we have displayed the average intensity that is conducted, per age group, for an entire day. It is important to note that, for example, if a participant had a period of nonwear from 12 PM until 1 PM, we will not count this in this graph; hence, only periods that wear considered as active or wear time will be averaged out.

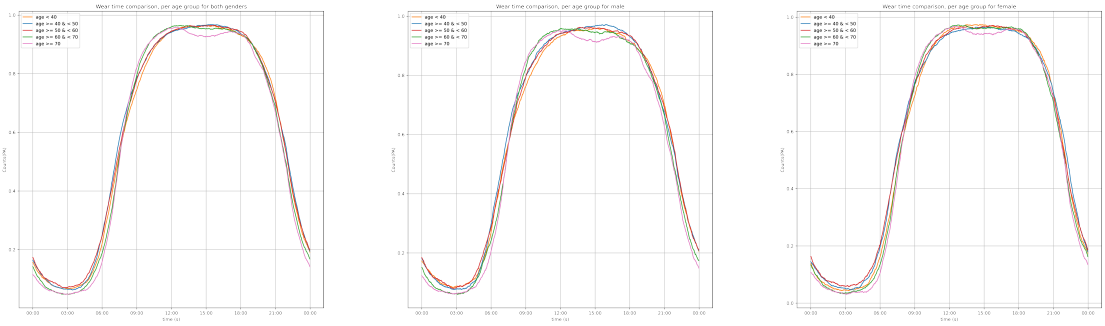


(a) Activity count (b) Male activity count (c) Female activity count

Figure 3.5: Average activity per age groups on one day of PA, only including wear time data.

The interesting thing we observe from figure 3.5a is that there is a clear difference in PA intensity over the day per age group. People of lower ages conduct more intense PA than people of older ages. It also appears that there is a more significant drop at older ages than at lower ages, given that the orange and blue lines still follow closely. When we look at the specific plots per gender, as displayed in Figures 3.5b and 3.5c, we also observe, as previously conducted from the raw data, males conduct more intense activity than females. This conclusion, however, seems less apparent at older ages between the two genders and is more present at younger ages. We observe that there is also less of a difference in the age groups for females than the males. This difference could indicate that the effect of age on the activity is less present for females than males.

To further observe the effect of activity at different time points, we will take a look at Figure 3.6. This figure displays the wear time across the different age groups, per gender. Hence, the y-axis represents the ratio of wear time, compared to the maximum available wear time at that point. Thus, this indicates that at 12:00, 95 % of the data points are considered wear time. Hence, we observe that there is no time in a day where each participant has valid data. Consequently, there are always some participants that have an area of nonwear.



(a) Ratio both genders (b) Ratio males (c) Ratio females

Figure 3.6: Ratio of wear time per age group of one day.

Furthermore, we also observe that at midnight, such as 3 AM, most people are supposedly

sleeping, as then only around 7 % of the participants have some valid wear time. Lastly, an interesting observation we have made from Figure 3.6 is that we observe a dip in wear time, for people above the age of 70, from around 1 PM until 5 PM. We suspect that this indicates a nap for older people. The wear time algorithm detects these periods as nonwear, hence supposedly when the participants are sleeping. We cannot draw a concrete conclusion from this, as it could also have been that older people took a bath around that time. However, we suspect that this is more likely related to a nap in the afternoon. People above the age of 70 are logically more home as people who are still working, and perhaps also more fatigued, due to being older. Additionally, Cheng et al. [21] also stated that an afternoon nap is also common among the Chinese population, especially of older ages, and it is considered tradition to take a nap around noon. We reason that a similar occurrence could be present in the US, and we reason it is more likely that the dip in wear time is correlated to an afternoon nap than to an activity that involved water.

3.4 Features Derived from Cut-Point Analysis

This section explains the features used derived using the cut-point analysis utilized in our machine learning models. The baseline features consist of the number of counts in each respective intensity group, as referred to in Table 2.1. This feature list is most often used in literature and is also easy to use. Some researchers have introduced additional features, such as bouts of activity and the amount of activity spend in the week compared to the weekend. We have gathered all the available features and have extended the features list with the following new features that we have not seen used before based on our literature review.

- Features that describe at which part of the day an activity was conducted.
- Extension of features regarding bouts to other intensity categories.
- Normalized features.

We decided to split a day into four areas: midnight, 0-6; morning, 6-12; afternoon, 12-18; and evening, 18-24. This split was motivated by various researchers indicating that the time of day activity is conducted could be of importance for CVD [36]. Conducting late-night activities could cause more arousal during sleep and degrade a person's sleep quality. Activity during the morning could also allow for a better restorative process to reduce our blood pressure. Secondly, we often only see bouts related to MVPA; however, we observe that lifestyle activity is also essential, especially in females. Hence, we have extended the bouts and various other features to be more informative about other intensity categories instead of only MVPA. Lastly, we have also normalized certain features based on the wear time. A normalized feature scales a feature based on the amount of wear time a person has daily. For example, if a person conducts 200 moderate counts every day, with 800 active minutes each day, we obtain a moderate feature of 200. For a normalized feature, this gets scaled; hence, in the same scenario, we would have $200/800 = 0.25$. If another person also has 200 moderate counts but 1000 active minutes, the count's feature stays the same. However, the normalized feature now changes to $200/1000 = 0.2$. Hence, the normalized feature takes into account the wear time compared to the counts.

In Table 3.9 we have listed all the available features that can be used in our models. The last four columns represent the different sets of features used in this analysis. Before doing the experiments, we did a feature selection process by observing the Pearson correlations between features and using the recursive feature elimination module from Scikit-learn to obtain the most relevant features, resulting in the advanced feature set [4]. This advanced feature set outperforms using all features in our models, which could be explained by having less correlated features in the advanced feature set. Having many high correlated features could degrade the performance and overfit our models. Some features in Table 3.9 express the activity for a certain time period or category. For instance, the wear time feature is calculated for morning, afternoon, evening, and

midnight, resulting in four different features. For reasons of brevity, we write the replaceable part of the name between square brackets (wear_[time]), where the replacements can be as follows:

time ∈ {midnight, morning, afternoon, evening}
period ∈ {week, weekend}
intensity ∈ {sedentary, light, lifestyle, moderate, vigorous}
activity ∈ {MVPA, sedentary, lifestyle}
ordinal ∈ {1(st), 2(nd), 3(rd), 4(th)}

Features	Feature description	Baseline	Baseline norm	Advanced	Advanced norm
valid_days	Number of valid days for that participant				
wear_minutes	Number of valid wear minutes for that day				
wear_[time]*	Amount of wear_minutes in [time] / valid_days				
total_activity	Total amount of activity from all valid days				
total_[time]*	Total amount of activity in the [time] from all valid days				
mean	Sum means per day / valid_days				
mean_norm	total_activity / total wear_minutes for all valid days				
mean_[time]*	Sum total_[time] for al valid days / valid_days				
mean_[time]_norm*	Sum total_[time] for al valid days / wear_[period] for all valid days				
mean_[period]*	total_activity for all valid days in [period] / wear_minutes for valid days in [period]				
c_[intensity]*	Total amount of [intensity] counts / valid_days	X			
c_[intensity]_norm*	Total amount of [intensity] counts / wear_minutes for all valid days		X		
c_[intensity]_[time]*	Total amount of [intensity] counts in the [time] / valid_days			X	
c_[intensity]_[time]_norm*	Total amount of [intensity] counts in the [time] / wear_[time] for all valid days				X
c_MVPA_[period]*	Amount of moderate and vigorous counts in the [period]				
c_lifestyle_[period]*	Amount of lifestyle activity counts in the [period]				
tot_wk_mv_dif_min	c_MVPA_week - c_MVPA_weekend				
tot_wk_li_dif_min	c_lifestyle_week - c_lifestyle_weekend				
avg_wk_mv_dif_min	tot_wk_mv_dif_min / valid_days				
avg_wk_li_dif_min	tot_wk_li_dif_min / valid_days				
perc_[period]_MVPA*	c_MVPA_[period] / total_c_MVPA			perc_week_MVPA	perc_week_MVPA
perc_[period]_lifestyle*	c_lifestyle_[period] / total_c_lifestyle			perc_week_lifestyle	perc_week_lifestyle
num_MVPA_bouts	Number of MVPA bouts of length ≥ 10 / valid_days			X	X
MVPA_min_bouts	Number of minutes accumulated in MVPA bouts / valid_days				
num_sed_bouts	Number of sedentary bouts of length ≥ 60 / valid_days				
sed_min_bouts	Number of minutes accumulated in sedentary bouts / valid_days			X	X
num_lif_bouts	Number of lifestyle bouts of length ≥ 10 / valid_days			X	X
lif_min_bouts	Number of minutes accumulated in lifestyle bouts / valid_days				
num_light_bouts	Number of light bouts of length ≥ 10 / valid_days				
light_min_bouts	Number of minutes accumulated in light bouts / valid_days				
num_vig_bouts	Number of vigorous bouts of length ≥ 10 / valid_days				
vig_min_bouts	Number of minutes accumulated in vigorous bouts / valid_days				
MVPA_bouts_[period]*	Amount of MVPA bouts in the [period] / valid_days				
wk_mv_dif_bout	MVPA_bouts_week - MVPA_ bouts_weekend				
lif_bouts_[period]*	Amount of lifestyle bouts in the [period] / valid_days				
mv_wk2	2 MVPA bouts of length 10 (True or False)				
[activity]_max_[ordinal]*	Number of [activity] counts in [ordinal] longest [activity] bout			MVPA_max_1	MVPA_max_1
[activity]_min_max_[ordinal]*	Number of [activity] counts on a day with [ordinal] most [activity] counts				

Table 3.9: All available features, where features with * have one or more replaceable [variable(s)]

Chapter 4

Classification Task Experiments

This chapter will discuss the results on the classification task, where a threshold of 10 % is used to classify someone as high-risk. Each section will explain the goal and display the results which are then also discussed in that section. The first section will explain the data transformation steps for the cut-point analysis. Then several experiments follow to answer some of our research questions about wear time, normalized features and imputation. Lastly, we discuss the effect of adding other factors to the models, and dive into the performance of SOTA approaches and the advanced feature set compared to the baseline (cut-point analysis with basic features).

4.1 Data Transformation

This section explores data transformation techniques, which we have to consider before doing our experiments. It will not be possible to re-evaluate every scaling or sampling technique for each experiment. Hence we explore it once and then use the same techniques on all other experiments.

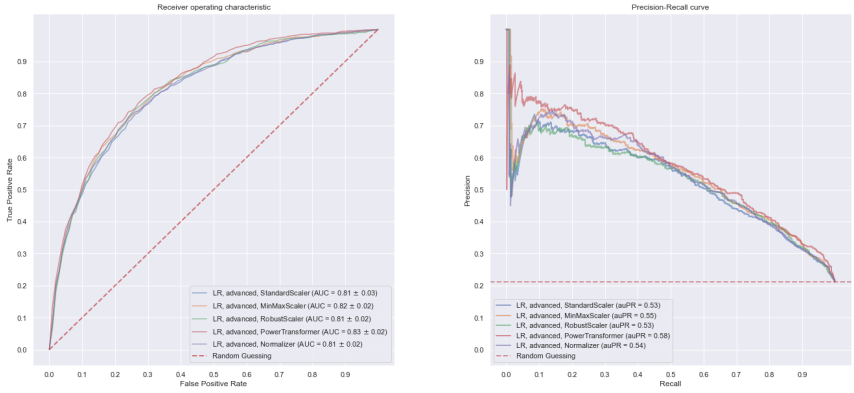
4.1.1 Scaling Techniques

This section will discuss and experiment which scaling technique is most appropriate in terms of the cut-points analysis. It is important to note that, before discussing the results, that we have ran the experiments for the RFC and the DEC; however, we found no impact in performance. Observing no differences is expected, as decision tree models do not expect or change in a decision if the underlying data distribution changes. These are rule-based models, and a scaling technique is similar to applying a scale to the value, which does not change the decision rule. Therefore, we present no results regarding these models in this section.

Scaler A scaler is a preprocessing technique that allows us to re-scale our raw data values into new values. This scaling allows the values to be internally consistent, which many machine learning models assume, such as the data to be normally distributed. This distribution, however, is often not present. Therefore, a scalar could be applied, which would scale the values to represent the desired distribution. Moreover, it is also often undesired to use large values in models, which can become very expensive. Hence, a good scaler is critical for having a good model. We have opted to test five different scaling methods: the StandardScaler, MinMaxScaler, RobustScaler, Power transformer, and Normalizer. Each of these scalers is explained on the Scikit-learn website [4]. We fit the scaler with our training data and then transform the training data accordingly. Then we transform the validation and test data accordingly.

Results and Conclusion After inspecting the results, we observe that a normally distributed dataset created by the PowerTransformer results in the best performance. As an example, we show the ROC and PR curves for our LR advanced model in Figures 4.1a and 4.1b, respectively.

We observe that both curves are closer to their optimal curve for the PowerTransformer. This difference is especially evident in the PR curve, and we observe that using a PowerTransformer on our dataset is most beneficial. We observe an increase in precision at most recall values than the other scaling techniques, especially at low recall values. The other ROC and PR curves are available in Appendix E.1, where we draw a similar conclusion for the other models. The summarized metrics are also available in Table 4.1 for all configurations. The metrics are displayed in rows, and the scaling options as columns per machine learning model. When looking at the distribution of many features, we observe that they are indeed very skewed and not normally distributed. After applying the PowerTransformer, we observed that most features become more normally distributed, based on the Shapiro-Wilk statistic.



(a) ROC curve (b) PR curve

Figure 4.1: Logistic Regression scaling comparison.

Baseline features	**Logistic Regression**				
	Standard	MinMax	Robust	Power	Norm
AVG AUC	0.748	0.741	0.755	**0.767**	0.736
AVG BS 0	0.214	0.208	**0.175**	0.176	0.188
AVG BS 1	**0.204**	0.243	0.244	0.219	0.233
AVG auPR	0.49	0.402	0.49	**0.513**	0.465

Advanced features	**Logistic Regression**				
	Standard	MinMax	Robust	Power	Norm
AVG AUC	0.809	0.817	0.814	**0.827**	0.813
AVG BS 0	0.165	**0.148**	0.149	0.15	0.156
AVG BS 1	0.197	0.204	0.21	**0.191**	0.201
AVG auPR	0.529	0.551	0.531	**0.575**	0.544

Baseline features	**Support Vector Classifier**				
	Standard	MinMax	Robust	Power	Norm
AVG AUC	0.778	0.761	0.76	**0.782**	0.754
AVG BS 0	0.088	0.092	0.092	**0.085**	0.097
AVG BS 1	**0.346**	0.356	0.357	0.35	0.355
AVG auPR	0.504	0.503	0.499	**0.513**	0.496

Advanced features	**Support Vector Classifier**				
	Standard	MinMax	Robust	Power	Norm
AVG AUC	0.831	0.824	0.82	**0.836**	0.812
AVG BS 0	0.083	0.085	0.086	**0.08**	0.089
AVG BS 1	**0.297**	0.301	0.304	**0.297**	0.308
AVG auPR	0.563	0.551	0.542	**0.585**	0.537

* Note: best metrics for each model are **bold**. Maximize AUC and auPR and minimize BS 0 and BS 1.

Table 4.1: Scaling comparison for LR and SVC.

4.1.2 Sampling Techniques

Many machine learning techniques often give a misleading performance on binary classification datasets where the dataset is imbalanced. The machine learning algorithm will put a higher emphasis on the classes that are more present. Although we have chosen several metrics capable of handling this imbalance, we are still interested in balancing the dataset and observing if this can increase predictive power. It is also argued that it highly depends on the machine learning techniques used. Sun et al. [111] argues that SVC models are less prone to imbalanced data compared to other models. One of the techniques that are used to tackle an imbalanced dataset is called sampling. The class imbalance of the test set is 4 low risk against 1 high risk participant.

Sampling Sampling is a technique of creating a balanced dataset by adding or removing data. We will be using various methods, namely, a weighted classifier, which will penalize the classifier more heavily if mistakes are made on the underrepresented class. Moreover, we apply under-sampling, which removes some of the majority classes. This method should generally only be applied when we have enough data, and we can take the loss of removing data. Additionally, we can oversample our data by simply duplicating some of our minority classes. A downside of this method is that it could cause the model to overfit the minority classes, especially if these are very high dimensional. Lastly, we can apply a technique called Synthetic Minority Oversampling TEchnique (SMOTE) [20]. SMOTE selects a certain amount of samples, similar to the k-nearest neighbors algorithm, and then draws a new sample based on the selected feature space.

Figure 4.2 shows the structure of how the data is being sampled and being processed including the order. Step 1, which is adding non-healthy participants only to the training data, is optional; however, as we will see in the next section, this is beneficial for the results. We have tested the sampling method with and without adding the non-healthy participants; however, given the conclusion of the next section, the results below will already include the non-healthy participants.

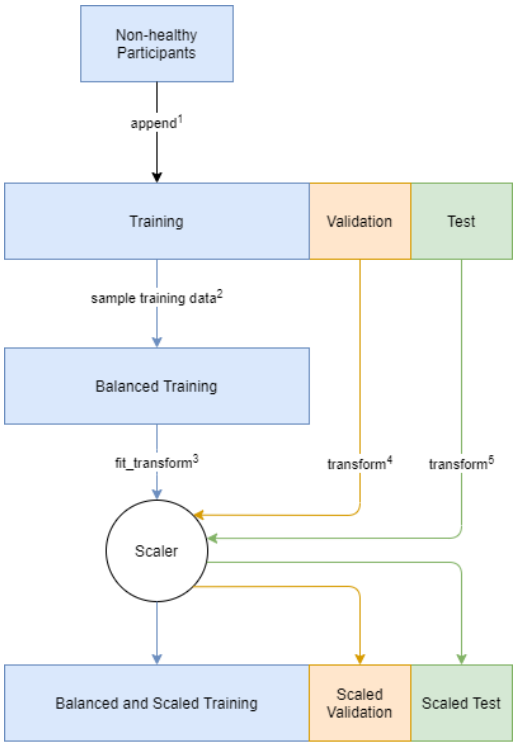
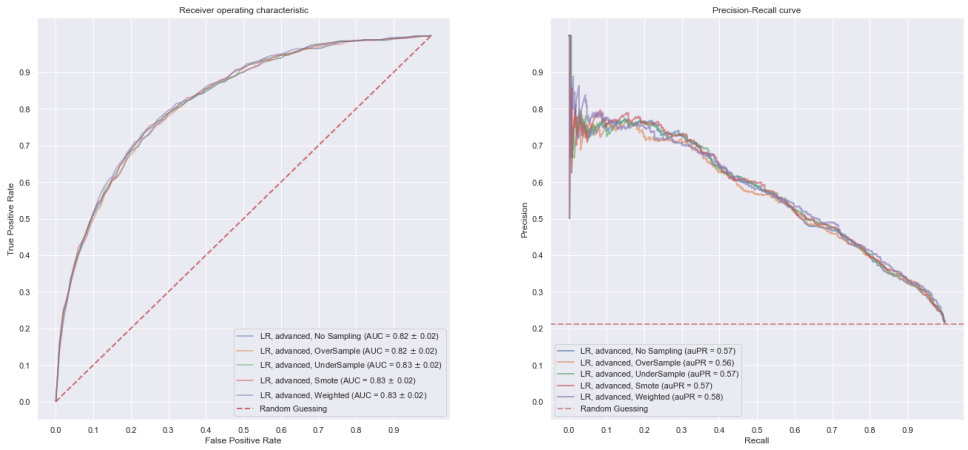


Figure 4.2: Sampling procedure including steps, where step 1 is optional.

Results and Conclusion We observe that sampling has an effect but does not drastically change the ROC and PR curves for any of the models. This can be viewed in the ROC and PR curves for the advanced Logistic Regression model in Figure 4.3a and 4.3b. As can be seen in both curves, the changes are minimal. This minimal change is also partly expected, given that these metrics make use of thresholds and thus do not mind how confident the model is in predicting high or low-risk. We are, therefore, to argue that sampling has little to no impact on the performance. However, we observe with our BS that the model's confidence changes with the sampling methods. We observe for all methods that the models start becoming more confident in predicting the high-risk group than the low-risk groups. As seen in Table 4.2, we observe that the AVG BS 1 is better than without sampling, compared to BS 0, which is now worse for all of the models.



(a) ROC curve

(b) PR curve

Figure 4.3: Sampling curves advanced Logistic Regression model

Advanced features	**Logistic Regression**				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	0.821	0.82	0.826	0.826	**0.827**
AVG BS 0	0.098	0.149	0.142	0.141	0.15
AVG BS 1	0.28	0.204	0.204	0.205	**0.191**
AVG auPR	0.568	0.56	0.572	**0.575**	**0.575**

Advanced features	**Support Vector Classifier**				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	0.833	0.831	**0.836**	0.832	**0.836**
AVG BS 0	**0.073**	0.142	0.136	0.134	0.08
AVG BS 1	0.318	**0.197**	0.199	0.207	0.297
AVG auPR	**0.588**	0.575	0.581	0.576	0.585

Advanced features	**Random Forest Classifier**				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	**0.828**	0.822	0.829	0.826	**0.828**
AVG BS 0	**0.078**	0.14	0.124	0.13	0.122
AVG BS 1	0.316	**0.211**	0.226	0.222	0.23
AVG auPR	0.571	0.569	0.569	0.562	**0.573**

Advanced features	**Decision Tree Classifier**				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	0.775	**0.788**	0.775	0.775	0.777
AVG BS 0	**0.092**	0.166	0.155	0.154	0.157
AVG BS 1	0.347	**0.216**	0.246	0.257	0.239
AVG auPR	0.505	0.49	0.502	0.483	**0.523**

Table 4.2: Sampling results on advanced features.

We have decided to apply a weighted classifier, to all our models, except for both SVC and the baseline DEC. We observed a slightly better PR curve for these models when not using sampling techniques. For the other models, the changes were marginal, or the usage of the re-weighted classifier was slightly better in terms of ROC and PR curves. We also reason that given the minimal changes, artificially changing the data is not worth it, and boosting the model's confidence with a re-weighted classifier is then the best option. The ROC and PR curves for the other models are also visible in Appendix E.2. The baseline metrics, are also available in Table E.1.

4.1.3 Including Non-Healthy Participants in the Training Data

In this subsection, we discuss the impact of including the non-healthy participants in our training data and discuss if it is beneficial for the performance of our models. Adding additional data could increase the performance of the models due to additional information. Furthermore, it would give us a more balanced dataset given that non-healthy people are generally older, resulting in more people in the high-risk group. These non-healthy participants can only be added to the training data, given that the RRS was conducted on healthy participants, as explained in Section 3.1.3. Previous diseases could increase the risk of a cardiovascular event in the future, for example, weakened or damaged blood vessels, diverted blood flow, or increased weight gain [104]. However, we argue in Appendix E.3 that this will not be an issue for the RRS. We argue that non-healthy participants have similar PA patterns, and have a similar risk than healthy people, due to previous CVD events not being accounted for in the RRS.

In this experiment, we are again using five-fold cross-validation to be more confident of our results. We use the same folds as before; however, we add the non-healthy participants to our training data for each fold. This structure ensures that the distribution of the validation and test

set is not affected. We want to compare the same distribution before and after adding non-healthy participants to the training data. This structure is visualized in Figure 4.4.

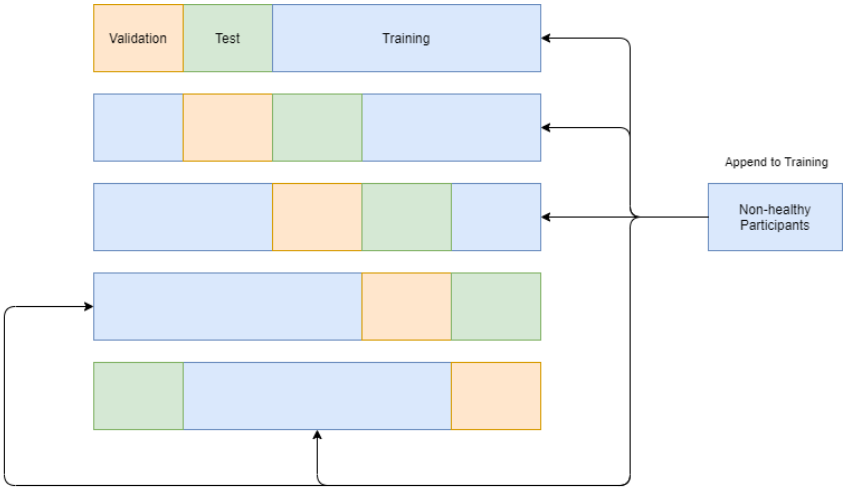
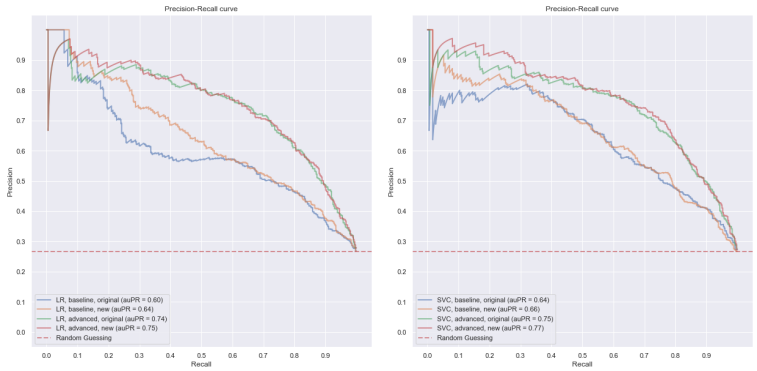
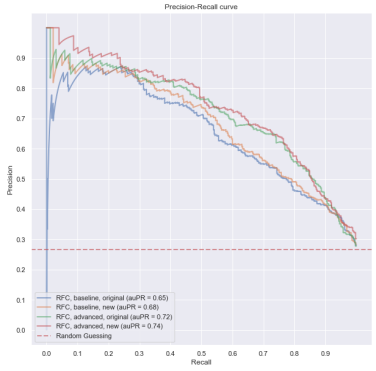


Figure 4.4: Five-fold split after adding non-healthy participants to the training data.

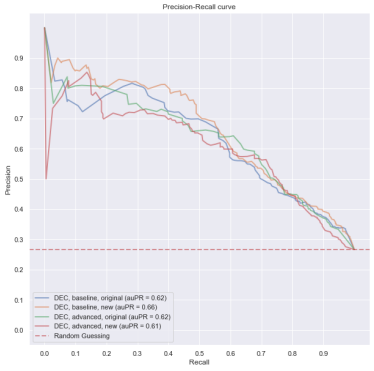


(a) LR

(b) SVC



(c) RFC



(d) DEC

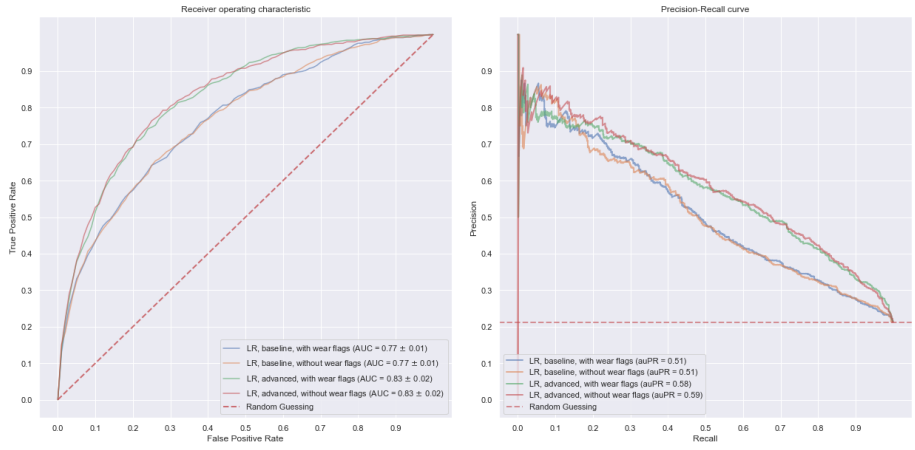
Figure 4.5: Comparison of inclusion and exclusion of non-healthy participants to training data, males only, where *original* = without non-healthy participants in the training data and *new* = inclusion of non-healthy participants in the training data.

Results and Conclusion We observe from the PR curves, displayed in Figure 4.5 that including non-healthy participants allows the performance to improve. This performance increase happens for both males and females. We observe that the PR curves are better for all the models at most points in the PR curve, especially at recall values below 0.5. We observed that the ROC curves differ little compared to the PR curves, hence why they are available in Appendix E.3. Hence, we argue and conclude that adding these non-healthy participants to our training data is beneficial. It allows us to obtain more information regarding certain risk groups, and based on our evaluation, the PA does not significantly differ between healthy and non-healthy participants.

4.2 Wear Time Flags Effect

This experiment aims to observe how the wear time flags affect the end-to-end prediction of the classification task. We have previously observed from Tudor-Locke et al. [115] that many different wear time algorithms have been used, some more used than others. The question thus arises, what is the actual impact of these wear time flags, and how reliable are the results of researchers not using a wear time algorithm or using deviations of the most used algorithm.

Results From the results, we observe that the end-to-end prediction of the classification task is affected in a minor manner. From the ROC curve of the LR model in Figure 4.6a, we observe no significant difference, by including or excluding the wear flags, for both the baseline and advanced models. We observe similar effects for all models. From the PR curve, displayed in Figure 4.6b, we observe a slight difference in the PR curves. We observe that not using wear flags at most points in the curve would benefit the advanced features, while for the baseline, including wear flags is better.



(a) ROC curve (b) PR curve

Figure 4.6: Comparison of inclusion and exclusion of wear time flags, advanced LR.

These results can also be viewed in Table 4.3, where we observe that including wear time flags in the majority of cases is beneficial. All the other ROC and PR curves are available in Appendix E.5. Additionally, it is important to denote that we observe an issue when not including wear time flags. We observe from the LR coefficients, as displayed in Figure 4.7 that the model without wear flags puts more focus on an unreliable feature, the sedentary time. Hence, without wear flags, having much sedentary time would result in being more likely to be in the low-risk group, which is odd. With wear flags, we observe that the sedentary feature is the only positive coefficient. This then indicates that the more sedentary time a participant has, the more likely they are in the high-risk group. Additionally, the moderate count, which is often seen as the most critical variable [97]

has the largest negative coefficient with wear flags, indicating to be the most important feature for being low-risk when wear flags are included. Hence, the coefficients with wear flags, based on literature, are more logical. It is also interesting to observe that vigorous activity seems to have little effect in both models. We reason that this has little effect because it happens very sporadically. This is also what we observed from the average vigorous count in Table 3.4. We observe, in general, that there is some variability between the coefficients; however, there are no significant deviations.

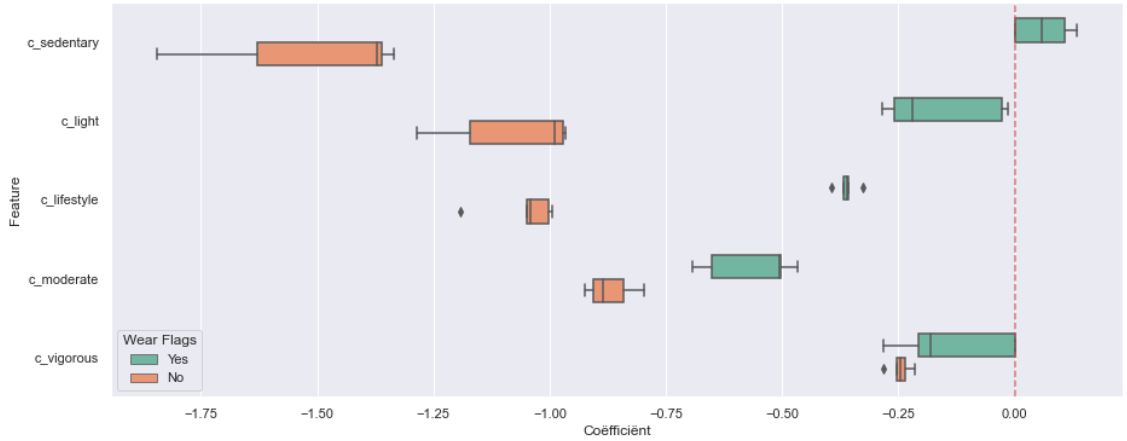


Figure 4.7: Coefficients LR model with and without wear flags.

Conclusion We conclude that, given that our wear time algorithm defines nonwear periods faster than some other versions, that the impact of wear time flags is minimal. Hence, the different versions used, as discussed by Tudor-Locke et al. [115] will likely be similar in terms of performance. We thus expect the results from researchers not using a wear time algorithm to differ little in terms of performance. We do, however, advise the usage of a wear time algorithm, as it is a simple method to allow the model to perform slightly better and rely on a more reliable feature. Lastly, it would be best to optimize the wear time algorithm to observe which settings are best and most reliable for the models. However, given that we will also be using other techniques in this work, optimizing this algorithm would take too much time, and we expect the improvement to be marginal. We have provided the wear time algorithm as a package, for Philips, visible in Appendix A.2. This package allows Philips researchers to experiment with different algorithm versions and easily observe the difference in performance when using different settings.

	Logistic Regression		**Support Vector Classifier**		**Random Forest**		**Decision tree**	
Baseline	With flags	Without flags	With flags	Without flags	With flags	Without flags	With flags	Without flags
AVG AUC	**0.767**	0.766	**0.764**	0.763	**0.783**	0.772	**0.765**	0.753
AVG BS 0	0.176	**0.16**	**0.077**	0.079	**0.142**	0.148	**0.086**	0.088
AVG BS 1	**0.219**	0.238	0.381	**0.377**	**0.246**	0.247	**0.362**	0.371
AVG auPR	**0.513**	0.512	**0.515**	0.512	**0.524**	0.514	**0.505**	0.481

	Logistic Regression		**Support Vector Classifier**		**Random Forest**		**Decision tree**	
Advanced	With flags	Without flags	With flags	Without flags	With flags	Without flags	With flags	Without flags
AVG AUC	0.827	**0.831**	**0.833**	0.828	**0.828**	0.819	**0.777**	0.774
AVG BS 0	0.15	**0.141**	**0.073**	0.074	**0.122**	**0.122**	0.157	**0.155**
AVG BS 1	**0.191**	0.199	**0.318**	0.32	**0.23**	0.238	**0.239**	0.248
AVG auPR	0.575	**0.586**	0.588	**0.589**	**0.573**	0.569	**0.523**	0.498

Table 4.3: Comparison of inclusion and exclusion of wear time flags.

4.3 Normalized Features

In the previous section, we have tested the effect of the nonwear algorithm and observed that including the wear flags resulted in most cases resulted in more reliable results. However, this does mean wear time is affecting the models. In this section, we will be testing the effect of normalized features compared to counts. We want to observe if this allows for a performance increase of the models. Counts calculate the occurrence of a feature happening, compared to a normalized feature, which scales that feature based on the wear time of that participant. This scaling could be necessary because the wear time differs per participant, and thus the amount of counts a person has on a day differs. For example, someone with 800 wear time on a day will have 800 counts divided over the intensity categories, as shown in Table 2.1. However, someone with 600, or 1000 wear time, will have 600 and 1000 counts divided over the intensity categories. We will still count the minutes in each intensity category with normalized features but divide it by the amount of wear time on the day. This division allows us to obtain a percentage, per intensity category, about how much activity was conducted in each category, percentage-wise, compared to simple counts. These percentages should, across the intensity groups, sum to one.

Results Based on the model performances, as displayed in Table 4.4, we can conclude that normalized features are always resulting in worse performance if considering the best model configurations. It is, however, also essential to look at the curves, given that they could tell a different story. We, however, argue that the curves show no significant additional value to the metrics and therefore are displayed in Appendix E.6. For the advanced models, the 'count' curve is generally better at most points in the ROC and PR curves than the 'normalized' curve. It seems that using raw counts is beneficial for the models. The BS, the model's confidence, also appears to be better for counts than normalized features. From a performance perspective, we see no reason for using normalized features compared to counts.

	Logistic Regression		**Support Vector Classifier**		**Random Forest**		**Decision Tree**	
Baseline	Counts	Normalized	Counts	Normalized	Counts	Normalized	Counts	Normalized
AVG AUC	0.767	**0.769**	**0.764**	0.759	**0.783**	**0.783**	0.765	**0.767**
AVG BS 0	0.176	**0.169**	**0.077**	0.078	**0.142**	0.143	**0.086**	0.088
AVG BS 1	**0.219**	0.227	0.381	**0.38**	0.246	**0.245**	0.362	**0.358**
AVG auPR	**0.513**	0.502	0.515	**0.516**	**0.524**	0.519	**0.505**	0.491

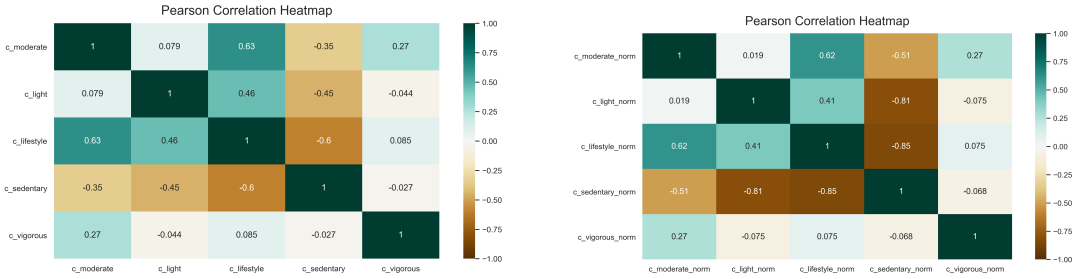
	Logistic Regression		**Support Vector Classifier**		**Random Forest**		**Decision Tree**	
Advanced	Counts	Normalized	Counts	Normalized	Counts	Normalized	Counts	Normalized
AVG AUC	**0.827**	0.817	**0.833**	0.816	**0.828**	0.82	0.777	**0.779**
AVG BS 0	**0.15**	0.156	**0.073**	0.078	**0.122**	0.128	0.157	**0.154**
AVG BS 1	**0.191**	0.195	**0.318**	0.329	**0.23**	0.232	**0.239**	0.241
AVG auPR	**0.575**	0.552	**0.588**	0.553	**0.573**	0.553	**0.523**	0.484

Table 4.4: Comparison between counts and normalized features.

When further investigating the differences between counts and normalized features, we observe a similar conclusion as before. Using normalized features, the model prioritizes the normalized sedentary feature based on the LR coefficients, which can be seen as unreliable. We observe that the normalized sedentary time is seen as a negative coefficient, which would mean that the more sedentary time, the lower the chance of being in the high-risk group. When using counts, the sedentary time is seen as a positive correlation, which is more logical. Although more sedentary time does not have to indicate high risk, long periods sitting without breaks is known to be harmful to our health. Spending a long time in sedentary bouts is also a feature in our advanced model, which is positively correlated to a higher CVD risk based on the LR coefficients in most tests. From the RFC feature importance, in Table E.21, we also observe it can be seen as in important feature in the RFC model.

Based on the heat-maps displayed in Figures 4.8a and 4.8b, which measures the Pearson correlation between the features, we observe that normalized features are more correlated between

each other compared than counts. Although we can not conclude if this is the cause for the performance differences, it could be the reason for the decrease in performance. Furthermore, it is logical if we think about it why the correlation increases for normalized features. We know that, as the intensity increases, the fewer counts we will have, as was also displayed in Table 3.4. Hence, if the sedentary percentage daily is very high, we know that the other intensity categories, such as lifestyle and light, have to be small. Hence, this explains the negative correlation between sedentary time and light and lifestyle. Given that higher intensity counts are less present (MVPA), this will impact the correlation less for those categories. This high correlation is less present for the raw counts, as displayed in Figure 4.8a.



(a) Counts (b) Normalized features

Figure 4.8: Heatmap comparison counts and normalized features.

Conclusion We conclude that we advise using counts, given the performance and the more logical coefficients of the models. Additionally, from a recommendations perspective, minutes spent in an activity group will also be more helpful to recommend to the public. Normalized features make it harder to relate to actual minutes, as the wear time is different per participant. Being able to state that 25 minutes of moderate activity is helpful is more understandable than 0.02 % of moderate activity per day based on that person's wear time.

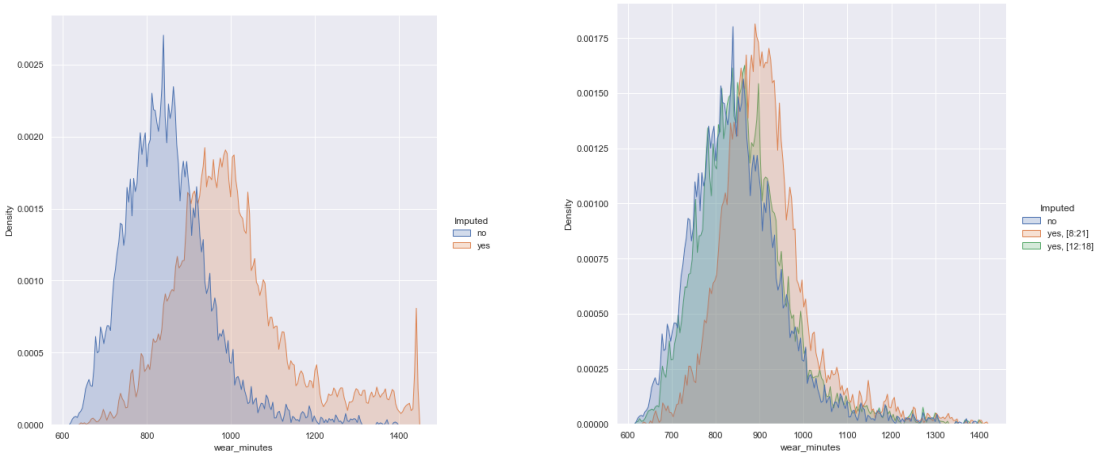
4.4 Imputing Nonwear Time

This section will explore the effect of imputing the nonwear areas in the PA time series. Imputing the nonwear areas could allow for an increase in the predictive performance of the models. It could also remove some bias introduced by the wear time; however, this is difficult to control and check. So, our goal is not to impute missing data, as they are not present anymore in our data after our inclusion and exclusion criteria; however, we want to impute the nonwear areas. As an example, it could be that participants take a walk every day from 1 PM until 2 PM during our lunch break. However, on a particular day, participants could decide to take off the accelerometer during that period and only put it back on at around 3 PM. This scenario would then result in problematic data; however, the wear flags solve this issue by indicating that we are unsure what activity was conducted during that period. It would, however, be better to know the actual values at those periods, as it would paint a complete picture of someone's PA profile. An important assumption we make in this section is that nonwear areas are not equal to sleeping behavior. However, we assume that areas of nonwear are areas where participants took off the device due to, for example, discomfort.

For our imputation strategy, we impute nonwear values by the mean of the other values of that person. If a participant has no data present at a certain point in the day (1 PM), for all valid days, this point will still be classified as nonwear. Otherwise, it will be imputed by the mean of the values at that point. As a concrete example, if a person only has five valid days but values at 1 PM for four days, we will impute the value at 1 PM for the day with the nonwear timestamp, with the average of the other four days. If no values were present at 1 PM for all days, no values

would be imputed for 1 PM. This imputation strategy also separates per week and weekend, given that we find it logical that the activity conducted during the week will be different compared to the weekend. Hence, weekdays will only be imputed by other values from weekdays and weekends only from values of other weekends.

We have applied the imputation to different times of the day, given that we know that there is less wear time at certain areas, as displayed in Figure 3.6, and it may make less sense to impute the entire day. Figure 4.9a displays the difference between wear minutes for no imputation and imputation on an entire day. We argue that there seems to be a heavy increase in wear minutes, especially above 1200 minutes, which we classify as strange. These numbers would indicate that, after imputation, many participants would be active for almost the entire day. We reason that it can thus be less wise to impute during 0:6 and 21:24, given that we know the wear time begins to increase and dip at those points in time. Otherwise, if the participants would wake up early once or stay up late once, it seems they woke up early or stayed up late all the time due to imputation. In Figure 4.9b, the difference between wear minutes for no imputation, imputation between 12 PM and 6 PM and 6 PM and 9 PM is shown. We argue that the increase in wear minutes is more reasonable than the original amount of wear time. Therefore, we reason that these two periods are the most reliable for imputations.



(a) Original and 0-24

(b) Original, 12-18 and 8-21

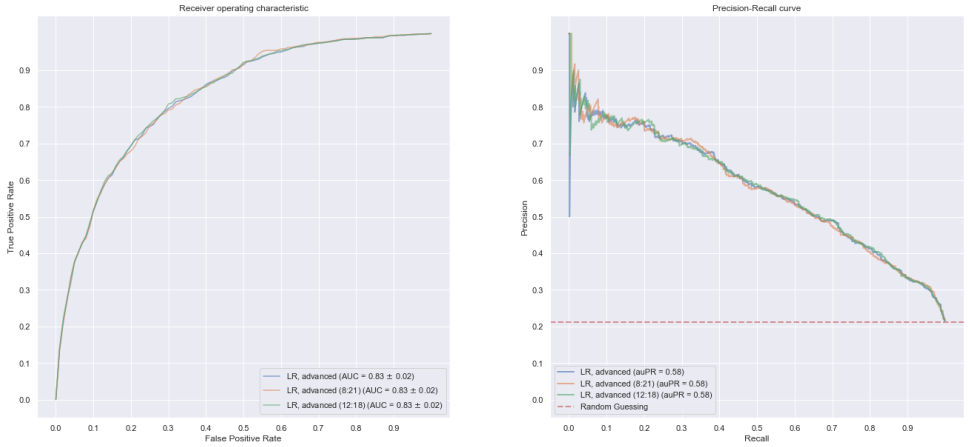
Figure 4.9: Difference in average wear time after imputation. *Original* means no imputation is applied.

Results and Conclusion From the results depicted in Table 4.5, which also display other areas of imputations to emphasize the impact of imputation, and the ROC and PR curves of the LR model in Figures 4.10a and 4.10b respectively, we observe that the results have only changed marginally. The other ROC and PR curves are available in Appendix E.7. We observe that there are sometimes increases in our metrics and ROC and PR curves; however, we argue that the differences are negligible. Furthermore, the most significant increase is when imputing unreliable time-zones and as previously argued, this would result in an undesirable amount of wear time. This could also discriminate against certain participants, as participants who would normally sleep at 9 PM, but stay up once until 11 PM, could obtain a more positive PA profile than in reality. Furthermore, when observing the 10 top % participants with the lowest amount of wear time, we also observed that the difference in performance is negligible. Hence, we conclude that imputing the nonwear areas does not seem beneficial to our models. We also argue that it can create unusual data if done incorrectly, as participants would have entire days of data.

	Logistic Regression							Support Vector Classifier						
Advanced	Original	[0:24]	[8:21]	[0:6]	[6:12]	[12:18]	[18:24]	Original	[0:24]	[8:21]	[0:6]	[6:12]	[12:18]	[18:24]
AVG AUC	0.827	0.826	0.828	0.826	0.827	0.829	**0.83**	**0.833**	0.829	0.831	0.827	0.828	0.831	0.831
AVG BS 0	0.15	0.153	0.148	0.155	0.148	0.148	**0.147**	**0.073**	0.077	0.076	0.076	0.076	0.075	0.075
AVG BS 1	0.191	0.192	0.194	**0.188**	0.195	0.191	0.191	0.318	**0.313**	0.315	0.317	0.316	0.314	0.315
AVG auPR	0.575	0.576	0.577	0.568	0.575	0.577	**0.582**	**0.588**	0.581	0.581	0.581	0.582	0.586	0.587

	Random Forest Classifier							Decision Tree						
Advanced	Original	[0:24]	[8:21]	[0:6]	[6:12]	[12:18]	[18:24]	Original	[0:24]	[8:21]	[0:6]	[6:12]	[12:18]	[18:24]
AVG AUC	**0.828**	0.825	0.824	0.826	**0.828**	0.826	0.825	0.777	**0.786**	0.782	0.779	0.791	0.778	0.783
AVG BS 0	0.122	0.123	0.124	0.124	0.121	**0.12**	0.127	0.157	**0.154**	**0.154**	0.157	**0.154**	0.159	**0.154**
AVG BS 1	0.23	0.23	0.23	0.228	0.232	0.235	**0.226**	0.239	0.233	0.236	0.239	**0.226**	0.242	0.239
AVG auPR	**0.573**	0.572	0.559	0.57	0.571	**0.573**	0.568	**0.523**	0.513	0.504	0.506	0.507	0.5	0.516

Table 4.5: Imputation results on different time-zones.



(a) ROC curve

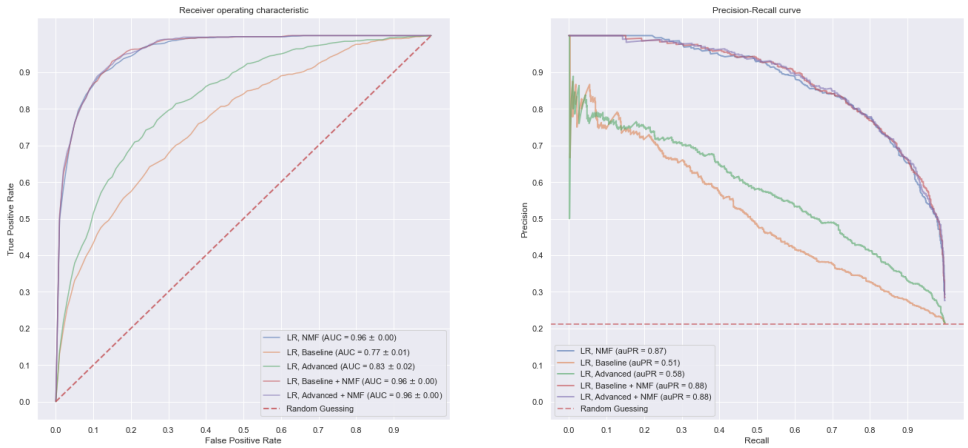
(b) PR curve

Figure 4.10: Imputation comparison Logistic Regression.

4.5 Effect of Adding Risk Factors to the Model

Based on our literature review and several statistics we computed, we observe that age and gender are highly correlated with PA. Hence, we want to test PA's predictive power alongside several variables when included as input to our models. We will test the following configurations, PA with age and gender and with the Non-Modifiable Factors (NMF). The NMF consist of age, gender, and family history of a heart attack before the age of 60. Although some also regard diabetes as a non-modifiable risk factor, given that medicine is available to reduce the negative effect of diabetes, it is often viewed as a modifiable risk factor [35]. We also use a glycohemoglobin % to observe how strong the effect is in our risk formula. Therefore, we classify diabetes as a modifiable risk factor.

Results and Conclusion We observe, that the additional predictive performance of PA alongside NMF, as displayed in the ROC and PR curves in Figures 4.11a and 4.11b respectively, can be seen as negligible. Although we are interested in getting the most value out of the PA data, knowing that the predictive power alongside the NMF is small/negligible, which are easily obtainable factors, is troublesome. The results are also available in Table 4.6. We conclude that adding PA features alongside NMF does improve the PR curve add several points compared to only knowing NMF; however, the additional predictive power of PA alongside NMF is small or can even be seen as negligible. Moreover, the RFC feature importance is visible in Figure E.22. We observe similar feature importance with and without the inclusion of NMF.



(a) ROC curve

(b) PR curve

Figure 4.11: Logistic Regression with non-modifiable risk factors.

	Logistic Regression			**Support Vector Classifier**		
	NMF	Baseline + NMF	Advanced + NMF	NMF	Baseline + NMF	Advanced + NMF
AVG AUC	0.955	**0.957**	**0.957**	0.955	**0.956**	0.955
AVG BS 0	0.069	**0.055**	0.058	0.036	**0.032**	0.035
AVG BS 1	**0.112**	0.12	0.115	**0.177**	0.189	0.182
AVG auPR	0.859	**0.873**	**0.873**	**0.87**	0.869	0.869

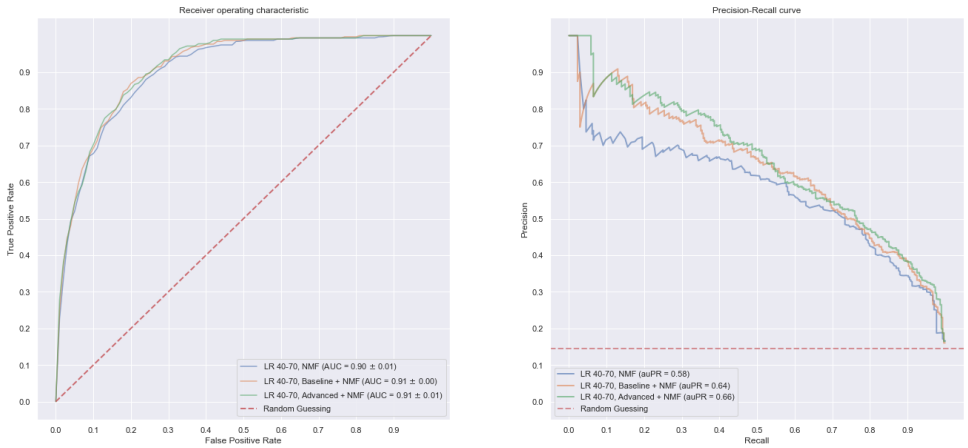
	Random Forest Classifier			**Decision Tree**		
	NMF	Baseline + NMF	Advanced + NMF	NMF	Baseline + NMF	Advanced + NMF
AVG AUC	**0.953**	0.952	0.947	**0.95**	0.944	0.946
AVG BS 0	0.061	**0.056**	**0.056**	0.061	0.063	**0.06**
AVG BS 1	**0.116**	0.13	0.152	**0.119**	0.127	0.127
AVG auPR	**0.865**	0.864	0.842	**0.859**	0.844	0.854

Table 4.6: Performance of PA alongside non-modifiable risk factors

Given these results, we conduct a further investigation into if PA has predictive power alongside the NMF. It could be that the PA effect differs per age group; therefore, we also experimented on the age group 40 till 70. We have chosen this age group, as we observed that PA's predictive power was most noticeable in this sub-group based on the PR curve. We will only show the difference between a model with NMF and NMF with PA to observe the predictive power of PA alongside those NMF.

We observe from the ROC and PR curves, as displayed in Figures 4.12a and 4.12b respectively, that the PA now shows clear additional predictive power when the NMF are known. We observe from the PR curve that at any precision value, picking either the baseline features or the advanced feature set with the NMF would be a better choice than only a model knowing the NMF. Hence, we conclude that knowing the PA features of a person is valuable in addition to the NMF.

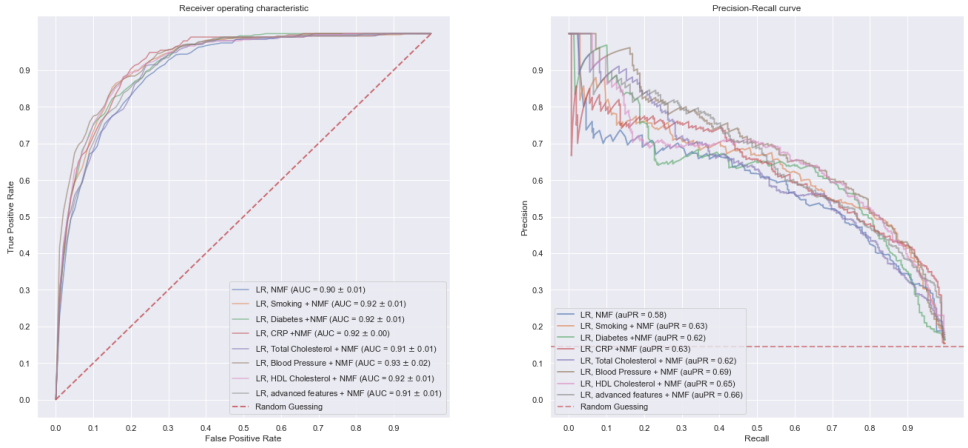
To evaluate the predictive power of PA compared to other variables included in the RRS, we will use the NMF as a baseline and add several factors of the RRS to this baseline. This evaluation enables us to observe the predictive power of PA compared to well-known risk factors, such as smoking and diabetes. From the ROC curve, displayed in Figure 4.13a, we observe that the PA at several points in the curve is equal to the predictive power of knowing other risk factors. From the PR curve, as displayed in Figure 4.13b, we observe that PA contains more predictive power compared to several variables. We observe that the classifier with PA and NMF obtains the highest precision values between recall values 0.2 and 0.4 from any of the curves. We observe



(a) ROC curve (b) PR curve

Figure 4.12: Predictive power of PA alongside NMF, age group 40-70.

that at several recall values that the PA obtains more predictive power than other biomarkers. However, at most points in the curve, knowing the blood pressure it is most valuable, which is also a well-known risk factor for CVD due to hypertension. We thus conclude that PA features are more predictive than well-known risk factors in the age group 40-70, such as diabetes, smoking, CRP, total cholesterol, and HDL-cholesterol in large parts of the PR-curve. In Appendix D we go more into depth about the correlation of PA against individual biomarkers.



(a) ROC curve (b) PR curve

Figure 4.13: Predictive power of other RRS risk factors alongside the NMF factors, in age group 40-70.

4.6 State-Of-The-Art Approaches

In this section, we explore the value of State-Of-The-Art (SOTA) approaches compared to the baseline cut-point analysis. This comparison will include several deep learning models and the MiniRocket approach.

4.6.1 MiniRocket

In our literature review, we conducted that MiniRocket was a SOTA time-series analysis algorithm. It builds further on convolutional network's success while also solving the long-running time that other algorithms have. We have used the implementation from Dempster et al. [32] which is publicly available. In this section, we want to explore the predictive power of MiniRocket, compared to our baseline, the cut-point analysis. As previously discussed, we only use one fold to validate the results from this section onwards.

It is important to note that, in this approach, we have to make use of the entire seven days of data, if they were seen as valid or not. MiniRocket does not require equal length time series as input, which means we can replace the nonwear areas with NaN values. However, we observe a poor performance when replacing nonwear areas with NaN values. Hence, we do not identify the areas of nonwear compared to the cut-point analysis, as this model will incorrectly interpret these areas. However, we investigated the effect of including or excluding the wear flags and observed no significant impact on the performance. We reason that there is no significant impact due to how the MiniRocket approach interprets the nonwear areas. The nonwear areas are indicated by values of -1 in the data. Given that MiniRocket looks purely at the intensity values, we believe the MiniRocket will interpret a -1 value very similar to a 0 value, even if a scalar is applied. Hence we believe the MiniRocket will regard the nonwear areas as sedentary time. Therefore, we have decided to give the approach the raw data.

Lastly, compared to standard machine learning models, it is more difficult to include other variables, such as NMF. For MiniRocket, we identified two methods that we can attempt. First, we can add each NMF as a time series to our original time series. Hence, in this manner, we are creating a multivariate time series (see Definition 2). In this manner, the kernel knows the NMF while going over the intensity values. Another method is simply adding the NMF to the list of features that we obtain from the MiniRocket method. As we also observed during testing, a downside of this approach is that MiniRocket does not know these factors while making the features compared to the multivariate approach. We observed that adding the features afterward resulted in significantly worse ROC and PR curves than in the multivariate variant. We observe that MiniRocket can interpret all needed variables in the multivariate fashion; hence, we use MiniRocket in a multivariate fashion when including multiple variables.

Scaling For the MiniRocket, we apply normalization to the outputted features. Dempster et al. [32] argued that this should work best; however, scaling was not a requirement. As previously mentioned, the features are already outputted on a scale from 0 to 1. Hence normalization, between -1 and 1, will only slightly differ the results. We tested the scaling effect regardless and concluded that scaling the outputted features is slightly better for the ridge classifier. Therefore, as also advised in the paper, we normalize the features. Lastly, transforming the data before obtaining features seemed to have little to no effect; hence, no sampling is applied to the data.

Results and Conclusions

Non-Healthy Participants We have tested the effect of including the non-healthy participants in the training data and concluded that including non-healthy participants is also beneficial for the MiniRocket approach. The ROC and PR curves are available in Appendix Figures E.20a and E.20b respectively.

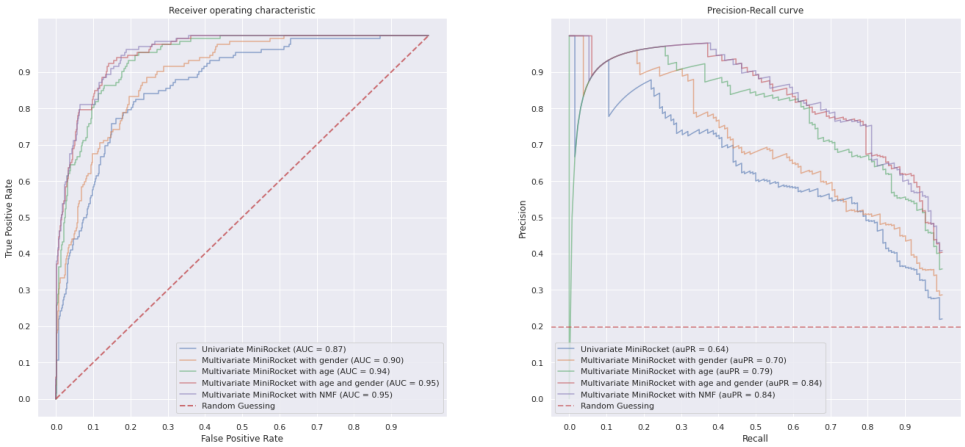
Imputation with MiniRocket We observe that imputation again has minimal effect on the predictive power of the MiniRocket approach. We observed no significant improvements based on the ROC and PR curves, which are available in Appendix Figures E.19a and E.19b respectively.

Univariate and Multivariate MiniRocket Performance Given our previous results, we observe the performance of the MiniRocket method in several settings, namely, solely on the PA

data, together with age, together with gender, together with age and gender, and together with the NMF. We conclude, as seen before, that including the additional variables, drastically boosts the performance, as can be seen from the ROC and PR curves in Figures 4.14a and 4.14b respectively. The results are also displayed in Table 4.7. We observe that MiniRocket performs quite well, even when using other variables. We observe that adding additional variables is comparable to the machine learning methods based on the AUC and auPR. Moreover, it is interesting to observe that the stratified BS, when any variable is included, is very extreme from Table 4.7. We conclude that the model is not confident anymore in predicting high-risk classes when used in a multivariate fashion. When closely inspecting the threshold values, we observe that they are very small. From a clinical point of view, we would like to have a well-calibrated model, which MiniRocket does not achieve. However, it performs quite well from a predicting perspective, based on the ROC and PR curves. In the next paragraph, results in comparison to the machine learning models are presented.

	PA	PA + Gender	PA + Age	PA + Gender + Age	PA + NMF
AUC	0.87	0.896	0.941	0.953	0.954
BS 0	0.137	0.0	0.0	0.0	0.0
BS 1	0.237	0.995	0.994	0.994	0.994
auPR	0.639	0.699	0.786	0.839	0.841

Table 4.7: MiniRocket results with and without NMF alongside PA features.

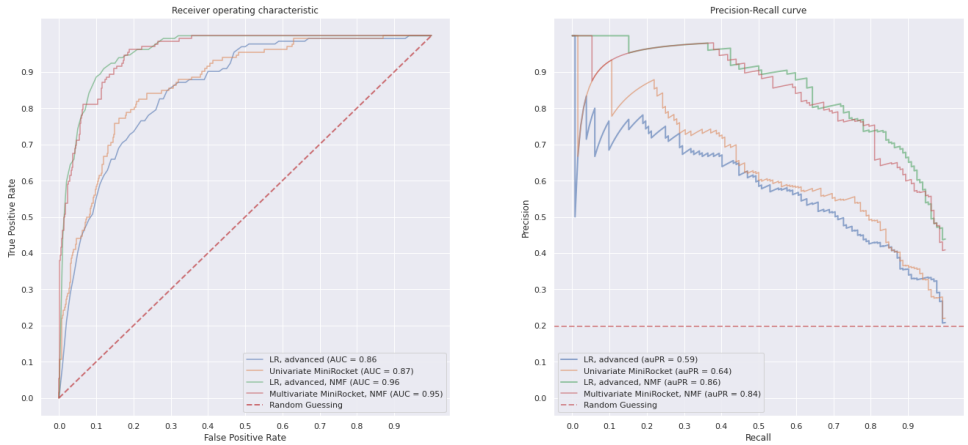


(a) ROC curve

(b) PR curve

Figure 4.14: Univariate and multivariate MiniRocket performance.

MiniRocket and Logistic Regression Comparison Results We reason that the performance between the machine learning models on the cut-point analysis features are very similar between the LR, SVC, and RFC models. Given that we only use the LR in multiple experiments, due to it achieving some interpretable and faster results than the other methods, we decide to use the LR model to compare to the MiniRocket method. The comparison between the LR and MiniRocket in multiple configurations is visible in the ROC and PR curves in Figures 4.15a and 4.15b respectively. A discussion of these results will follow after the deep learning section. In the deep learning section, we will compare the SOTA approaches against the baseline approach and determine which approach obtains the most predictive power.



(a) ROC curve

(b) PR curve

Figure 4.15: Comparison of MiniRocket and LR.

4.6.2 Deep Learning

As previously discussed, we have chosen a wide variety of deep learning models to observe if any of those architectures can obtain more predictive power out of the time series compared to the baseline. Each architecture has been run multiple times, with different settings, such as layers, loss, batch size, and more, to optimize each model for this specific task. We closely investigated the loss on the training and validation data to follow if the data was underfitting or overfitting. We observed that the latter was occurring quite often. The specific architectures of each of the models are visible in the Appendix F. To add other variables to the deep learning models, we use the same approach as the MiniRocket approach and add the variables as a time series, resulting in a multivariate time series. Lastly, similar to the MiniRocket method, we take the raw data and assume that the deep learning models can figure out the nonwear areas themselves.

Loss Function We have experimented with two different loss functions, namely, binary cross-entropy and focal loss. These loss functions are visualized in Figure 4.16, where the blue line follows cross-entropy loss, and the other lines, based on gamma, follow focal loss.

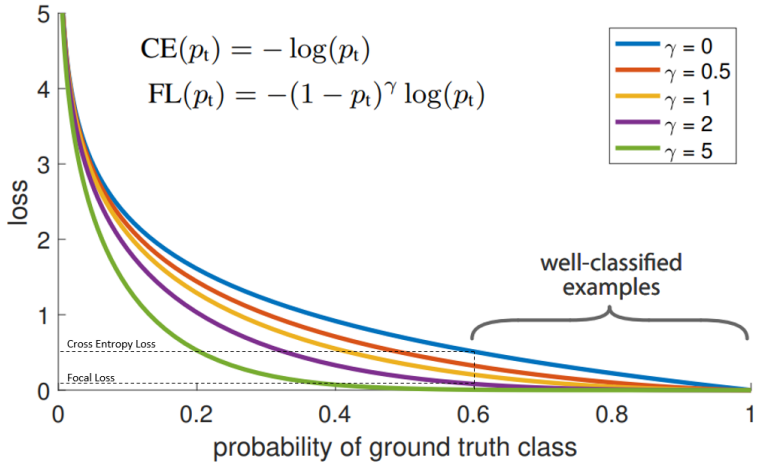


Figure 4.16: Focal loss comparison against binary cross entropy. Image taken from [75].

We obtained slightly better ROC and PR curves during our experiments using the binary cross-entropy than focal loss. The focal loss, as can be seen in Figure 4.16, should, in theory, be better for imbalanced classification tasks. Using binary cross-entropy, even when the model is not super confident, it will already assign a loss to that prediction. For example, at a prediction probability of 0.6, the focal loss would assign almost no loss with a gamma of 5. In contrast, it would assign a loss for cross-entropy because the model is not confident in its prediction. If there is an extreme imbalance, the model could heavily focus on predicting the majority classes correctly, as this could minimize the overall loss of the model. This focal loss can ensure that the model takes more risks in its predictions due to given lower losses at specific probabilities, as can be seen in the curves in Figure 4.16. In practice, to our extent of testing, we experienced that the focal loss did not improve our ROC and PR curves. Similar to before, using the ROC and PR curves is already a good metric for imbalance, given that using the focal loss would only shift the threshold along the curve. In the end, we obtained better results with the binary cross-entropy.

Scaling We are using the StandardScaler to scale the data for the deep learning techniques. We observed that this allowed for faster convergence of the results and resulted in better performance.

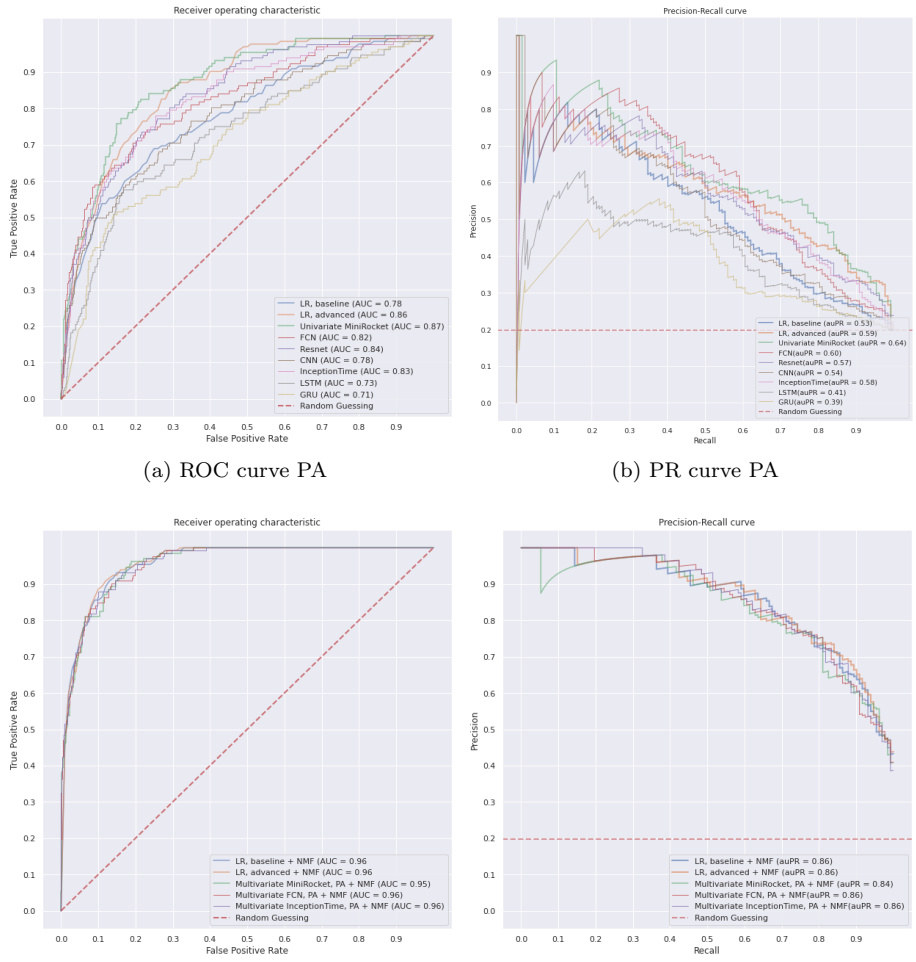
Results The ROC and PR curves are displayed in Figures 4.17a and 4.17b respectively of all the deep learning models and other approaches. Additionally, Table 4.8 shows the AUC, auPR and BS for all methods. When comparing the deep learning models in PR curves, we observe that the FCN is the best performing from 0.25 until 0.6 from all models. Additionally, we argue that the InceptionTime model, based on the PR-curve, is the second-best model out of the deep learning models. The performance is comparable to the ResNet model; however, given its better receptive field, we argue that the FCN and InceptionTime are the two best performing deep learning models. Given that optimizing and running each model and including all the variables will be very costly, we have decided to only run these two models for the upcoming experiments. Hence, when including the NMF we only run those two models, as we observed the best performance for those models when only using PA. The results including the NMF are displayed in Figure 4.17c and 4.17d, for the ROC and PR curves respectively. Table 4.9 also shows the BS and other metrics.

PA only	LR, baseline	LR, advanced	MiniRocket	FCN	ResNet	CNN	InceptionTime	LSTM	GRU
AUC	0.78	0.856	**0.87**	0.838	0.83	0.781	0.83	0.73	0.714
BS 0	0.17	0.147	0.137	0.054	0.061	0.094	0.059	0.096	**0.029**
BS 1	0.212	**0.164**	0.237	0.373	0.347	0.337	0.363	0.361	0.594
auPR	0.526	0.586	**0.639**	0.603	0.57	0.535	0.58	0.412	0.39

Table 4.8: SOTA performance using only PA.

4.6.3 Discussion

First, we will discuss the results when only using PA data. We observe that there are explicit models that perform well and poorly. When we look at the ROC curve, in Figure 4.17a, we observe that the MiniRocket approach has arguably the best ROC curve. Although MiniRocket loses at higher TPR values ($\geq$0.95), we argue that this model performs best compared to the advanced cut-point analysis. We observe that only the GRU and LSTM models perform worse when compared to the baseline (blue line). When taking a look at the PR-curve, displayed in Figure 4.17b, we observe again that these two models perform worse than the baseline. The other models show improvements compared to the baseline at large parts of the ROC and PR curves. Only the simple CNN can be seen as an arguable similar performance compared to the baseline. This performance could thus indicate that a simple CNN does not have enough layers to capture the data's richness. The other convolutional networks have better performance than the simple CNN models, and thus we observe that more complex models are required to capture the richness of the PA data.



(a) ROC curve PA

(b) PR curve PA

(c) ROC curve PA + NMF

(d) PR curve PA + NMF

Figure 4.17: Comparison of SOTA against cut-point analysis technique.

When comparing the baseline approach with the advanced cut-point analysis, we observe, as seen in all previous sections, significant improvements are gained in terms of performance, both in the ROC and PR curves. The advanced features, such as features regarding the time of day activity is conducted, sedentary bouts, and moderate bouts allow for additional predictive power out of the PA data, compared to the baseline. The feature importance can also be seen in Table E.21 for the RFC. Moreover, we thus observe that using the advanced cut-point analysis is more beneficial than using the baseline features. Furthermore, let us dive deeper into the best-performing models compared to the advanced cut-point analysis. The MiniRocket approach performs best in the PR curve at early recall values, 0 till 0.25 and 0.6 till 0.95. The FCN performs best between 0.25 and 0.6 and the advanced LR between 0.95 and 1. Hence, if we purely care about the recall values below 0.9, we would argue that combining the FCN and MiniRocket approach would be best. Depending on the use case, we would then either use the FCN or MiniRocket approach. Thus, we conclude that SOTA approaches better predict cardiovascular risk than the baseline approaches and the advanced cut-point analysis technique when only using physical activity. Additionally, we conclude that RNNs seem less applicable given their poor performance. We speculate that due to RNNs focusing on local dependencies and being unable to track long-term dependencies, they lose sight of the global pattern. Additionally, it could also be that more data is required to train these models; however, this is difficult to conclude.

Including NMF	LR, base	LR, adv	MiniRocket	FCN	InceptionTime
AUC	0.957	**0.958**	0.954	0.955	0.956
BS 0	0.058	0.058	**0.0**	0.037	0.041
BS 1	0.118	**0.113**	0.994	0.181	0.166
auPR	0.859	**0.862**	0.841	0.856	**0.862**

Table 4.9: SOTA performance PA and NMF.

Lastly, we discuss the results, including the non-modifiable risk factors. From the ROC curve in Figure 4.17c, we observe that the differences are marginal and that generally, the cut-point analysis is better. There is one point (0.18-0.22 recall) where the Multivariate-MiniRocket method would be better; however, the cut-point analysis performs better on average. When looking at the PR-curve in Figure 4.17d, we observe that the MiniRocket is rarely a good choice. The MiniRocket model is surpassed at most points in the curves by the other models. However, we observe that the deep learning models, especially InceptionTime, can obtain very high precision values at low recall values ($<$ 0.3) compared to the other methods. Moreover, from a recall value of 0.5 onwards, we reason that using the baseline approach, or the advanced cut-point analysis, would be beneficial. Hence, even when including NMF, we observe small improvements when using SOTA approaches in specific scenarios. However, it is challenging to validate using these SOTA approaches, given that we lose the interpretability of the models and the improvements, which can be seen as marginal. Additionally, the running complexity of these SOTA methods is exponentially longer than the machine learning methods. Hence, if interpretability is vital, we strongly advise the cut-point analysis, in both cases, PA only and PA with NMF. If it is not essential and only used to select people for treatment, we would argue that using the SOTA approaches is beneficial in several cases, as previously discussed.

Chapter 5

Regression Task Experiments

This chapter shows and discusses the results of trying to predict the RRS. Compared to the classification task, predicting the RRS allows for more insight into the actual risk of a patient, which allows a health professional to make more informed decisions on whom to treat. First, we compare machine learning methods on the advanced feature set, given that this performed best based on the classification task. Then, we explore how well SOTA approaches can predict the RRS and discuss an alternative approach to predict the RRS. Lastly, we compare the classification and regression tasks and discuss the best approach.

5.1 Comparison of Machine Learning Methods

In this section, we will make a comparison of the previously mentioned machine learning methods in Section 3.2 and observe which approach can best be used to predict CVD risk based on PA data. We will perform this comparison in two ways, namely, compare the performance purely on the PA alone and together with the NMF. These comparisons allow us to obtain a good representation of the model performance and the predictive power of the PA independently from the NMF.

From the results, as displayed in Table 5.1, we observe that when including the NMF, the RFR has the best metrics. Based on the correlation metric, R2, the RFR model fits the data best and has the lowest MAE. However, it can be argued that when we want to focus on the top 10 % of the highest risk participants, and only PA would be available, using the ADA model would be preferred, given that the MAE is lowest for the top 10 % highest risk participants. However, the correlation metric and MAE are worse for the ADA than the RFR. Given the small differences in the top 10 % MAE between RFR and ADA and the better metrics when including the NMF for RFR, we believe that using the RFR model is preferable. To further investigate the difference, we investigate the BA-diff plots, including PA and NMF.

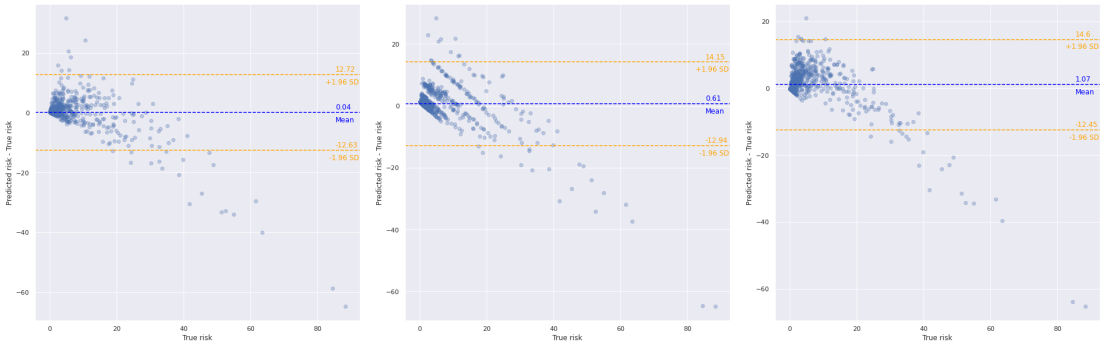
Features	**PA only**			**PA and NMF**		
Model	RFR	ADA	Ridge	RFR	ADA	Ridge
MAE, top 10 %	15.508	**14.925**	15.386	**11.74**	11.817	11.979
MAE	**5.964**	6.453	6.042	**3.042**	3.695	3.92
R2	**0.259**	0.199	0.243	**0.605**	0.546	0.54

* Note: best metrics for each model are **bold**. Maximize R2 and minimize MAE.

Table 5.1: Comparison of machine learning models on regression task.

In Figures 5.1a, 5.1b, and 5.1c we observe the BA-diff plots for the RFR, ADA, and Ridge model respectively. We can observe that the models make different predictions. The ADA model seems to be unable to capture the richness of the data, resulting in prediction lines. We reason that each line stands for a different PA status, and the higher the age, the more we move to the

right of the line. The other two models, RFR and Ridge, have a more spread-out prediction; however, we argue that the RFR BA-diff plot looks slightly better. Although the RFR under-predicts sometimes more heavily than the ridge model, the ridge model seems to under-predict a lot more frequently and higher on average. This argumentation can also be confirmed with the standard deviation lines. Hence, based on the metrics displayed in Table 5.1, and the BA-diff plots, we argue that the RFR model is performing best. We do, however, note that the differences can be seen as marginal, and the differences are difficult to justify. Given that we argue that the RFR model performs best, we will only use the RFR model to improve and compare to other models.



(a) BA-diff RFR (b) BA-diff ADA (c) BA-diff Ridge

Figure 5.1: Machine learning model comparison, BA-diff plots PA and NMF.

Non-Healthy Participants In the previous experiment, we already included the non-healthy participants; however, from Table 5.2, we observe that including the non-healthy participants in the training data improves all our metrics. Hence, we conclude that including non-healthy participants seems beneficial for the performance of our models similar to before.

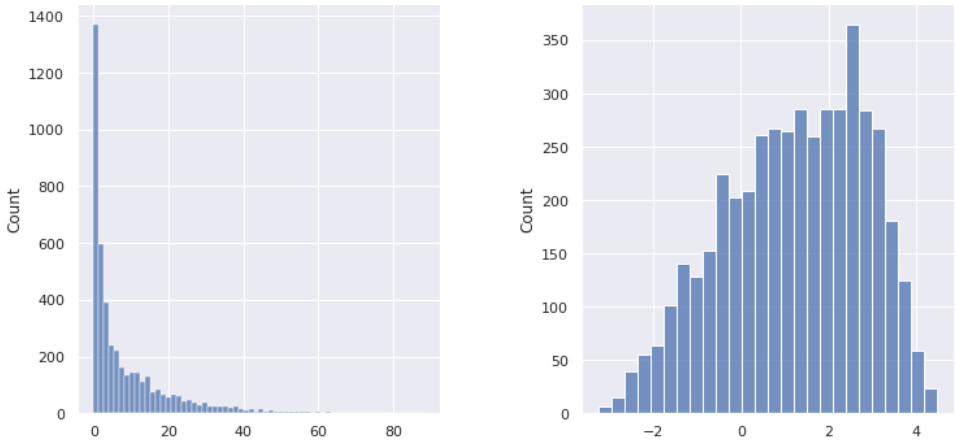
Features	**PA only**		**PA and NMF**	
Training data	Without non-healthy	With non-healthy	Without non-healthy	With non-healthy
MAE, top 10 %	15.921	**15.508**	11.865	**11.74**
MAE	6.07	**5.964**	3.073	**3.042**
R2	0.238	**0.259**	0.587	**0.605**

Table 5.2: RFR with and without non-healthy participants in the training data.

5.1.1 Log-Transform

When we inspect the data more closely, we observe that the data is skewed towards the left. This can also be seen in Figure 5.2a. To deal with this issue, we argue that transforming our predictor variable, the RRS, with a log could cause the data to become less skewed. This result of the log transform on our predictor variable can be seen in Figure 5.2b.

We observe from Figure 5.2b that the data seems more Gaussian distributed after the log transform. This Gaussian distribution could improve the model's performance, due to predicting less close to zero risk. From the results, as displayed in Table 5.3, we observe that the MAE for the top 10 % highest risk has increased significantly for both model configurations. We conclude from these metrics that it is difficult to conclude if transforming the predictor variable helps the performance. However, it appears that applying a log transform causes the top 10 % participants with highest risk to be classified worse, although we would expect those to be better predicted. We argue, however, that this is an artifact of the loss function, which will now punish mistakes at higher risk scores less.



(a) Distribution of RRS

(b) Distribution of log(RRS)

Figure 5.2: RRS distribution before and after log transform on outcome variable.

Features	**PA only**		**PA and NMF**	
Outcome variable	RRS	Log(RRS)	RRS	Log(RRS)
MAE, top 10 %	**15.508**	19.853	**11.74**	13.879
MAE	5.964	**4.55**	3.042	**2.936**
R2	0.259	**0.283**	**0.605**	0.566

Table 5.3: RFR with and without log transform on RRS.

The BA-diff plots, as displayed in Figures 5.3a and 5.3b, for the RRS and log(RRS) as outcome variable, respectively, do not differ significantly. However, we notice that the log-transformed RFR under-predicts on average by 1 %, while the original model slightly underpredicts with 0.04 %. We observe that the log-transformed RFR model also under-predicts more heavily for higher risk, which was also displayed in Table 5.3. The advantage of the log-transformed model is that the over-predicts less at lower risk groups. We also observe this from the standard deviation line. We argue that the BA-diff plots are comparable and that log-transforming the predictor variables

(a) BA-diff RRS

(b) BA-diff log(RRS)

Figure 5.3: RFR BA-diff plots with PA and NMF on the RRS and log(RRS).

does not solve the under and over-predicting of participants. We even argue that the metrics are slightly better for the non-log transformed models. Using the log-transformed models, the prediction of higher risk participants gets significantly worse, especially for PA only, where the MAE of the top 10 % increases from 15.5 to 19.9. Hence, we conclude that transforming the predictor variable, at least for the RFR model, does not solve the under and over-predicting of certain participants. Depending on the use case, it could be interesting to use the log-transform; however, we reason that given the significant increase in MAE for those high-risk participants, it is not worth it. However, we want to emphasize that the differences can still be seen as marginal, especially when inspecting the BA-diff plots.

5.2 State-Of-The-Art Approaches

In this section, we explore the predictive power of SOTA approaches compared to the cut-point analysis. This comparison will include the MiniRocket approach, the FCN, and the InceptionTime model. We argued that these models obtained the best results based on PR curves in the classification task. We reasoned that the FCN and InceptionTime Model performed best among the tested deep learning models in the classification task. Optimizing and running each deep learning model would take too much time, hence why only those two are tested. After showing the results for the MiniRocket approach, we show the results for the deep learning models and then discuss the results of these approaches against the baseline and advanced RFR model.

5.2.1 MiniRocket

In this section, we explore the usage of the MiniRocket approach in predicting the RRS. We want to evaluate if MiniRocket can obtain more predictive power from the PA data as the RFR. As before, we test this in two settings, namely, PA only and alongside the non-modifiable risk factors. As input, we take the raw time series, and as before, we do not indicate the wear time areas. We also use normalization in the experiments using MiniRocket, and we now use the Ridge model instead of the RidgeClassifier model on the outputted features of MiniRocket.

Nonwear The MiniRocket method has an appealing feature that we can use; namely, it does not assume the time series are of equal length, allowing us to remove the nonwear time from the time series. We replace the nonwear with NaN values, simply indicating no values were present for those timestamps. However, we observe that MiniRocket is unable to make any good predictions using this approach. The model only predicts one risk score, namely 9.573. We observe that the issue is still occurring after multiple tests, using different folds, and setting specific hyperparameters. Thus, we conclude that using this property of the MiniRocket method causes the model to learn 'nothing' from the time series, as it always predicts close to the average risk. The poor performance could be explained by the fact that it is harder for the kernels to learn their weights with NaN values. These kernels could then make suboptimal feature maps, in which obtaining informative features is impossible. The model will then ignore most features and simply predict the average risk it has encountered during training.

Non-Healthy Participants For the MiniRocket method, we have decided to include the non-healthy participants. From Table 5.4, we observe for the first time that including the non-healthy participants has a slight negative impact with only PA data. The differences when using PA and NMF can be seen as marginal. However, we would argue that there is a small improvement when including non-healthy participants, given that the MAE has only increased by 0.001, but the other metrics have improved. To keep the results consistent, we have decided to keep the non-healthy participants in our training data. It is, however, interesting to observe that the performance decreases when including the non-healthy participants in our model when only considering PA data. We observe that the model with non-healthy participants in the training data predicts the higher risk participants better; however, the other participants seem to be predicted less well. We

reason that, when only using PA data, the performance decreases more heavily for people at lower risks, given that the dataset will start to include more people from higher risk groups. Hence, it will make slightly more mistakes at younger ages, where the risk is low, at the cost of predicting the higher risks/ages better. This can also be observed from the BA-diff plots in Figures E.18a and E.18b, for RFR with and without non-healthy participants in the training data, respectively.

Features	**PA only**		**PA and NMF**	
Training data	Without non-healthy	With non-healthy	Without non-healthy	With non-healthy
MAE, top 10 %	17.498	**15.453**	12.522	**11.949**
MAE	**5.151**	5.695	**3.429**	3.43
R2	**0.303**	0.259	0.571	**0.585**

Table 5.4: MiniRocket with and without non-healthy participants in the trianing data.

Log-Transform From the results, we again observe that log-transforming the outcome variable does not provide us with consistent improvements, as displayed in Table 5.5. We again observe that for both model configurations, log transforming the output causes worse MAE values for the higher risk participants. Additionally, the data fit the models better when not transforming the data, based on the R2 statistic.

Features	**PA only**		**PA and NMF**	
Outcome variable	RRS	Log(RRS)	RRS	Log(RRS)
MAE, top 10 %	**15.453**	20.389	**11.949**	13.986
MAE	5.695	**4.772**	3.43	**2.826**
R2	**0.259**	0.151	**0.585**	0.575

Table 5.5: MiniRocket log-transformation of RRS comparison.

The BA-diff plots for RRS and log(RRS) as outcome variables, for PA and NMF, are displayed in Figures 5.4a and 5.4b respectively. We observe that the log-transformed model is over predicting a lot more than the non-log transformed model. This can also be seen by the mean lines and standard deviations lines in the plots. Therefore, we argue that the non-log transformed model is better. A similar conclusion was also drawn from the RFR using the cut-point analysis features.

(a) BA-diff MiniRocket RRS

(b) BA-diff MiniRocket log(RRS)

Figure 5.4: MiniRocket BA-diff plots with PA and NMF on the RRS and log(RRS).

5.2.2 Deep Learning

For our deep learning methods, as previously discussed, we only use the FCN and InceptionTime model, given we concluded they obtain the best performance in the classification task in Section 4.6.2. It would also be interesting to test the other models; however, optimizing and testing all those models would take too much time in the available time. We use the same structure as before, only with different metrics, as discussed in the experimental setup section. Given that log-transforming the output resulted in worse performance for the RFR and MiniRocket, we do not log-transform the output for the deep learning models for consistency reasons.

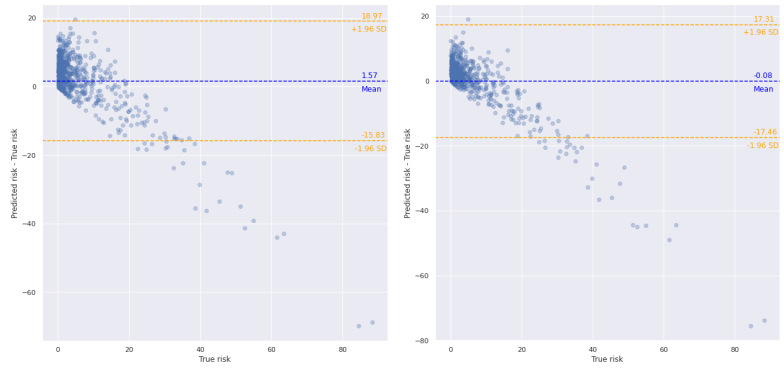
Results

In Table 5.6 the results for our deep learning models, including MiniRocket and RFR models are displayed. Additionally, in Figures 5.5a, 5.5b, 5.5c, and 5.5d are the BA-diff plots for our deep learning models. In the next section we will discuss these results together with the results on our MiniRocket approach.

Features	**PA only**					**PA and NMF**				
Models	Baseline*	Advanced*	MiniRocket	FCN	InceptionTime	Baseline*	Advanced*	MiniRocket	FCN	InceptionTime
MAE, top 10 %	16.411	**15.508**	17.498	16.779	19.565	11.907	**11.74**	12.522	11.832	11.927
MAE	6.332	5.964	**5.151**	6.009	5.209	3.081	3.042	3.429	3.035	**3.015**
R2	0.204	0.259	**0.303**	0.233	0.257	0.596	0.605	0.571	**0.621**	0.619

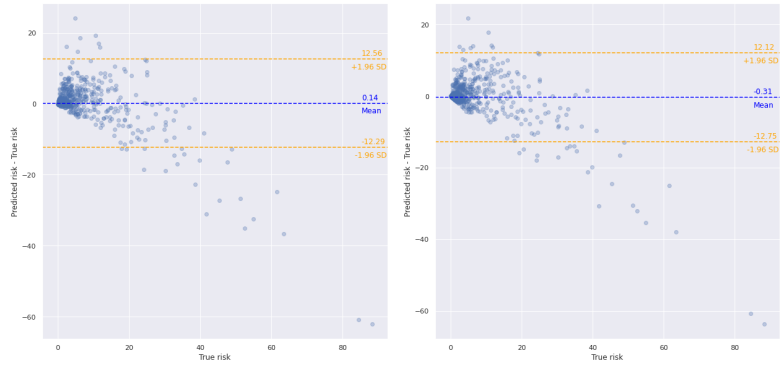
* Baseline or Advanced PA features using the RFR model.

Table 5.6: Regression results on SOTA models.



(a) BA-diff FCN PA only

(b) BA-diff InceptionTime PA only



(c) BA-diff FCN PA and NMF

(d) BA-diff InceptionTime PA and NMF

Figure 5.5: BA-diff plots from the FCN and InceptionTime models.

5.2.3 Discussion

First, we will observe the results when only using PA data. From Table 5.6, we observe that based on the R2 statistic, that the MiniRocket fits the data best. It also has the lowest MAE, with only 5.1 error. This indicates that, on average, the model predicted 5.1 % of risk too high or too low. However, the MiniRocket performance is worse for the high-risk participants based on the top 10 % MAE than other models. We, however, observe a pattern. When the MAE drops, the MAE for the top 10 % will increase. Hence, it is difficult to argue which model performs best. We, however, argue that MiniRocket performs best due to its increase in R2 statistic and a minor increase in MAE for the top 10 % compared to the advanced PA model. We observe that the deep learning methods, compared to the baseline, perform slightly better for certain metrics. However, compared to the advanced RFR model, the deep learning models are not an improvement. For the FCN model, all metrics are worse compared to the advanced RFR model. However, the InceptionTime model has a better MAE while having a significant increase in MAE top 10 %. Compared to MiniRocket, the InceptionTime is still performing worse. Hence, we argue that the MiniRocket method should be used when only using PA to predict the RRS. However, it is difficult to observe a clear winner out of the available models for the regression task, the differences can also be seen as marginal based on the BA-diff plots.

Next, we discuss the results on PA, including the non-modifiable risk factors. From Table 5.6, we observe that the MiniRocket method, although performing well when only using PA data, it now has the worst metrics for all available metrics. Similar to the classification task, it appears that the MiniRocket method does not work well together with other variables in its current configuration. Furthermore, it seems that, based on the correlation metrics, that the FCN and InceptionTime model both provide a better model fit to the data than the RFR model with the advanced features. However, the differences are minimal, and if inspected with the BA-diff plots (Figures 5.3a, 5.5c, and 5.5d), we observe that the differences between those models are marginal. Moreover, given the cost of training and the loss of interpretability that the deep learning models have, we argue that using the cut-point analysis is more beneficial in this case. When this is not important, we encourage the usage of the FCN model due to its slightly better metrics (MAE and R2) than the advanced cut-point analysis model. We, however, again want to emphasize that there is no clear winner from the results.

5.3 RRS Residual of Age and Gender as Outcome Variable

In this section, we explore the predictive power of individual biomarkers alongside the NMF. We are interested in how the predictive power of PA alongside the NMF relates to the predictive power of individual biomarkers. Given the results of this experiment, we explore predicting the risk independently from age and gender, referred to as the RRS residual.

From Table 5.7, we observe the effect of including other biomarkers that were used in the RRS formula as model inputs. We observe that PA brings comparable predictive power compared to certain well-known risk factors such as diabetes, total cholesterol, or HDL-cholesterol for specific metrics. We saw a similar conclusion in the classification task. We reason thus that, although the value of PA seems small in addition to the NMF, the value of knowing actual biomarkers used in the formula is also small. Furthermore, even when using all factors available in the risk score, the RFR model cannot predict each participant correctly. This could be due to the RRS being a complex formula, for which not enough data is available to learn its complexity. Hence, this also makes it more difficult for other factors, such as PA, to give additional power in addition to the NMF. We conclude that age and gender are two very dominant factors, as previously observed. The other risk factors, such as blood pressure or CRP, allow for a more clear additional predictive power in addition to the NMF.

Model inputs	NMF	PA Advanced	Smoking	Diabetes	Glycohemoglobin	Blood pressure	CRP	Total cholesterol	HDL-cholesterol	All RRS factors
MAE, top 10 %	12.624	11.74	11.485	12.54	12.71	10.462	10.922	11.101	11.967	**7.667**
MAE	3.212	3.042	3.115	3.076	3.037	2.463	2.778	3.239	2.968	**1.719**
R2	0.591	0.605	0.666	0.617	0.607	0.742	0.656	0.622	0.619	**0.852**

Table 5.7: Performance of including risk factors as model inputs in addition to the NMF using the RFR. All the models above always include the NMF.

Given the previous results, we will experiment with a different strategy to counteract the dominant age and gender variables, namely, the residual age risk per participant. We will calculate the average risk, per gender, per age, only on the training data, as displayed in Figure 5.6. This setup will create negative and positive residuals, where a negative residual indicates that someone is below the average risk, and a positive residual means someone is above the average risk. Hence, having a good physical health, in an ideal case, would result in negative residuals. We hope this strategy mitigates the dominant effects of age on our models and allows for additional predictive power of PA alongside the NMF.

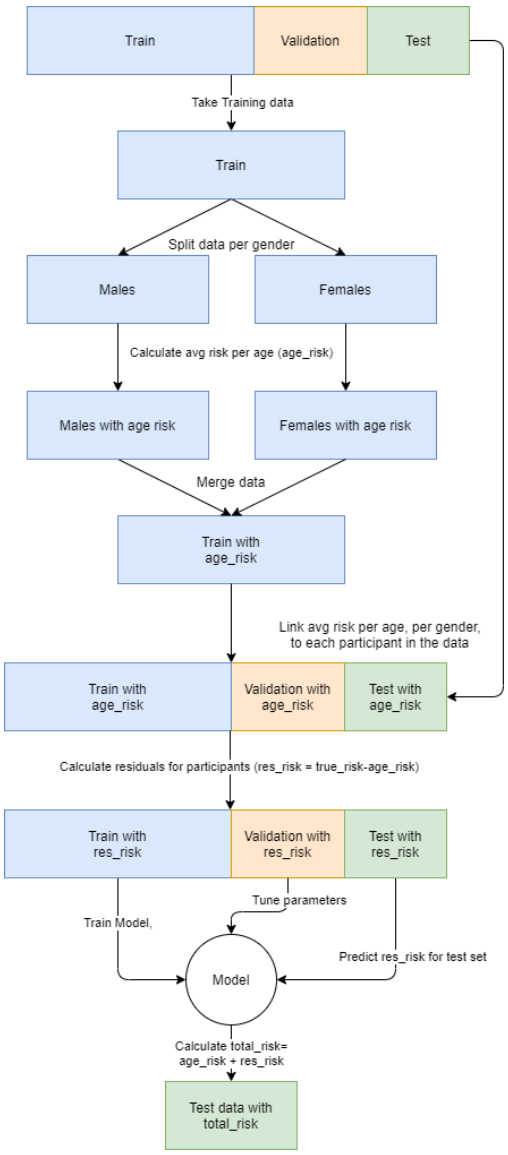


Figure 5.6: Creation of residual RRS based on age and gender.

From Table 5.8, we conclude that this structure does not improve the RFR model. Moreover, we also observe that the MiniRocket approach slightly improves; however, it still does not surpass the performance of the other methods, such as the advanced RFR model, as seen in Table 5.8. Hence, we conclude that this structure does not provide additional predictive power that surpasses the best model. It is also important to recognize that, compared to the average risk, from the residual, the improvement of PA is marginal. It simply appears that, compared to the classification task, the regression task seems a lot more difficult for the models to obtain helpful information out of the data independently from age and gender.

Model, Inputs, Outcome	Average risk from res_risk	RFR, Advanced and NMF, RRS	RFR, Advanced, res_risk*	RFR, Advanced and NMF, res_risk*
MAE, top 10 %	12.308	**11.74**	12.027	12.412
MAE	3.284	**3.042**	3.186	3.149
R2	0.565	**0.605**	0.597	0.601

Model, Inputs, Outcome	Average risk from res_risk	MiniRocket, PA and NMF, RRS	MiniRocket, PA only, res_risk*	MiniRocket, PA and NMF, res_risk*
MAE, top 10 %	12.308	12.522	**12.016**	12.077
MAE	3.284	3.429	**3.255**	3.259
R2	0.565	0.571	0.576	**0.587**

* Outcome variable is res_risk then add age_risk to obtain total_risk as in Figure 5.6.

Table 5.8: RRS Residual RFR and MiniRocket results.

Given that we observed only a small improvement in the previous experiments when using PA, we are interested in how the PA is correlated against the residual RRS (res_risk). We will use a scatter plot between the moderate and res_risk to investigate the correlation. The moderate count is chosen because it is one of the most important features. Against the res_risk, we plot the Z-score of the moderate count per age. This Z-score shows how active, in terms of moderate activity, a participant is compared to peers of the same age. The Z-score is used to counteract the effect of age due to moderate activity also being correlated to age, as seen in Figure D.1. The res_risk is created as displayed in Figure 5.6. The scatter plot, with correlation line, is displayed in Figure 5.7.

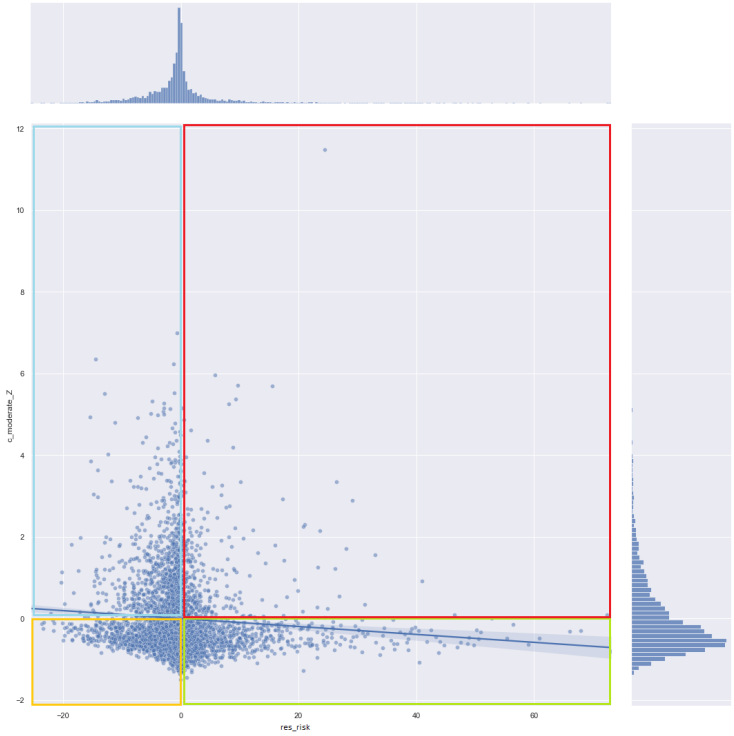


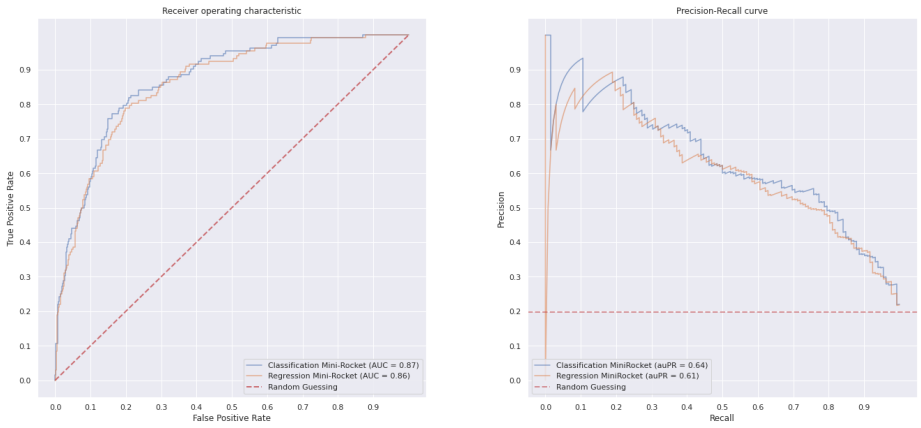
Figure 5.7: z_moderate and age_risk scatter plot

A positive c_moderate_Z indicates that participants were more active than their peers (participants of similar age). Hence, in the red box, displayed in Figure 5.7, are participants who have a positive res_risk, with a positive c_moderate_Z. Although those participants move more on average than their peers, they are still at higher risk than the average risk of their peers. The orange box indicates the opposite, namely participants who move less than their peers but still have a lower risk score. The blue and green boxes are the ideal boxes, where the blue box indicates active participants who also have a lower risk than their peers. Moreover, the green box includes participants who move less than their peers, who also have a higher risk score than their peers. From this plot, we observe a slight trend line, which indicates the less activate someone is, compared to their peers, the higher their risk for CVD compared to their peers; however, there are also many points in the data that trouble the model to make good predictions, independently from age and gender. Due to many points troubling the model, this could explain that PA has little predictive power in addition to age and gender. However, it is essential to denote that this graph only focuses on c_moderate, and models could make different decisions and use multiple features.

5.4 Comparison of Classification and Regression Task

Given our regression results, we are interested in how they relate to our use-case, namely, how well can we identify high-risk people. The classification task only knows if someone is above 10 % risk or below; hence it does not know the difference between someone at 10 % risk or 50 % risk. The regression model does know this, and we are thus interested if this causes the regression model to make a better prediction model. We have decided to focus on our best-performing models on PA data only; hence, we will compare the classification MiniRocket and regression MiniRocket model. To compare these models, we will use the ROC and PR curves.

Results and Conclusion From the ROC curve, displayed in Figure 5.8a, we observe that at most points in the curve, the classification model outperforms the regression model. When looking at the PR curve, displayed in Figure 5.8b, we observe a more significant difference. We observe that the classification model has significant improvements over the regression model. Between recall vales, 0 till 0.1, 0.3 till 0.45, and 0.6 till 0.85, the classification model has significantly better precision values than the regression model. At the other recall values, the performance is comparable. Hence, we observe it is better first to apply a threshold value than training a regression model and then applying a threshold value to the output. Although the regression model should have more information during training, it appears the model cannot create a more valuable prediction model from it.



(a) ROC-curve

(b) PR-curve

Figure 5.8: MiniRocket classification and regression comparison.

Chapter 6

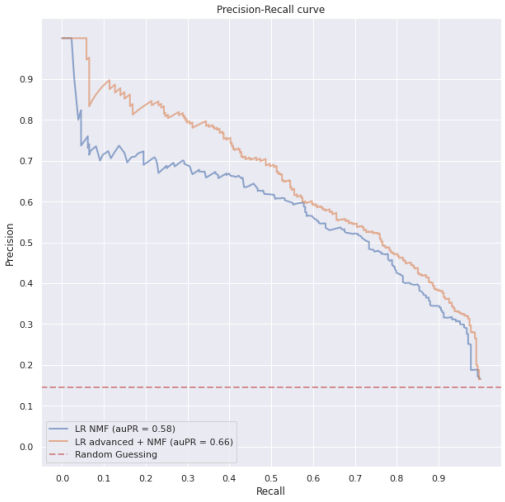
Model to Clinical Value

This chapter will translate our model to clinical value and argue how it can be used in a CVD risk management program in an example. A CVD risk management program aims to find people at high risk for CVD ($> 10\%$) and lower their CVD risk. A selection procedure is used to select people at high risk for CVD based on the organization's specifications. After the selection procedure, each individual is then screened to determine who is genuinely at high risk for CVD, given that the selection procedure tries to estimate the risk. After the selection procedure, personalized treatment is given based on several risk factors of the patient [34]. This personalized treatment can range from PA recommendations, the recommendation to stop smoking, and the application of a treatment, such as hypertension.

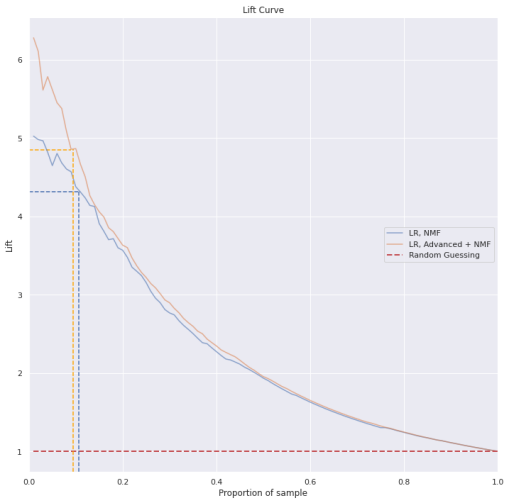
For this use-case, we will only look at the age group 40-70, given that in a real-world setting, we find it more reasonable to focus on a smaller sub-group, where the benefit of CVD risk reduction is most beneficial [34]. In this business example, we will compare the value of a baseline selection approach, only using the NMF (age, gender, and history of heart attack of close relative before the age of 60) against the NMF with the inclusion of PA features. We reason that an interpretable model would be more valuable in this specific case because we are simulating a CVD risk management program. The treatment for a person will be personalized based on several factors. Hence, the knowledge of harmful PA habits can then be of great importance.

Figure 6.1a displays the PR curve for the baseline selection approach (LR, NMF) against the selection approaching, including PA features (LR, Advanced + NMF). A healthcare professional can define their optimal precision and recall point for a selection approach from the PR curve. Using this approach, they can observe the threshold that should be used for the selection procedure. For example, if the precision of 0.8 is desired, the model including PA features would have a larger recall of 0.3 compared to 0.05 of the baseline. Hence, we would be able to select more people at expected high risk of CVD. Hence, using a model that includes PA features would be a better selection approach, as we can achieve higher precision values for each recall value. Additionally, in Figure 6.1b the lift curve is displayed for the baseline and model including PA features. A health professional can use the lift curve to observe the lift ratio, on the y-axis, against the sorted population percentage, on the x-axis, sorted based on the highest probability of being high-risk from the model output, from left to right. The definition of lift is given in Equation 6.1. For example, for the model including PA features, a lift ratio of 4.3 means it has selected 4.3 times the proportion of high-risk participants using 9 % of the proportion sample compared to randomly guessing.

$$Lift = \frac{Ratio\ of\ high\text{-}risks}{Ratio\ of\ high\text{-}risks\ in\ test\ data} \tag{6.1}$$



(a) PR curve



(b) Lift curve, including lift curve selection for 10,000 high risk participants based on lift ratio and proportion of sample

Figure 6.1: Selection procedure comparison.

Using the lift curves displayed in Figure 6.1b, we will sketch a real-world cost-benefit analysis example. In this example, a hospital wants to treat 10,000 patients at high risk for CVD of their hospital, which has a population of 175,000 patients. We assume the hospital has the same low and high-risk distribution as in the PR-curve (1:8). The goal of the hospital is to minimize the costs per prevented CVD event. To obtain 10,000 high-risk patients, we need to select a different lift ratio and proportion samples for each selection procedure. The total amount of selected participants, the average CVD risk among the 10,000 high-risk participants, and the amount of CVD events from the high-risk group based on the average CVD risk are displayed in Table 6.1. We observe that more people need to be selected using the baseline approach to achieve the desired 10,000 risk participants than the selection procedure including PA features. The baseline approach will thus result in more screening cases due to having more false positives than the selection procedure, including PA features. Moreover, the average CVD risk of the 10,000 actual high-risk participants is higher when also PA is used in the selection procedure.

	LR, NMF	**LR, NMF + PA**
Total selected participants [A]	18,550	16,490
Total of high-risk participants	10,000	10,000
Average CVD risk [B]	14 %	15.3 %
Number of CVD events [B]	1400	1530

[A] Based on lift-ratio and proportion sample to obtain 10,00 high-risk participants.

[B] From high-risk selected participants.

Table 6.1: Selection statistics per selection model.

Given the statistics of Table 6.1, we also need to know the costs and efficiency of a CVD risk management program to make a cost-benefit analysis. A cost-benefit analysis will show the additional benefit a better selection procedure has on the overall reduction in costs. We will assume that the costs of the screening, treatment, CVD event costs, and efficiency of treatment based on Table 6.2, which displays the costs and efficiency under three different use-cases based on several sources. Using several sources, we have estimated the costs and efficiency of treatment, which could differ in reality due to various factors. Therefore, we use three different use-cases,

which differ in treatment efficiency, to simulate possible use-cases. Other researchers can use this example with their believed costs. The main intention of this example is to show how a model including PA features can be used.

	Use case 1	**Use case 2**	**Use case 3**
Screening costs ($) [127], [110]	250	250	250
Treatment costs ($) [18], [15], [24]	750	750	750
Efficiency of CVD treatment (%) [34], [116]	10	20	50
CVD event costs ($) [66]	16,000	16,000	16,000

Table 6.2: Assumptions on costs and efficiency for CVD risk management in different use-cases.

Based on Tables 6.1 and 6.2, we can create a cost-benefit analysis which is displayed in Table 6.3. We observe in all use cases that the total costs would be lower when including PA features in the model. We observe that the main differences between the two selection procedures are in the total screening costs and the prevented CVD costs. We have to select fewer participants for the model, including PA features, to obtain our desired 10,000 high-risk participants than only using the NMF. Due to selecting fewer participants, the total screening costs (determining who is actually at high risk for CVD) will be lower. Moreover, the prevented CVD costs also differ between the two selection procedures. The selection procedure that includes PA features has, on average, selected high-risk participants that are of higher risk than the model only considering NMF, which explains the difference in prevented CVD costs. In an optimistic case, where we consider use-case 3, we can observe that a selection procedure that includes PA features would result in a net-benefit gain per prevent CVD events. Using the baseline would result in a net loss per prevented CVD events in all use-cases. When considering any use-case, the cost per-prevented CVD event would still be lower than many hospitals are willing to pay to prevent CVD events given the additional quality of life they provide. The willingness to pay to prevent one CVD event, based on [24], is on average $ 20,000 for one quality of life that is gained. Hence, even when using the baseline model, when hospitals use the willingness to pay system, doing a CVD risk management program is beneficial.

	Use Case 1		**Use Case 2**		**Use Case 3**	
	LR, NMF	**LR, NMF + PA**	**LR, NMF**	**LR, NMF + PA**	**LR, NMF**	**LR, NMF + PA**
Total screening costs ($)	4,637,500	4,122,500	4,637,500	4,122,500	4,637,500	4,122,500
Total treatment costs($)	7,500,000	7,500,000	7,500,000	7,500,000	7,500,000	7,500,000
Total prevented CVD costs ($)	-2,240,000	-2,448,000	-4,480,000	-4,896,000	-11,200,000	-12,240,000
Total costs/revenue ($)	9,897,500	9,174,500	7,657,500	6,726,500	937,500	-617,500
Cost per prevented CVD event ($)	7069.64	5996.41	5469.64	4396.41	669.64	-403.60

Table 6.3: Cost-benefit analysis under different use-cases.

Conclusion We conclude that we have shown the additional benefits of the usage of PA alongside NMF in this business example. Using a selection procedure including NMF and PA features, lower total costs were obtained in all presented use-cases than using a selection procedure based on NMF only. It also allowed for lower costs per prevented CVD event, on average $ 1000, based on the example given. In use case three, we observed that a net benefit could be obtained based on an optimistic efficiency ratio. Hence, based on this example, we hope to have shown the benefit a model using PA can obtain in a CVD risk management program based on several assumptions.

Chapter 7

Conclusions and Future Work

In this work, we have considered the problem of using objective physical activity data to predict CVD risk. Hence, we have defined the following two main research questions to answer in this work:

'Can we predict CVD risk based on one week of univariate objective physical activity data?'

'Which techniques can obtain the most predictive power out of one week of physical activity data?'

7.1 Conclusions

To answer our first research question stated above, we discovered that we could predict low and high-CVD risk well based on the 10 % clinical threshold. Adding features such as the activity conducted at a specific moment, bouts of activity, and maximum MVPA bout have shown to better predict CVD risk in both the classification and regression task than using the baseline features. However, predicting the actual risk is still difficult. Based on the regression task, we observed that the models mostly overpredict people under 20 % risk and underpredict people at higher than 20 % risk, which also occurs when the NMF are included. The cause could be that too little data is available for specific risk scores, especially at higher age groups where the risk is more spread out. Moreover, when comparing the classification and regression task, detecting participants at high risk for CVD works better when treating the problem as a classification task than first predicting the risk and applying a threshold on the predicted risk.

We experimented with more sophisticated approaches, such as MiniRocket and deep learning, to observe which technique can obtain the most predictive power out of the PA data. We observed that we obtained better PR curves than the baseline when only PA data was used on the classification task. However, due to the loss of interpretability of the models, the improvements are difficult to justify. If interpretability is not a concern, we advise combining the MiniRocket and FCN model to select people for treatment. We also reason that RNNs seem to perform worse than CNNs due to RNNs losing sight of the bigger picture and focusing too much on local dependencies.

Besides the main research questions, we also had several sub-goals, including questions regarding wear time, which are unreliable parts of the time series. We observed that the effect of wear time flags on the performance could be seen as marginal. However, we concluded they are essential to include. If they are omitted, the models will mainly use sedentary time to predict CVD risk, which is an unreliable feature. The same conclusion was drawn when using normalized features. The models using normalized features obtained a worse performance but also focused on the unreliable feature sedentary time. Lastly, we observed that imputation harmed the performance of the models for both the cut-point analysis and the MiniRocket method. We reasoned that it did not reliable impute the nonwear areas and thus did not solve nonwear areas.

We experimented with the NMF as model inputs alongside the PA, answering our sub-goals regarding the value of PA, observing that PA still has predictive power independently from these factors. This predictive power was especially noticeable in the age group 40-70, where the PA provided significant improvements alongside the NMF. With a more profound analysis, we observed that the predictive power of PA is comparable to the value of knowing well-known factors such as diabetes, smoking, total cholesterol, CRP, or HDL-cholesterol. Lastly, we presented a cost-benefit analysis example, where we showed that including PA features in a model allows for a better selection tool for a CVD risk management program.

We conclude that objective PA data is a good predictor for CVD risk and provides additional predictive power alongside known NMF. Moreover, when only using PA data, we reason that the MiniRocket approach obtained the most predictive power from the PA data to predict CVD risk.

7.2 Limitations and Future Work

Models We have experimented with a large set of models based on our literature review. However, there are still other models available that could obtain more predictive power out the PA data. For instance, the BOSS model or a CNN-LSTM model. Given time constraints, we are unable to further experiment with more models to evaluate their performance. However, we encourage using other methods to observe if they can obtain the PA data's richness and in a more interpretable manner. Moreover, more data could also increase the performance of the tested models, given that the data is high dimensional (i.e., deep learning models may require more data to properly train their parameters) and in certain age groups the amount of data available is scarce.

MiniRocket Using only the PA data, we have seen that the MiniRocket resulted in the best prediction model, compared to the baseline and deep learning models. However, when including other variables, the MiniRocket method starts to perform worse than the other models. Thus, it seems MiniRocket can obtain the most predictive power out of the PA data; however, the method seems unable to work best for multiple variables. As this was also not part of the main scope of the thesis, it would thus be interesting to observe if this method can be extended to work better in a setting using multiple variables as inputs.

Long-Term Effects Given the structure of the data and the risk score, it seemed less applicable to people of younger ages, specifically below the age of 50. From the results, we would never treat anyone under the age of 50; however, this does not mean those participants should not be advised a different lifestyle. We are unable to measure the long-term effects of being physically inactive from this dataset. Therefore, it would be interesting to conduct a follow-up study that focuses on the long-term effects of physical activity. A higher risk score at a younger age, which is still below the threshold of being treated, could already indicate a worse risk score in the future due to being physically inactive at a younger age for long periods.

Accelerometer From our results, we concluded that we could not correctly predict if certain participants were at low or high risk for CVD. We observed several participants who were more active than their peers but still at a larger risk. It would be interesting to observe if this is also the case for the 2011-2014 PA data of NHANES. We question if the poor relationship for certain participants can be an artifact of the accelerometer used in this dataset. The accelerometer was limited to vertical acceleration and placed on the right hip of the participant, limiting itself to specific movements (e.g., lower body movements). We reason it would be interesting to extend this analysis on a triaxial accelerometer, for instance, the 2011-2014 dataset of NHANES when more information is made available about health outcomes and biomarkers. Perhaps certain age groups engage in unrecognized activities by the accelerometer compared to activities conducted by other age groups. This reasoning could also explain the predictive power in the age group 40-70.

Bibliography

[1] Asthma control questionnaire, accessed on 12 march 2021. https://media.mycme.com/documents/171/w16-11_asthma_q_42563.pdf. 17, 81

[2] Centers for disease control and prevention. national diabetes statistics report: Estimates of diabetes and its burden in the united states, 2014. atlanta, ga: Us department of health and human services; 2014. Available from: https://www.cdc.gov/diabetes/pdfs/data/2014-report-estimates-of-diabetes-and-its-burden-in-the-united-states.pdf. 82

[3] National health and nutrition examination survey, accessed on 15 february 2021. https://www.cdc.gov/nchs/nhanes/index.htm. 19, 22

[4] Scikit-learn machine learning in python, accessed on 23 july 2021. https://scikit-learn.org/stable/. 26, 28, 32, 34

[5] National trends in total cholesterol obscure heterogeneous changes in hdl and non-hdl cholesterol and total-to-hdl cholesterol ratio: a pooled analysis of 458 population-based studies in asian and western countries. *International journal of epidemiology*, 49(1):173–192, 2020. 30

[6] Donna K Arnett, Roger S Blumenthal, Michelle A Albert, Andrew B Buroker, Zachary D Goldberger, Ellen J Hahn, Cheryl Dennison Himmelfarb, Amit Khera, Donald Lloyd-Jones, J William McEvoy, et al. 2019 acc/aha guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 74(10):e177–e232, 2019. 18

[7] Makoto Ayabe, Hideaki Kumahara, Kazuhiro Morimura, and Hiroaki Tanaka. Interruption in physical activity bout analysis: an accelerometry research issue. *BMC research notes*, 7(1):1–6, 2014. 7

[8] Mario Renato Azevedo, Cora Luiza Pavin Araújo, Felipe Fossati Reichert, Fernando Vinholes Siqueira, Marcelo Cozzensa da Silva, and Pedro Curi Hallal. Gender differences in leisure-time physical activity. *International journal of public health*, 52(1):8–15, 2007. 29

[9] Anthony Bagnall, Michael Flynn, James Large, Jason Lines, and Matthew Middlehurst. On the usage and performance of the hierarchical vote collective of transformation-based ensembles version 1.0 (hive-cote 1.0). *arXiv preprint arXiv:2004.06069*, 2020. 8

[10] Philip Barter, Antonio M Gotto, John C LaRosa, Jaman Maroni, Michael Szarek, Scott M Grundy, John JP Kastelein, Vera Bittner, and Jean-Charles Fruchart. Hdl cholesterol, very low levels of ldl cholesterol, and cardiovascular events. *New England Journal of Medicine*, 357(13):1301–1310, 2007. 30

[11] Emelia J Benjamin, Paul Muntner, Alvaro Alonso, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Alexander R Chang, Susan Cheng,

Sandeep R Das, et al. Heart disease and stroke statistics—2019 update: a report from the american heart association. *Circulation*, 139(10):e56–e528, 2019. ii, 1, 82, 83

[12] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994. 9

[13] Elena Boiarskaia. *Recognizing cardiovascular disease patterns with machine learning using NHANES accelerometer determined physical activity data.* PhD thesis, University of Illinois at Urbana-Champaign, 2016. 7, 16, 83

[14] Søren Brage, Niels Wedderkopp, Paul W Franks, Lars Bo Andersen, and Karsten Froberg. Reexamination of validity and reliability of the csa monitor in walking and running. *Medicine & Science in Sports & Exercise*, 35(8):1447–1454, 2003. 6

[15] Michael Brandle, Mayer B Davidson, David L Schriger, Brett Lorber, and William H Herman. Cost effectiveness of statin therapy for the primary prevention of major coronary events in individuals with type 2 diabetes. *Diabetes Care*, 26(6):1796–1801, 2003. 65

[16] Wendy J Brown, Adrian E Bauman, Fiona Bull, and Nicola W Burton. Development of evidence-based physical activity recommendations for adults (18-64 years). report prepared for the australian government department of health, august 2012. 2013. 8

[17] John Burn, Martin Dennis, John Bamford, Peter Sandercock, Derick Wade, and Charles Warlow. Long-term risk of recurrent stroke after a first-ever stroke. the oxfordshire community stroke project. *Stroke*, 25(2):333–337, 1994. 94

[18] Vilma Carande-Kulis, Judy A Stevens, Curtis S Florence, Bonita L Beattie, and Ileana Arias. A cost–benefit analysis of three older adult fall prevention interventions. *Journal of safety research*, 52:65–70, 2015. 65

[19] Robert M Carey and Paul K Whelton. Prevention, detection, evaluation, and management of high blood pressure in adults: synopsis of the 2017 american college of cardiology/american heart association hypertension guideline. *Annals of internal medicine*, 168(5):351–358, 2018. 18

[20] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 36

[21] Tsung O Cheng. Afternoon nap is good for the elderly. *Archives of internal medicine*, 160(5):711–711, 2000. 32

[22] Leena Choi, Zhouwen Liu, Charles E Matthews, and Maciej S Buchowski. Validation of accelerometer wear and nonwear time classification algorithm. *Medicine and science in sports and exercise*, 43(2):357, 2011. 24

[23] Frances Chung, Hairil R Abdullah, and Pu Liao. Stop-bang questionnaire: a practical approach to screen for obstructive sleep apnea. *Chest*, 149(3):631–638, 2016. 17

[24] Margaret Constanti, Christopher N Floyd, Mark Glover, Rebecca Boffa, Anthony S Wierzbicki, and Richard J McManus. Cost-effectiveness of initiating pharmacological treatment in stage one hypertension based on 10-year cardiovascular disease risk: A markov modeling study. *Hypertension*, 77(2):682–691, 2021. 65

[25] Nancy R Cook, Julie E Buring, and Paul M Ridker. The effect of including c-reactive protein in cardiovascular risk prediction models for women. *Annals of internal medicine*, 145(1):21–29, 2006. 18

[26] Nancy R Cook, Nina P Paynter, Charles B Eaton, JoAnn E Manson, Lisa W Martin, Jennifer G Robinson, Jacques E Rossouw, Sylvia Wassertheil-Smoller, and Paul M Ridker. Comparison of the framingham and reynolds risk scores for global cardiovascular risk prediction in the multiethnic women's health initiative. *Circulation*, 125(14):1748–1756, 2012. 17

[27] Marie Therese Cooney, Alexandra L Dudina, and Ian M Graham. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. *Journal of the American College of Cardiology*, 54(14):1209–1227, 2009. 17

[28] Laura Cordova-Rivera, Peter G Gibson, Paul A Gardiner, Heather Powell, and Vanessa M McDonald. Physical activity and exercise capacity in severe asthma: key clinical associations. *The Journal of Allergy and Clinical Immunology: In Practice*, 6(3):814–822, 2018. 16, 81

[29] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. The ucr time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/. 8

[30] Andrew P DeFilippis, Rebekah Young, Christopher J Carrubba, John W McEvoy, Matthew J Budoff, Roger S Blumenthal, Richard A Kronmal, Robyn L McClelland, Khurram Nasir, and Michael J Blaha. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Annals of internal medicine*, 162(4):266–275, 2015. 17

[31] Angus Dempster, François Petitjean, and Geoffrey I Webb. Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020. 14

[32] Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. Minirocket: A very fast (almost) deterministic transform for time series classification. *arXiv preprint arXiv:2012.08791*, 2020. 10, 14, 47

[33] Brett A Dolezal, Eric V Neufeld, David M Boland, Jennifer L Martin, and Christopher B Cooper. Interrelationship between sleep and exercise: a systematic review. *Advances in preventive medicine*, 2017, 2017. 16, 80

[34] Martin Duerden, Norma O'Flynn, and Nadeem Qureshi. Cardiovascular risk assessment and lipid modification: Nice guideline. *British Journal of General Practice*, 65(636):378–380, 2015. 63, 65

[35] Evaluation Expert Panel on Detection et al. Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). *Jama*, 285(19):2486–2497, 2001. 18, 44

[36] Kimberly Fairbrother, Ben Cartner, Jessica R Alley, Chelsea D Curry, David L Dickinson, David M Morris, and Scott R Collier. Effects of exercise timing on sleep architecture and nocturnal blood pressure in prehypertensives. *Vascular health and risk management*, 10:691, 2014. 16, 32, 80, 83

[37] Kathryn Farni, David A Shoham, Guichan Cao, Amy H Luke, Jennifer Layden, Richard S Cooper, and Lara R Dugas. Physical activity and pre-diabetes—an unacknowledged mid-life crisis: findings from nhanes 2003–2006. *PeerJ*, 2:e499, 2014. 16, 82

[38] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019. 11, 12, 13, 16, 105

[39] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. 8, 11, 12, 13, 25, 105, 106

[40] Cheryl D Fryar, Te-Ching Chen, and Xianfen Li. *Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010*. Number 103. US Department of Health and Human Services, Centers for Disease Control and . . . , 2012. ii, 1

[41] Ben D Fulcher and Nick S Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037, 2014. 9

[42] Louisa Gnatiuc, William G Herrington, Jim Halsey, Jaakko Tuomilehto, Xianghau Fang, Hyeon C Kim, Dirk De Bacquer, Annette J Dobson, Michael H Criqui, David R Jacobs Jr, et al. Sex-specific relevance of diabetes to occlusive vascular and other mortality: a collaborative meta-analysis of individual data from 980 793 adults from 68 prospective studies. *The lancet Diabetes & endocrinology*, 6(7):538–546, 2018. 95

[43] Maria Carolina Gongora and Nanette K Wenger. Cardiovascular complications of pregnancy. *International journal of molecular sciences*, 16(10):23905–23928, 2015. 23

[44] Philip Greenland, Sidney C Smith Jr, and Scott M Grundy. Improving coronary heart disease risk assessment in asymptomatic people: role of traditional risk factors and noninvasive cardiovascular tests. *Circulation*, 104(15):1863–1867, 2001. 18

[45] Robert T Greenlee, Taylor Murray, Sherry Bolden, and Phyllis A Wingo. Cancer statistics, 2000. *CA: a cancer journal for clinicians*, 50(1):7–33, 2000. 82

[46] Nicole Gruber and Alfred Jockisch. Are gru cells more specific and lstm cells more sensitive in motive classification of text? *Frontiers in artificial intelligence*, 3:40, 2020. 15

[47] Scott M Grundy, James I Cleeman, C Noel Bairey Merz, H Bryan Brewer, Luther T Clark, Donald B Hunninghake, Richard C Pasternak, Sidney C Smith, Neil J Stone, and Coordinating Committee of the National Cholesterol Education Program. Implications of recent clinical trials for the national cholesterol education program adult treatment panel iii guidelines. *Journal of the American College of Cardiology*, 44(3):720–732, 2004. 17

[48] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–28, 2017. 11, 15

[49] Lu Han, Chongchong Yu, Kaitai Xiao, and Xia Zhao. A new method of mixed gas identification based on a convolutional neural network for time series classification. *Sensors*, 19(9):1960, 2019. 12

[50] Haibo He and Yunqian Ma. Imbalanced learning: Foundations, algorithms, and applications. 2013. 26

[51] National Health and Nutrition Examination Survey. 2011-2012 examination data page, November 2020. https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Examination&CycleBeginYear=2011. 1, 19, 20

[52] National Health and Nutrition Examination Survey. 2011-2012 physical activity monitor - day, November 2020. https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/PAXDAY_G.htm. 78

[53] National Health and Nutrition Examination Survey. 2013-2014 examination data page, November 2020. https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Examination&CycleBeginYear=2013. 1, 19

[54] National Health and Nutrition Examination Survey. 2013-2014 physical activity monitor - day, November 2020. https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/PAXDAY_H.htm. 78

[55] Julia Hippisley-Cox, Carol Coupland, Yana Vinogradova, John Robson, Margaret May, and Peter Brindle. Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study. *Bmj*, 335(7611):136, 2007. 23

[56] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998. 11

[57] SUZI Hong and Joel E Dimsdale. Physical activity and perception of energy and fatigue in obstructive sleep apnea. *Medicine and science in sports and exercise*, 35(7):1088–1092, 2003. 80

[58] Yoichi Izumi, Koichi Matsumoto, Yukio Ozawa, Yuji Kasamaki, Atsusi Shinndo, Masakatsu Ohta, Madet Jumabay, Tomohiro Nakayama, Eise Yokoyama, Hiroaki Shimabukuro, et al. Effect of age at menopause on blood pressure in postmenopausal women. *American journal of hypertension*, 20(10):1045–1050, 2007. 86

[59] Dinesh John, Qu Tang, Fahd Albinali, and Stephen Intille. An open-source monitor-independent movement summary for accelerometer data processing. *Journal for the Measurement of Physical Behaviour*, 2(4):268–281, 2019. 78

[60] Dinesh John, Brian Tyo, and David R Bassett. Comparison of four actigraph accelerometers during walking and running. *Medicine and science in sports and exercise*, 42(2):368, 2010. 20

[61] EF Juniper, PM O'byrne, GH Guyatt, PJ Ferrie, and DR King. Development and validation of a questionnaire to measure asthma control. *European respiratory journal*, 14(4):902–907, 1999. 81

[62] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Lstm fully convolutional networks for time series classification. *IEEE access*, 6:1662–1669, 2017. 15

[63] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. In *Data mining in time series databases*, pages 1–21. World Scientific, 2004. 8

[64] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371, 2003. 9, 11

[65] Eamonn J Keogh and Michael J Pazzani. Derivative dynamic time warping. In *Proceedings of the 2001 SIAM international conference on data mining*, pages 1–11. SIAM, 2001. 9

[66] Meredith Kilgore, Harshali K Patel, Adrian Kielhorn, Juan F Maya, and Pradeep Sharma. Economic burden of hospitalizations of medicare beneficiaries with heart failure. *Risk management and healthcare policy*, 10:63, 2017. 65

[67] Roland Klingenberg. The heart in rheumatic, autoimmune and inflammatory diseases. *European Heart Journal*, 38(40):2985–2985, 2017. 17

[68] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 11

[69] Jeroen Lakerveld, Anne Loyen, Nina Schotman, Carel FW Peeters, Greet Cardon, Hidde P van der Ploeg, Nanna Lien, Sebastien Chastin, and Johannes Brug. Sitting too much: a hierarchy of socio-demographic correlates. *Preventive medicine*, 101:77–83, 2017. 1

[70] Simona Lattanzi, Mauro Silvestrini, and Leandro Provinciali. Elevated blood pressure in the acute phase of stroke and the role of angiotensin receptor blockers. *International journal of hypertension*, 2013, 2013. 94

[71] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014. 11

[72] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. Data augmentation for time series classification using convolutional neural networks. 2016. vii, 13

[73] NY Leenders, W Michael Sherman, HN Nagaraja, and C Lawrence Kien. Evaluation of methods to assess physical activity in free-living conditions. *Medicine and science in sports and exercise*, 33(7):1233–1240, 2001. 6

[74] Andrew Leroux, Junrui Di, Ekaterina Smirnova, Elizabeth J Mcguffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K Urbanek, and Ciprian Crainiceanu. Organizing and analyzing the activity data in nhanes. *Statistics in biosciences*, 11(2):262–287, 2019. 19

[75] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. vii, 49

[76] Jaana Lindström and Jaakko Tuomilehto. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes care*, 26(3):725–731, 2003. 17

[77] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, 2015. 9

[78] Jason Lines, Sarah Taylor, and Anthony Bagnall. Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data*, 12(5), 2018. 8

[79] Michael Littner, Clete A Kushida, W McDowell Anderson, Dennis Bailey, Richard B Berry, David G Davila, Max Hirshkowitz, Sheldon Kapen, Milton Kramer, Daniel Loube, et al. Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002. *Sleep*, 26(3):337–341, 2003. 80

[80] Donald M Lloyd-Jones, Lynne T Braun, Chiadi E Ndumele, Sidney C Smith Jr, Laurence S Sperling, Salim S Virani, and Roger S Blumenthal. Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the american heart association and american college of cardiology. *Circulation*, 139(25):e1162–e1177, 2019. 18

[81] Paul D Loprinzi and Bradley J Cardinal. Association between objectively-measured physical activity and sleep, nhanes 2005–2006. *Mental Health and Physical Activity*, 4(2):65–69, 2011. 8, 16, 80

[82] Paul D Loprinzi, Jonathan Sheffield, Brian M Tyo, and Jeanine Fittipaldi-Wert. Accelerometer-determined physical activity, mobility disability, and health. *Disability and health journal*, 7(4):419–425, 2014. 16, 82

[83] Amy Luke, Lara R Dugas, Ramon A Durazo-Arvizu, Guichan Cao, and Richard S Cooper. Assessing physical activity and its relationship to cardiovascular risk factors: Nhanes 2003-2006. *BMC Public Health*, 11(1):1–11, 2011. 16, 82, 83

[84] Brigid M Lynch, Christine M Friedenreich, Elisabeth AH Winkler, Geneviève N Healy, Jeff K Vallance, Elizabeth G Eakin, and Neville Owen. Associations of objectively assessed physical activity and sedentary time with biomarkers of breast cancer risk in postmenopausal women: findings from nhanes (2003–2006). *Breast cancer research and treatment*, 130(1):183–194, 2011. 82

[85] Steven Mann, Christopher Beedie, and Alfonso Jimenez. Differential effects of aerobic exercise, resistance training and combined exercise modalities on cholesterol and the lipid profile: review, synthesis and recommendations. *Sports medicine*, 44(2):211–221, 2014. 89

[86] Patricia Manns, Victor Ezeugwu, Susan Armijo-Olivo, Jeff Vallance, and Genevieve N Healy. Accelerometer-derived pattern of sedentary and physical activity time in persons with mobility disability: national health and nutrition examination survey 2003 to 2006. *Journal of the American Geriatrics Society*, 63(7):1314–1323, 2015. 82

[87] Izet Masic, Milan Miokovic, and Belma Muhamedagic. Evidence based medicine–new approaches and challenges. *Acta Informatica Medica*, 16(4):219, 2008. 1

[88] Louise C Masse, Bernard F Fuemmeler, Cheryl B Anderson, Charles E Matthews, Stewart G Trost, Diane J Catellier, and Margarita Treuth. Accelerometer data reduction: a comparison of four reduction algorithms on select outcome variables. *Medicine and science in sports and exercise*, 37(11):S544, 2005. 7

[89] Anne McTiernan. *Cancer prevention and management through exercise and weight control*. CRC Press, 2016. 2, 4, 7

[90] Matthew Middlehurst, William Vickers, and Anthony Bagnall. Scalable dictionary classifiers for time series classification. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 11–19. Springer, 2019. 10

[91] Elaine Morrato, Patrick Sullivan, Vahram Ghushchyan, Holly Wyatt, and James Hill. Physical activity and diabetes in us adults: The medical expenditure panel survey 2000-2002. *Diabetes*, 54:A254, 2005. 1

[92] Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover. Exact discovery of time series motifs. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 473–484. SIAM, 2009. 10

[93] Sherry L Murphy, Jiaquan Xu, and Kenneth D Kochanek. Deaths: final data for 2010. 2013. ii, 1

[94] Maisa Niemelä, Maarit Kangas, Vahid Farrahi, Antti Kiviniemi, Anna-Maiju Leinonen, Riikka Ahola, Katri Puukka, Juha Auvinen, Raija Korpelainen, and Timo Jämsä. Intensity and temporal patterns of physical activity and cardiovascular disease risk in midlife. *Preventive medicine*, 124:33–41, 2019. 7

[95] Josh Patterson and Adam Gibson. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.", 2017. 8

[96] Sanne AE Peters, Rachel R Huxley, and Mark Woodward. Diabetes as risk factor for incident coronary heart disease in women compared with men: a systematic review and meta-analysis of 64 cohorts including 858,507 individuals and 28,203 coronary events. *Diabetologia*, 57(8):1542–1551, 2014. 95

[97] Katrina L Piercy, Richard P Troiano, Rachel M Ballard, Susan A Carlson, Janet E Fulton, Deborah A Galuska, Stephanie M George, and Richard D Olson. The physical activity guidelines for americans. *Jama*, 320(19):2020–2028, 2018. 7, 8, 39

[98] Nicolaas P Pronk, Louise H Anderson, A Lauren Crain, Brian C Martinson, Patrick J O'Connor, Nancy E Sherwood, and Robin R Whitebird. Meeting recommendations for multiple healthy lifestyle factors: prevalence, clustering, and predictors among adolescent, adult, and senior health plan members. *American journal of preventive medicine*, 27(2):25–33, 2004. 17

[99] Irina Valeryevna Pustokhina, Denis Alexandrovich Pustokhin, Deepak Gupta, Ashish Khanna, Kannan Shankar, and Gia Nhu Nguyen. An effective training scheme for deep neural network in edge computing enabled internet of medical things (iomt) systems. *IEEE Access*, 8:107112–107123, 2020. 2

[100] Timothy V Pyrkov, Konstantin Slipensky, Mikhail Barg, Alexey Kondrashin, Boris Zhurov, Alexander Zenin, Mikhail Pyatnitskiy, Leonid Menshikov, Sergei Markov, and Peter O Fedichev. Extracting biological age from biomedical data via deep learning: too much of a good thing? *Scientific reports*, 8(1):1–11, 2018. 8, 10

[101] Thanawin Rakthanmanon and Eamonn Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *proceedings of the 2013 SIAM International Conference on Data Mining*, pages 668–676. SIAM, 2013. 10

[102] G Razay, KW Heaton, and CH Bolton. Coronary heart disease risk factors in relation to the menopause. *QJM: An International Journal of Medicine*, 85(2-3):889–896, 1992. 30

[103] MMWR Morb Mortal Wkly Rep. Acsm's exercise is medicine™: A clinician's guide to exercise prescription. *Prev Med*, 39:815–822, 2004. 1, 16

[104] Paul M Ridker, Julie E Buring, Nader Rifai, and Nancy R Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score. *Jama*, 297(6):611–619, 2007. 3, 17, 18, 22, 23, 30, 37

[105] Paul M Ridker, Nina P Paynter, Nader Rifai, J Michael Gaziano, and Nancy R Cook. C-reactive protein and parental history improve global cardiovascular risk prediction: the reynolds risk score for men. *Circulation*, 118(22):2243, 2008. 3, 17, 18, 22, 23

[106] Patrick Schäfer. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015. vii, 10

[107] Yiwey Shieh, Donglei Hu, Lin Ma, Scott Huntsman, Charlotte C Gard, Jessica WT Leung, Jeffrey A Tice, Celine M Vachon, Steven R Cummings, Karla Kerlikowske, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast cancer research and treatment*, 159(3):513–525, 2016. 17

[108] Ahmed Shifaz, Charlotte Pelletier, François Petitjean, and Geoffrey I Webb. Ts-chief: A scalable and accurate forest algorithm for time series classification. *Data Mining and Knowledge Discovery*, pages 1–34, 2020. 8

[109] B Shin, SL Cole, Sang-Joon Park, DK Ledford, and RF Lockey. A new symptom-based questionnaire for predicting the presence of asthma. *J Investig Allergol Clin Immunol*, 20(1):27–34, 2010. 81

[110] Sujha Subramanian, Robai Gakunga, Joseph Kibachio, Gladwell Gathecha, Patrick Edwards, Elijah Ogola, Gerald Yonga, Naftali Busakhala, Esther Munyoro, Jeremiah Chakaya, et al. Cost and affordability of non-communicable disease screening, diagnosis and treatment in kenya: Patient payments in the private and public sectors. *PloS one*, 13(1):e0190113, 2018. 65

[111] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009. 35

[112] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016. 12

[113] Keith M Thraen-Borowski, Keith P Gennuso, and Lisa Cadmus-Bertram. Accelerometer-derived physical activity and sedentary time by cancer type in the united states. *PloS one*, 12(8):e0182554, 2017. 16, 82

[114] Richard P Troiano, David Berrigan, Kevin W Dodd, Louise C Masse, Timothy Tilert, Margaret McDowell, et al. Physical activity in the united states measured by accelerometer. *Medicine and science in sports and exercise*, 40(1):181, 2008. 1, 8, 24, 80

[115] Catrine Tudor-Locke, Sarah M Camhi, and Richard P Troiano. Peer reviewed: a catalog of rules, variables, and definitions applied to accelerometer data in the national health and nutrition examination survey, 2003–2006. *Preventing chronic disease*, 9, 2012. 24, 39, 40

[116] Tjeerd-Pieter van Staa, Liam Smeeth, Edmond SW Ng, Ben Goldacre, and Martin Gulliford. The efficiency of cardiovascular risk assessment: do the right patients get statin treatment? *Heart*, 99(21):1597–1602, 2013. 65

[117] Paolo Verdecchia, Fabio Angeli, and Claudio Cavallini. Ambulatory blood pressure for cardiovascular risk stratification, 2007. 80

[118] Paul A Vodak, Peter D Wood, William L Haskell, and Paul T Williams. Hdl-cholesterol and other plasma lipid and lipoprotein concentrations in middle-aged male and female tennis players. *Metabolism*, 29(8):745–752, 1980. 30

[119] Byron C Wallace and Issa J Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and information systems*, 41(1):33–52, 2014. 27

[120] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019. 11

[121] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017. vii, 9, 11, 12, 105

[122] Qingsong Wen, Liang Sun, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*, 2020. 11, 15

[123] Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998. 17, 22

[124] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956, 2009. vii, 9, 16

[125] Agneta Yngve, Andreas Nilsson, Michael Sjostrom, and ULF Ekelund. Effect of monitor placement and of activity setting on the mti accelerometer output. *Medicine and science in sports and exercise*, 35(2):320–326, 2003. 6

[126] Asija Začiragić. The use of reynolds risk score in cardiovascular risk assessment in apparently healthy bosnian men and women: Cross-sectional study. *Recent Advances in Cardiovascular Risk Factors*, pages 341–358, 2012. 17

[127] Ping Zhang, Michael M Engelgau, Rodolfo Valdez, Stephanie M Benjamin, Betsy Cadwell, and KM Venkat Narayan. Costs of screening for pre-diabetes among us adults: a comparison of different screening strategies. *Diabetes Care*, 26(9):2536–2542, 2003. 65

Appendix A

Python Package for NHANES

A.1 Data Collection Function

The function *combine()* allows the user to combine a set of files effortlessly. This function is handy when combining all files for a data analysis. This function will check if the required files are locally available for the analysis, based on the users' settings. If certain files are not available, it will automatically fetch the files that are not locally available from the NHANES website, which saves the user a significant amount of time having to download each file separately. The user can specify the specific data files to be combined, including the relevant data headers. Moreover, the user can specify which years of data files should be downloaded and combined. Additionally, the user can also specify if the PA should be included in the data frame. Note, this can only be done for year cycles where PA was available. Lastly, there are several options to specify when a participant should be excluded from the analysis based on missing data. These options are further explained in the source code[1].

A.2 Wear Time Algorithm

The function *non_wear(data, year, count_max=2, value_max=100, interval_min=60)* allows the user to generate a wear_flags file. This wear_flags file will output a file that will indicate for each PA time-series, which areas are considered nonwear, and which areas can be considered wear. The user can define for which years to generate wear time flags. The count_max specifies how many consecutive intensities above 0 but below value_max are allowed. Lastly, interval_min specifies the length of an interval to be considered a valid period of nonwear. Further details are available in the source code[1].

A.3 Feature Calculation Function

The function *pa_features()*, allows the user to generate the set of features introduced in Section 3.4. These features are calculated based on the cut-point analysis technique. The user can generate the features based on several settings. These settings include, specification of the length of an allowance interval, length of a bout per activity type, different time settings (e.g., midnight [0:6], afternoon[12:18]), inclusion of wear flags during calculation of features, and more available in the source code[1].

[1] Only available for Philips

Appendix B

2011-2014 NHANES Data

This section explains the 2011-2014 cycles of data in more detail and explains why we have chosen the 2003-2006 data. For the first year of the newer data, namely 2011, participants six years or older were asked to participate, while in the other three years, this was extended to participants aged three and older. Participants were asked to wear a different actigraph for these cycles than in 2003-2006, namely a triaxial actigraph (Model GT3X+ and wGT3X+). Given some concerns in the 2003-2006 cycles, participants found that wearing the device on their hip was irritating and resulted in participants dropping out. NHANES opted for a different placement. During the 2011-2014 cycles, the actigraph was placed on the non-dominant wrist of the participants. For these cycles, the activity count was converted to an 'open source' algorithm, which would make it easier to compare new results for future use. It is, however, not clear how the algorithm works; although it is classified as open-source, no algorithm is available that states how to convert accelerometer counts to MIMS units [59]. An appreciated addition to the MINS units' calculation is that a wear time flag is pre-calculated, which is not present in the 2003-2006 data.

2011-2012 wave A total of 6917 participants were partaking in this cycle, namely 2011-2012. Of the 6917 participants, a total of 3705 participants have a quality flag triggered. These quality flags can be triggered by a long list of events, which are listed on [52]. No rules and decisions are available how to tackle these flags.

2013-2014 wave A total of 7776 participants were partaking in this cycle, namely 2013-2014. Of the 7776 participants, a total of 4266 participants have a quality flag triggered. These quality flags can be triggered by a long list of events, which are listed on [54]. No rules and decisions are available how to tackle these flags.

MIMS unit Unlike the widely used Actigraph activity counts, the MIMS unit is similar to the intensities in the 2003-2006 data however uses a different calculation. The calculation, however, does not allow us to convert the intensities to the MIMS unit or the other way around. MIMS units are less sensitive than intensity counts and focus on depicting visual movements, not vibrations from a car or train, which could account for a bias.

Actigraphs have different dynamic ranges, represented by their range (g) and frequency (Hz), making the summation of activity different per actigraph. The MIMS unit tries to solve this issue by extrapolating certain areas where the acceleration reached the max of the range. It then uses the frequency to determine how to extrapolate the values to make sure different accelerometers can be used in a cross-study. However, as argued, the ± 2 g range, often available in wearables and phones, is still challenging for the MIMS unit due to it being on the lower end of the dynamic range. However, an issue with the MIMS unit is that there are no current cut-points defined to classify when someone falls in a particular intensity group. The author also argues that using cut-point is flawed, and other methods should be used.

Dataset decision

We notice that the 2003-2006 cycles contain a lot more information in terms of medical conditions. Certain laboratory results, or questionnaires questions, are not yet added to the 2011-2014 cycles when writing this document. These, for example, include CRP levels, which is an important indicator for many diseases. Therefore, in terms of indicators for diseases, the 2003-2006 cycles would be better to use, given the richness of variables present. However, although the 2003-2006 is richer, there are still sometimes missing variables, such as sleep apnea, which is only available for the 2005-2006 cycle. Moreover, it is important to denote that results are still being added to the dataset. Therefore, this argument could become invalid at a later stage when more information is added.

We observe that for the first 2003-2006 cycles, the dataset has more changing variable names. This structure makes it more complicated to merge the needed information to conduct an analysis correctly. This merging will take additional time and is therefore also an essential consideration in the decision.

The upside of the 2011-2014 cycles is that the wear time is pre-calculated. However, although it is pre-calculated, it is unclear how the wear time flags are calculated for the 2011-2014 cycles. As we discuss in Section 4.2, there is some ambiguity regarding the wear time algorithm for the 2003-2006 cycles. Many different wear time algorithms are used for these cycles, which would change the results per analysis. Given the difficulty of defining nonwear areas, it would be beneficial to know how the nonwear was calculated for the 2011-2014 cycles to make a better decision if they are calculated appropriately.

Conclusion We have chosen to opt for the older dataset, hence, cycles 2003-2006. We have chosen these cycles due to the availability of more health outcomes and risk factors for health outcomes. It takes several years for all information to be added to the dataset due to processing and making sure the data is valid. This leads us to conclude that the newer data cycles do not contain enough information for analysis at this moment in time. However, this could change soon, as missing data could be added to the dataset. Another issue we observe with the newer dataset is the meaning of the MIMS unit. The calculation of the MIMS unit should be open source; however, the paper addressing this unit does not explain exactly how the MIMS unit is calculated with a specific formula. This ambiguity makes us unable to relate a MIMS unit to a certain activity group. Although this unit allows us to observe wear time, which is a nice property, we are unsure how reliable this is done. Hence, given the richness of the data, and more interpretable results, we have decided to go with the 2003-2006 cycles.

Appendix C

Health Outcomes Review

This section will investigate how the PA data has been used in relation to several health outcomes. This investigation should give us an idea of what features are of most importance and which health outcomes could benefit most from more advanced approaches. This section will also allow us to gain ideas on extending the cut-point analysis with additional features. We will discuss different health conditions and then discuss which health outcomes are of most interest. It is also important to address that the dataset being used, NHANES, is not a longitudinal dataset. Hence, although we have seven days of PA data available, we only have information for the health status of the person available for one point in time, typically on the day the PA starts. Hence, this means we cannot observe the effect of seven days of PA but are looking more at direct relationships and the predictive power of PA.

Sleep disorders Numerous studies have shown that there is a clear relationship between PA and the quality of sleep we obtain [33]. The inability to sleep well is also seen as a pervasive health concern. Loprinzi et al. [81], used the objectively measured PA from NHANES to find the relationship between PA and sleeping variables. A data analytic approach was taken, and Loprinzi used a regression approach to find the relationship between PA with sleep parameters available in the NHANES dataset. These sleep parameters consisted of a total of 20 questions. The PA was analyzed with the usage of a cut-point analysis, with similar thresholds as seen by Troiano et al. [114]. Based on a linear regression model, it was found that there is no difference in the mean hours slept per night for those meeting PA guidelines and not meeting guidelines. The number of hours slept was self-reported. When controlling for age, BMI, health status, smoking status, and depression in a multi-nominal model, only associations were found to feel overly sleepy during the day and having leg cramps.

These findings suggest that frequent participation in PA may improve sleep quality. It has been suggested that the potential beneficial effect of PA on sleep is moderated by the time of day in which PA occurs. The American Academy of Sleep Medicine [79] indicated that vigorous late-night exercise might lead to inadequate sleep hygiene by producing increased arousal. Kimberley Fairbrother et al. [36] further discussed the usage of PA in relation to several risk factors that have a relationship to high blood pressure caused by exercise at certain times of the day. Hypertension is seen as a significant modifiable risk factor associated with stroke, coronary artery disease, and congestive heart failure. The risk of CVD increases as blood pressure rises above a certain threshold. The blood pressure generally increases during exercise and decreases after exercise and during our sleep. The lowering of the blood pressure during the night has been suggested to be a restorative physiological process. Individuals whose blood pressure does not decrease by more than 10% are referred to as "nondippers". Nondipping has been identified in patients with sleep apnea, advanced kidney disease, and nocturia as argued by Verdecchia et al. [117]. Obstructive Sleep Apnea (OSA) is one of the most common sleep disorders, which causes our breathing to repeatedly stop and start during our sleep [57]. We conclude that the time of exercise has a

significant effect on the quality of sleep, which is connected to several risk factors, such as sleep apnea, hypertension, and CVD. Hence, the time of the exercise can be seen as a factor for the reduction of CVD and OSA risk factors.

OSA, however, is a disease that often goes unrecognized, given that it occurs during our sleep. It can be recognized from several symptoms, such as snoring loudly or feeling overly tired. However, the latter can also often be related to other illnesses, and OSA is thus often hard to recognize. Hence, although NHANES has information available regarding if a person has OSA, it is only available for one cycle, while two cycles are recommended due to low prevalence. This prevalence will also be explained later in section 3.1. It can thus often occur that the person has OSA but has not recognized it. Given the low prevalence and potential bias OSA can bring along, we reason OSA will not be the best variable to investigate. We reason that other health outcomes may benefit more from PA. However, it might thus be interesting to extend the cut-point analysis with the time activity is conducted. It is suggested that it is related to several risk factors, such as CVD and hypertension.

Respiratory Respiratory issues, such as asthma, are often researched by different researchers. One of these researchers is Laura Cordova-Rivera et al. [28], which investigated how physical inactivity and sedentary time are related to severe asthma. It was argued that more activity and less sedentary time would be associated with better clinical outcomes and biomarkers of systemic and airway inflammation in people with severe asthma. This study did not use the NHANES dataset; however, it used steps to relate to severe asthma. These steps were related to energy expenditure, similar to the intensity counts of the PA relating to different intensity categories. These steps were independently associated with better exercise capacity in participants with severe asthma. From the results, no significant differences in sedentary time between people with and without asthma were discovered. However, people with asthma conducted an average of 20 moderate minutes less per day. Although asthma thus seems like a good candidate, given that people with asthma conduct less moderate activity, we have an issue. Given that this paper used another dataset, we might not have the necessary information to determine the severity of asthma. We observe the study of Shin et al. [109] which argues about the Asthma Screening Questionnaire; however, we do not have the required information to complete the questionnaire. Additionally, Juniper et al. [61] also developed an asthma risk score, which also has been thoroughly validated. The Asthma Control Questionnaire is available at [1]. This questionnaire consists of seven questions, which would determine the severity of asthma. However, again we observe that the answers given to questions in the NHANES dataset are not rich enough to answer the questions in the questionnaire. These findings indicate that important variables are perhaps not present in the NHANES dataset, which could mean that another health outcome may be more appropriate. Although there seemed to be a connection with PA in the severity of asthma, it will be hard to identify the severity of asthma given the availability of variables in the NHANES dataset. To our length of search into respiratory, it seemed that little research had been done relating PA to respiratory factors for the NHANES dataset. We reason that this is due to the lack of certain variables, and thus another health outcome may be better to use.

Mobility Mobility limitations may relate to difficulty walking or falling, which are often correlated to the elderly. Being able to predict when someone falls or what the symptoms are when someone is going to fall can be of real benefit for health professionals. However, given that only aggregated data rather than raw data is available in the NHANES dataset, this will be difficult. The activity was averaged over each minute, and therefore some intense spikes of activity may have been averaged out. An example is climbing the stairs when an intense reading lasts only a few seconds, followed by a moment of rest. Hence, it could be hard to detect if a person has fallen or is going to fall. Hence, to predict if someone will fall, it would be more appropriate to access the RAW accelerator data instead of the summed data. With the summed data, we are also unable to detect postures or sudden spikes. Hence, we have to relate PA to other mobility issues, such as mobility in general.

Using the NHANES dataset, Loprinzi et al. [82] discovered that people with mobility restrictions participated in more sedentary and light PA and less moderate and vigorous exercise than those without mobility restrictions. Loprinzi's study used a linear and poisson regression model to relate the PA to several risk factors. Based on the models, increasing sedentary behaviors by 60 minutes would increase the ratio of chronic disease by 4 %. Similarly, an increase of 60 minutes in light PA would decrease the ratio of chronic disease by 5 %. We conclude that minimizing sedentary behavior and increasing physical activity among those with mobility limitations may help to improve their health outcomes. Additionally, Manns et al. [86] showed that people with mobility issues generally conduct longer sedentary activities.

Although there seems to be a clear correlation, there is another issue. When comparing mobility issues, comparing people who have and do not have a disability is made. The analysis from Loprinzi et al. [82] indicated that those with mobility limitations had less favorable biomarker levels when compared to individuals without mobility limitations. However, as mentioned in that analysis, care has to be taken, given that people with mobility issues are more likely to be unable to engage in much PA. Hence, the relationship between higher sedentary time may also be an artifact of simply being unable to move due to having issues with moving. Therefore, this will make it especially difficult to relate PA to mobility issues, and another health outcome may be more beneficial. However, we conclude that high sedentary time, in general, seems to result in worse biomarkers.

Diabetes Diabetes affects a wide range of people and is also is the 7th leading cause of death[2]. Finding patterns or detecting when people have diabetes could thus also be of great importance. Physical activity is strongly related to obesity and diabetes, as also argued by Luke et al. [83]. In Luke's study, we observe that the odds ratio for diabetes is 35% lower when considering patients' activity levels. Diabetes is also a risk factor for several diseases or issues, such as weight gain, high blood pressure, and CVD risk. Furthermore, Farni Kathryn et al. [37] argued that the prevalence of pre-diabetes (PD) among US adults had increased substantially over the past two decades. By current estimates, over 34% of US adults fall in the PD category. When controlling for BMI, no significant results were found to relate PA to PD. We conclude that, at least from the NHANES data, there seems to be no significant relationship.

Cancer Cancer is the second leading cause of death among Americans [45]. In the year 2000 alone, an estimated 552,200 Americans died of cancer. Cancer is thus a major burden, and therefore, Thraen-Borwoski et al. [113] suggested that PA may result in enhancements in quality of life and physical functioning in specific cancer types. However, based on statistical analysis, there seemed no significant difference between cancer survivors and people who did not have cancer at each intensity level. It argues, however, that breaks in long periods of sedentary time, among cancer types, could be important. Especially for certain cancer types, this could be interesting to look pattern. However, the only thing shown was that cancer survivors take significantly fewer breaks in sedentary time than do non-cancer patients.

Additionally, Lynch et al. [84] tried a logistic regression method between the NHANES data and breast cancer for women. Findings suggested that both PA and sedentary time accumulated through the day may have implications for breast cancer risk. Independently of both MVPA and sedentary time, accumulating activity in longer bouts was associated with lower waist circumference and CRP, while accumulating sedentary time in longer bouts was associated with higher BMI. It is possible that breaking up sedentary time may be a strategy to help prevent or reduce breast cancer risk in postmenopausal women. We have also seen this previously that longer periods of sedentary time are related to risk factors. We conclude that there are relationships present; however, they could be marginal.

Cardiovascular Many credible sources show that PA improves cardiovascular function and, as an effect, reduces the risk of CVD, as also stated by Benjamin et al. [11] in the heart disease and stroke statistics of 2019. Death from CVD is known to be the highest cause of death in

Americans [11]. Hence, reducing the risk for CVD would benefit many people. Luke Amy [83] uses the objective PA data and made use of cut-points to define the relationship between PA and Cardiovascular risk factors. In this analysis, also the NHANES dataset was used. It was discovered that the activity per gender differs; males are generally more active than females. Negative associations were observed between MVPA and systolic blood pressure, diabetes, hypertension, BMI, obesity, HDL-cholesterol, and blood glucose. These are several risk factors that contribute to the risk for a CVD event. It was, however, argued that Luke Amy was unsure if the result of self-reported PA, and objective PA, was an artifact of inaccurate measurement resulting from the accelerometers. To conclude, there have thus been signs that PA is related to biomarkers of CVD.

Elena Boiaskaia [13] used the cut-point strategy, together with several machine learning approaches, to relate the PA to a CVD risk score. It showed that PA is highly correlated to CVD factors and concluded that the amount of PA, not how and when it is performed, is important for CVD risk classification. However, we are to argue that this would contradict the previous statement of Kimberley Fairbrother et al. [36]. She argued that the time of activity could be of real importance; we also observe that Elena Boiaskaia did not use a metric that showed were on the day activity was conducted, only on the week and weekend. However, Elena showed a clear relationship between PA and CVD risk. However, due to numerous decisions taken and more sophisticated methods available, many improvements could be made, which would show additional interesting insights. Kimberley Fairbrother et al. [36] as also mentioned in Section C associated "nondippers" are more prone to CVD. This fact also shows that additional information can be obtained by adding information regarding the time of day activity is conducted. Together with our data analysis, we conclude that it indeed seems that CVD risk is associated with PA. We conclude that CVD seems highly correlated with PA, and thus seems a good health outcome to investigate further.

Appendix D

Physical Activity Correlation to Biomarkers of RRS

To gain more information, if the risk score can be predicted and the PA is of value, we conduct an experiment correlating individual risk factors of RRS to PA. It is important to note that we have classified biomarkers that are believed to be related to the PA as binary classification tasks. We have classified them as low or high risk, according to Table D.1.

Biomarkers	**Low Risk**	**High Risk**
CRP	<1	$\geqslant 1$
Systolic blood pressure	<140	$\geqslant 140$
Total cholesterol	<240	$\geqslant 240$
HDL-cholesterol male	>40	$\leqslant 40$
HDL-cholesterol female	>50	$\leqslant 50$

Table D.1: CVD factors as individual outcomes

We have attempted to correlate the biomarkers in a regression task; however, we observed that the R2 statistic resulted in 0, indicating it is hard to predict the actual biomarkers as regression tasks. Therefore, we have chosen to treat it as a binary classification task. This section aims to observe if PA is correlated with the biomarkers independently from age and gender, specifically, the moderate activity, given that this is the best indicator in the models. We have also tested the other intensity features; however, we observed they are not as valuable as the moderate values.

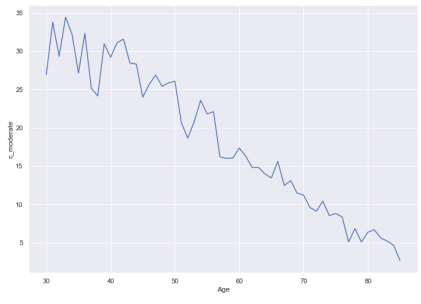


Figure D.1: Relationship moderate activity and age

We observe that the moderate count is highly correlated with age, as displayed in Figure D.1. Hence, to make good conclusions, we must remove the effect of age. To do this, we use lift curves and plot sub-groups of people to make sure the effect of age will be minimal. The x-axis on the lift curve displays the proportion sample, sorted from right to left based on the highest moderate activity or left to right based on highest age, and the y-axis the lift ratio (Definition 6.1).

D.1 Blood Pressure

We observe that there seems not to be a significant difference in blood pressure (BP) between males and females if we were to look at the median value in the boxplot in Figure D.2. However, it is interesting to observe that some of the outliers values from females are quite higher than of males, indicating that some females might experience much higher blood pressures than males. It is given that females, in general, have worse PA profiles than males. We are thus interested if PA is perhaps linked to these outliers values.

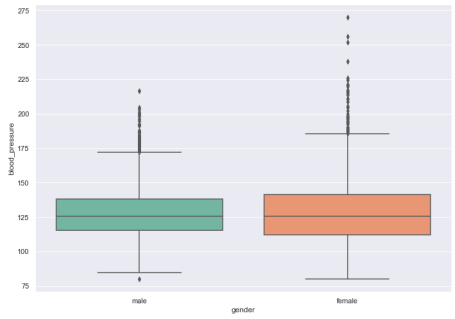


Figure D.2: Boxplot blood pressure per gender

To figure this out, we will classify high blood pressure, in this case, to be above 200. After a close inspection, we conclude that these blood pressure outliers are of older females, with the lowest aged female being 61, with an average age of 75,8. To make it a more fair comparison, we will take false extreme blood pressure of a similar age group and compare the results against the moderate count value to negate the effect of age.

From Figure D.3, it seems that females with very low moderate activity are more likely to have an extreme blood pressure value. The box plot median is lower compared to with no high blood pressure. It is also interesting to observe that females with extreme blood pressure do not condone in much moderate activity. From inspecting these individuals, nothing out of the ordinary was concluded. We conclude, which is also argued in an upcoming section, that females have higher blood values than males, especially at higher ages, resulting in these high blood values due to menopause.

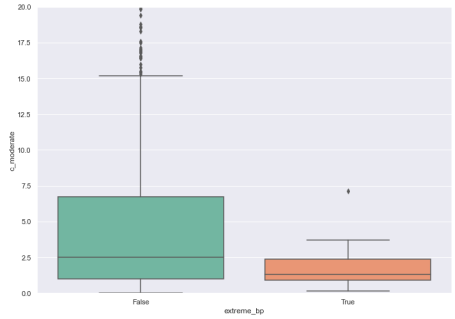


Figure D.3: Boxplot blood pressure females with 200 threshold

After these first insights, we will start looking at the actual binary task relating moderate activity to the high blood pressure group. We observe, that it is very difficult to conclude if moderate count is highly correlated with a high blood pressure, as can be seen in Figures D.4 and D.5. The age is highly correlated, and the box plots do not show us a clear difference. Therefore, we will look into the lift curve.

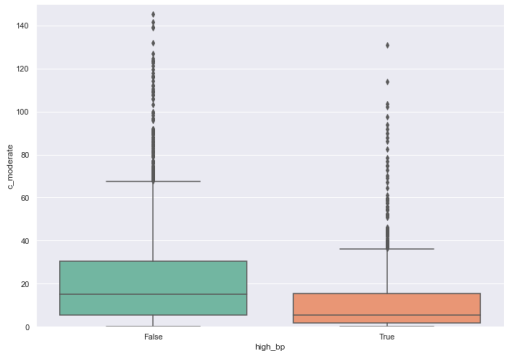


Figure D.4: Boxplot BP against moderate count

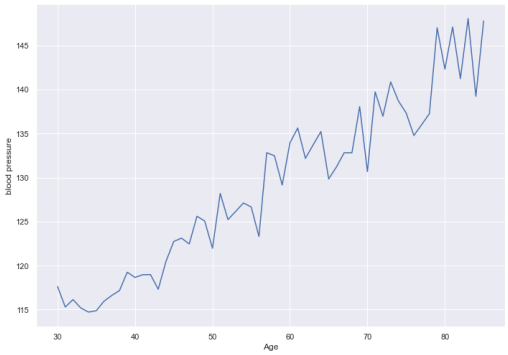
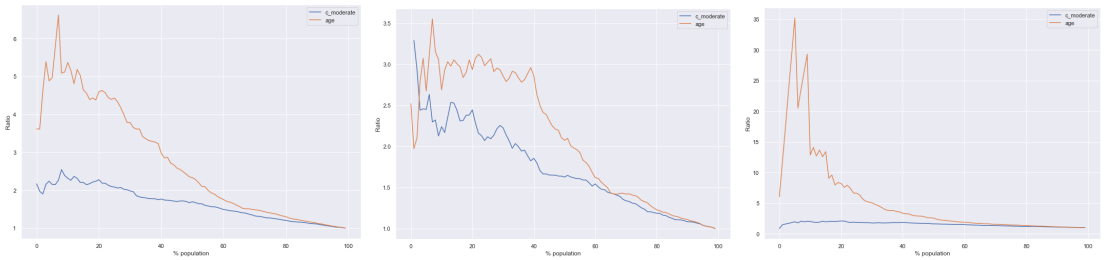


Figure D.5: BP-age correlation

If we look at the entire life plot in Figure D.6a, we would argue that the moderate parameter has some value; however, compared to age, it would be almost nothing. It would be better to use age to select people with high blood pressure than physical activity. Moreover, it is interesting to look at the following plots, which depict the difference between males and females in Figures D.6b and D.6c respectively.

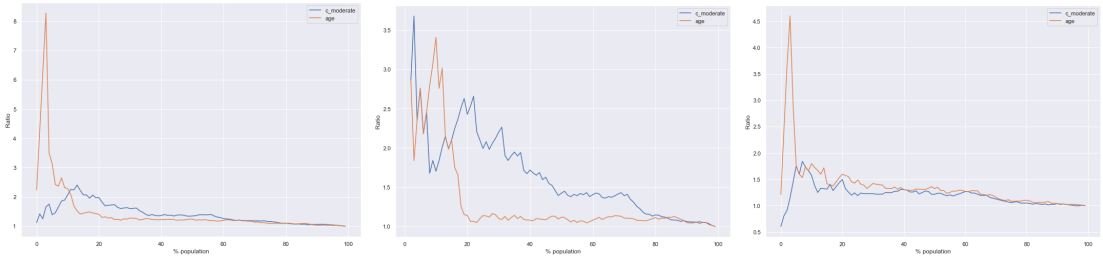


(a) All participants (b) Male (c) Female

Figure D.6: Blood pressure lift curves

We observe that the age variable has a significant impact on having a high blood pressure for females. For the males, this impact is significantly lower. The effect of menopause could explain this impact. Although age is generally a factor of higher blood pressure, as we observed in Figure D.5, menopause is to be believed to shift the effect of hormones and increase blood pressure for females. This would explain why females are less likely to be in the high group at early ages. In higher age groups, females generally have higher blood values than earlier ages due to menopause. The effect of menopause can be reached at young as 30 and as high as 60 [58]. However, in general, it is believed to be between 40 and 50. At 20 percent of the population, females are between 30 and 41. We could thus argue that the effect of menopause can explain the effect observed in the female graph.

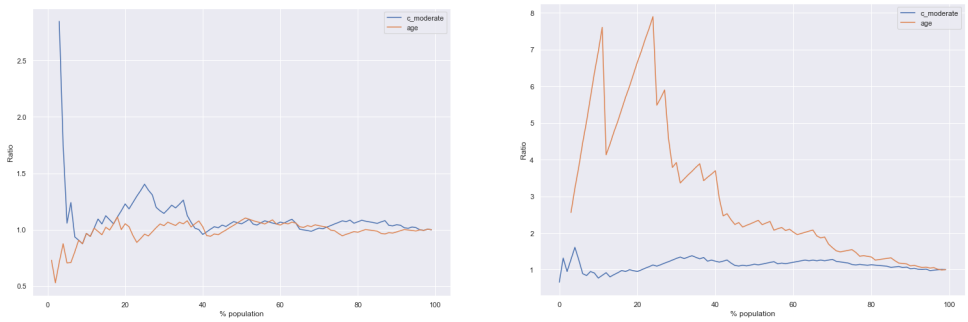
In Figure D.7a, we have plotted the lift curves for the age group between 50 and 60. In this scenario, of the 20 percent of the most moderate people in the subgroup, in the age group 50-60, we observe participants are twice as likely to be in the low blood pressure group compared to the average. Additionally, we could argue that the effect of age is still present to a certain degree, however, it is largely removed. Given that the age groups result in smaller groups, we have a larger bias, explaining the spike at around 3 percent of the population. Given that gender also has an impact, we also plot the results, when only trying plot the results per gender, as depicted in Figure D.7b and D.7c.



(a) All participants (b) Male (c) Female

Figure D.7: Blood pressure lift curve, age group 50-60

In multiple age groups, we observe that moderate activity is more effective in detecting males with high blood pressure than age in specific proportion samples. We could argue that, at lower age groups, the moderate value seems to have less value for both males and females. We observe, the following in the lift curves for age group 30 till 50, as depicted in Figure D.8a and D.8b.



(a) Male (b) Female

Figure D.8: Blood pressure lift curve, age group 30-50

For the age group 30-50, we observe that age still significantly impacts females, and the moderate count parameter has little value. For the males, the impact is marginal; however, we do observe the effect of age in that group is less apparent, and at 20-40 percent of the population, there is some value to be gained with the moderate activity count. The effect of menopause can again explain the effect for females. We conclude that age is highly correlated to high blood pressure, especially for females, due to menopause. Moreover, we observe that for males, moderate activity is correlated to high blood pressure at specific age groups, where the effect of age is mostly negated. Hence, there seems to be a relation between PA and blood pressure independently from age and gender; however, the correlation is small based on most of the lift curves. The most significant difference, based on moderate activity, was observed in age-group 50-60 for the males.

D.2 C-Reactive Protein

The C-Reactive Protein (CRP) values seem not correlated to age, as can be seen in Figure D.9. Given that the age is less correlated to CRP, we could look at the boxplot as displayed in Figure D.10, to gain a good first insight if physical activity, specifically, the moderate count, would be correlated to high risk of $CRP \geq 1$. We observe that there is a clear relationship, between moderate activity and CRP based on the boxplot. This would suggest, that although the differences are not big, there should be some correlation between physical activity and CRP.

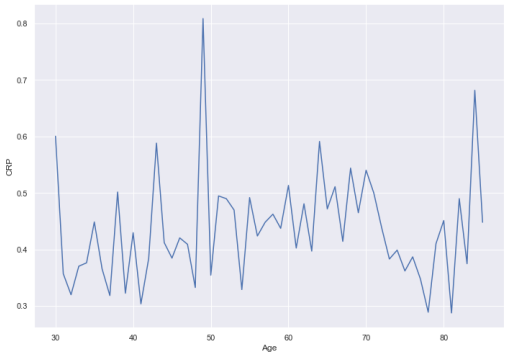


Figure D.9: CRP relationship to age

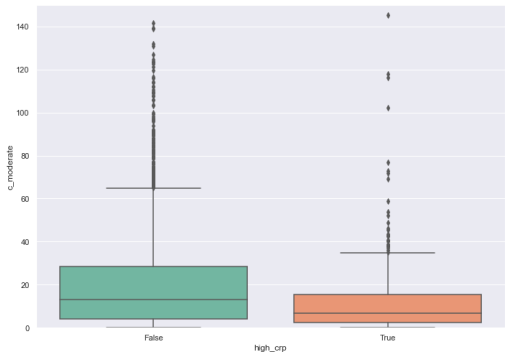
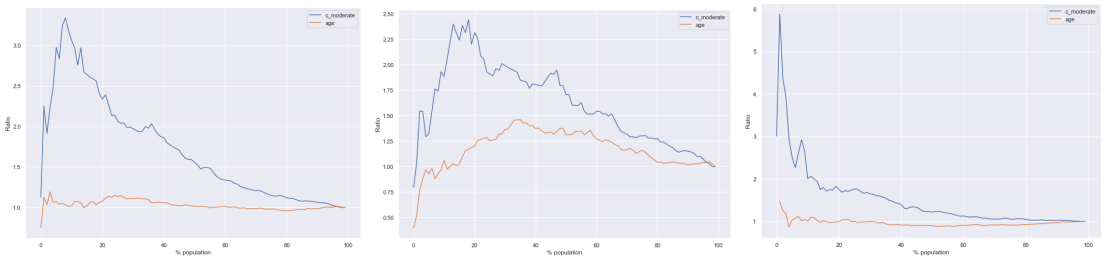


Figure D.10: Boxplot CRP-moderate count

When looking at the lift curve plot, in Figure D.11a, we observe that for CRP, the moderate value can be very valuable. We observe that age almost does not affect the CRP on the lift curve, allowing us to regard the moderate values as an independent variable. We could thus argue, as an example, that at 38 percent of the population, with the lowest amount of moderate activity, participants are twice as more likely to be in the high-risk group for CRP than the average population. This would show that moderate activity has some value for predicting CRP, especially because age seems to be a bad predictor. In Figure D.11b and D.11c are the lift curves for males and females respectively. We observe that still for both males and females, moderate activity is correlated to CRP. We, however, would argue that the difference between males and females is only present at the top 10 % of least activity participants. At that point, we observe a big spike for females, which is not present for males. Furthermore, we argue that the value of moderate activity for males and females to be similar in correlation with CRP. We thus conclude that moderate activity seems well correlated to CRP, independently from age and gender, given the observed lift curves.



(a) All participants (b) Male (c) Female

Figure D.11: CRP lift curves

D.3 Total Cholesterol

We observe that from the boxplot displayed in Figure D.12, that there seems to be no difference in being in the high or low-risk group for total cholesterol based on moderate activity conducted. If we look at the relationship between age and total cholesterol, displayed in Figure D.13, we observe that there seems to be a bell-curve shape present. This relationship, however, is unexpected, and we have no explanation for the bull-curved shape of the curve, which could be an artifact of the dataset.

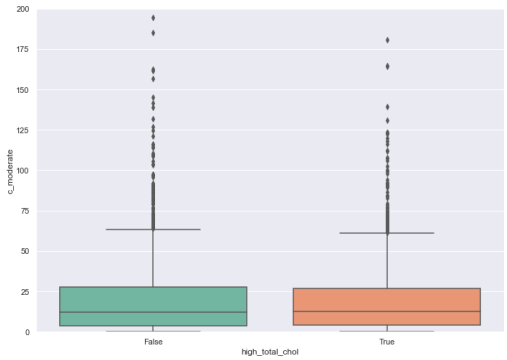


Figure D.12: Boxplot total cholesterol against moderate count

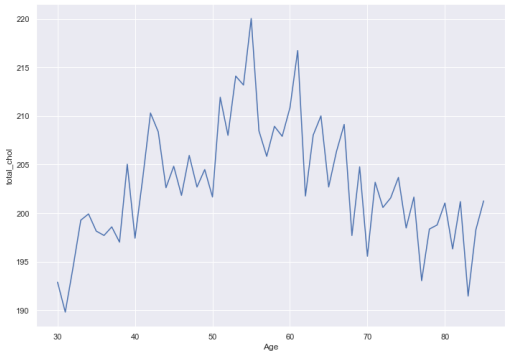
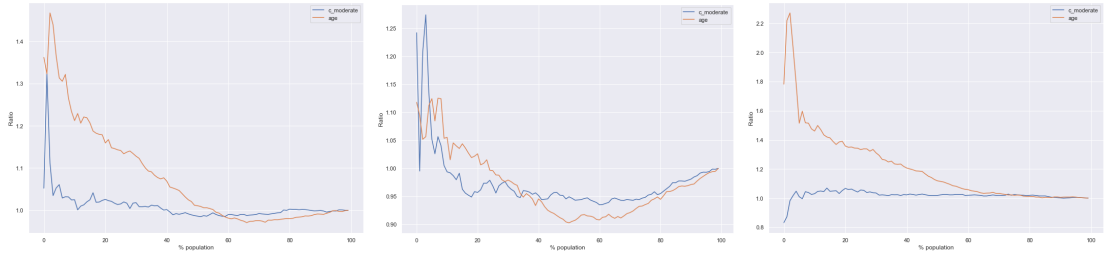


Figure D.13: Relationship total cholesterol and age

From the lift curve as displayed in Figure D.14a, we observe that age does still correlate with total cholesterol, however, only with minimal lift ratios. At the same time, this is less apparent for moderate activity, where the correlation seems very small. Even at smaller age groups, we observed that moderate activity is not highly correlated to total cholesterol. To observe the effect per gender, we also plotted the results per gender. We could argue that age is more important in females than in males for being in the high-risk group for total cholesterol, as we see a big difference in the lift curves between those genders, as displayed in Figures D.14b and D.14c. Further, for both gender, there seems to be no relationship between moderate values and total cholesterol present.



(a) All participants (b) Male (c) Female

Figure D.14: Total cholesterol lift curves

We also looked at the people with the highest moderate lengths of activity and the most moderate bouts of 10 minutes; however, there seems to be no relationship present. However, based on literature, this is also expected [85]. Hence, we conclude that age seems correlated to total cholesterol in a minor manner, while no apparent correlation for total cholesterol and moderate activity seems to be present.

D.4 HDL-Cholesterol

This relationship between HDL-cholesterol and age seems not immediately clear from Figure D.15; however, we could argue there is a small linear connection going upwards between age and HDL-cholesterol. It is important to remember we classify high-risk HDL, as having low HDL values, hence, HDL is seen as a good cholesterol. We also observe that from the boxplot in Figure D.16, it is hard to observe if there is a clear relationship present between the moderate count and HDL-cholesterol; the difference in median of the boxplot is minimal. This difference can also be explained by Figure D.17, where we can see the global relationship between the moderate count

and HDL-cholesterol. We would argue that there seems to be no strong relationship between age and HDL-cholesterol but also not between HDL and PA.

Figure D.15: Relationship of HDL and age

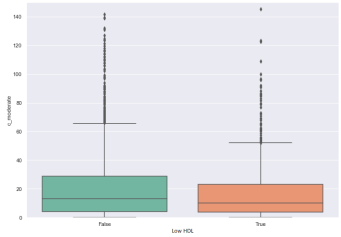


Figure D.16: Box plot HDL against moderate

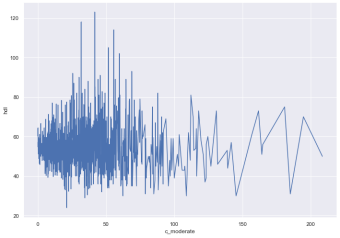
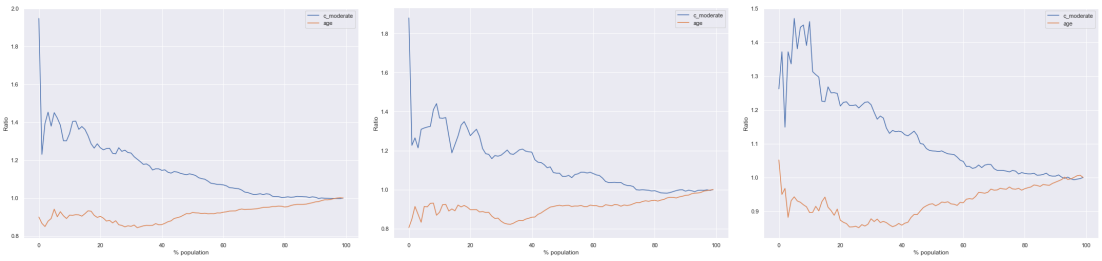


Figure D.17: hdl against moderate relationship

We observe that there is a relationship at all age groups in the lift curve, displayed in Figure D.18a, suggesting that there is some correlation between PA and HDL-cholesterol. However, we could argue that moderate activity has no effect giving its inverse association with age. From the age correlation, we observe that the older someone is, the less likely they are at high risk for HDL-cholesterol values. On the other hand, the more moderate counts someone has, the better the HDL values are; this was difficult to conclude from Figure D.17. The lift curves are also present for males and females, in Figures D.18b and D.18c respectively. For females, we observe that at low population samples, the moderate value is more correlated, based on lift ratio, than the age. We observe that after 10 % of the population, that the effect for males and females is similar, and that age has a similar inverse association. We observed similar results for a smaller age group, where the moderate activity would hover around a lift value of one. Hence, we conclude that there seems to be a correlation between age and HDL; however, a similar association is present for moderate activity. However, given the inverse association between age and moderate activity, we argue that the correlation between moderate and HDL-cholesterol is related to the correlation between age and moderate values, also given we observed no correlation at smaller age groups.



(a) All participants (b) Male (c) Female

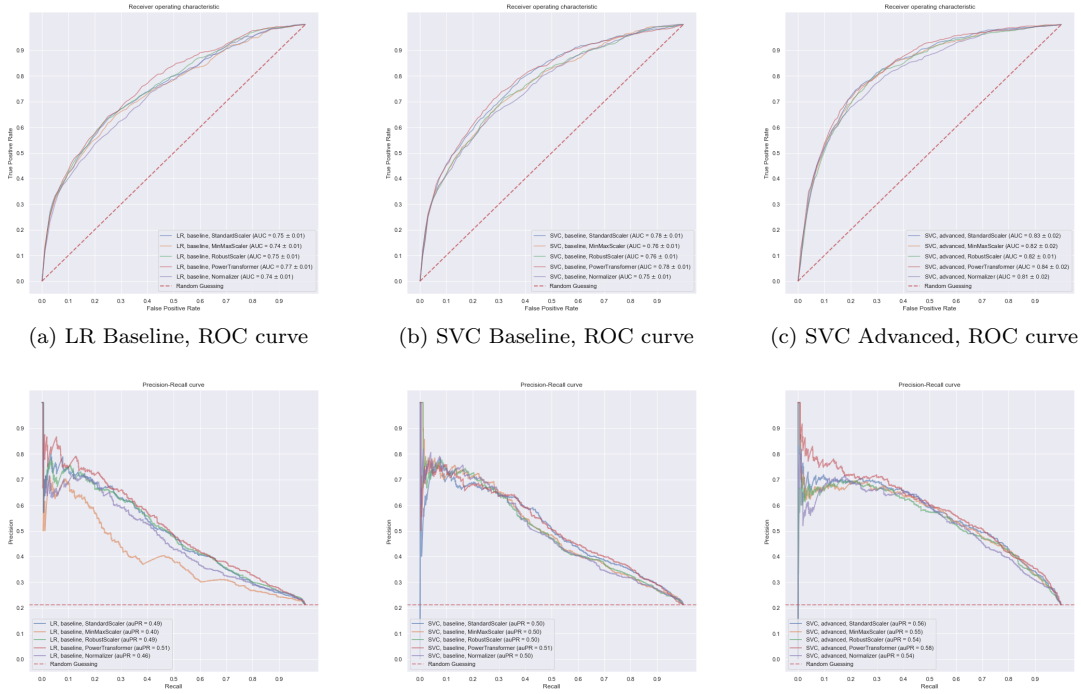
Figure D.18: HDL-cholesterol lift curves

Appendix E

Results

This Appendix Chapter displays the ROC and PR curves that were not displayed in the main text. Additionally, additional information can also be found here not to clutter the main work, such as the non-healthy analysis, which reasons why adding non-healthy participants to the training data is reasonable.

E.1 Scaler Results



(a) LR Baseline, ROC curve (b) SVC Baseline, ROC curve (c) SVC Advanced, ROC curve

(d) LR Baseline, PR curve (e) SVC Baseline, PR curve (f) SVC Advanced, PR curve

Figure E.1: Scaling ROC and PR curves.

E.2 Sampling Results

Baseline features	Logistic Regression				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	0.767	0.766	0.769	**0.768**	0.767
AVG BS 0	**0.104**	0.169	0.163	0.162	0.176
AVG BS 1	0.324	0.231	0.232	0.233	**0.219**
AVG auPR	**0.516**	0.504	**0.516**	**0.516**	0.513

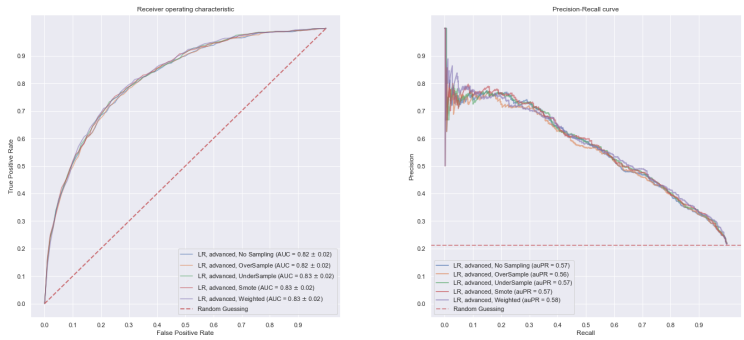
Baseline features	Random Forest Classification				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	**0.784**	**0.784**	0.783	0.782	0.783
AVG BS 0	**0.082**	0.152	0.147	0.149	0.142
AVG BS 1	0.353	**0.233**	0.24	0.24	0.246
AVG auPR	**0.528**	**0.528**	0.522	0.511	0.524

Baseline features	Support Vector Classifier				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	0.764	0.775	0.779	0.776	**0.782**
AVG BS 0	**0.077**	0.157	0.152	0.154	0.085
AVG BS 1	0.381	**0.236**	**0.236**	0.237	0.35
AVG auPR	0.515	0.506	**0.516**	0.515	0.513

Baseline features	Decision Tree				
	Normal	Oversample	Undersample	Smote	Re-weighted
AVG AUC	0.765	**0.768**	0.757	0.759	0.761
AVG BS 0	**0.086**	0.161	0.151	0.158	0.157
AVG BS 1	0.362	**0.239**	0.258	0.253	0.25
AVG auPR	**0.505**	0.487	0.491	0.488	0.474

Table E.1: Sampling results on baseline features

Logistic Regression For the Logistic Regression model, we observe that the ROC curve, as displayed in Figure E.2, is almost not affected by sampling. The PR-curve, displayed in Figure E.2, is slightly affected, however, very minimally. Given that it does not heavily influence the results but also increases the model's confidence for the high-risk class, we have decided to use the re-weighted classifier. We believe that the changes are so small that artificially changing the data is not worth it, and simply using the re-weighted classifier to boost confidence is best.

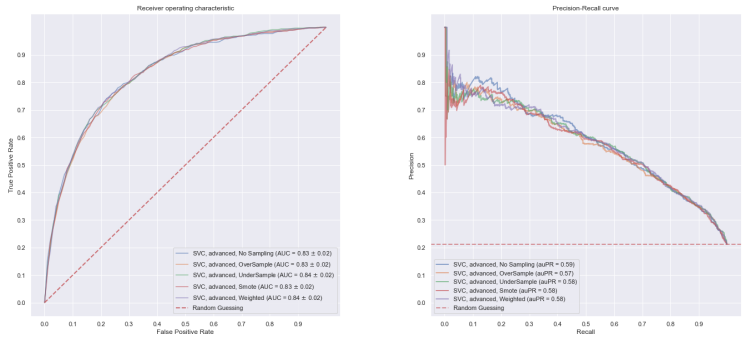


(a) ROC curve (b) PR curve

Figure E.2: Sampling result, roc and pr curve, logistic regression

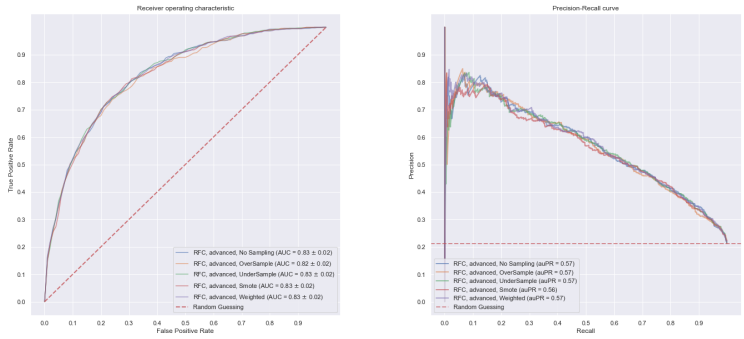
Support Vector Classifier For the SVM, we observe a more interesting case, as shown in Figure E.3. We argue that not using sampling is best, given the increase in precision values at lower recall values. Although the brier score does not decrease with the re-weighted class, this is not of great importance given that we obtain more performance out of the PR-curve. The brier-score is displayed in Table 4.2.

Random Forest Classifier For the RFC, we observe a similar case as for the Logistic Regression model. We argue that the difference in the curve is minimal, as displayed in Figure E.4 and there is no benefit in artificially changing the data with other sampling techniques. Therefore, as similar as before, we decide to use the re-weighted classifier to boost the brier score of the high-risk score. The brier-scores are visible in Table 4.2.



(a) ROC curve (b) PR curve

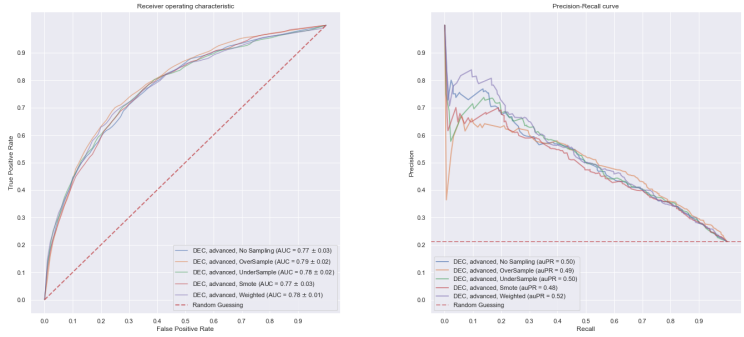
Figure E.3: Sampling result, roc and pr curve, svc



(a) ROC curve (b) PR curve

Figure E.4: Sampling result, roc and pr curve, rfc

Decision Tree Classifier Lastly, DEC, we have decided to again use a re-weighted classifier, given that, from Table 4.2, we could argue that the marginal gains are the best when using a re-weighted classifier. This decision is again mainly based on the gains on the PR curve, which can be seen in Figure E.5. For the baseline, we observed from the PR-curve that using no sampling is best. We observed that at lower recall values, the precision was again much higher than the other classifiers.



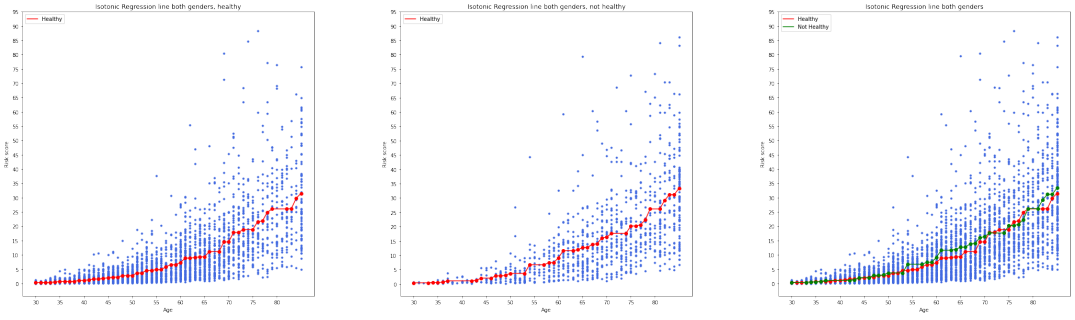
(a) ROC curve (b) PR curve

Figure E.5: Sampling result, roc and pr curve, dec

E.3 Non-Healthy Participant Analysis

This section will give further argumentation on why including non-healthy people in our training data is valid. Although these people are non-healthy, it has been shown that stroke has no statistical association with increased blood pressure, compared to people with no stroke, which is one of the major causes of having a stroke [17]. Also, Lattanzi et al. [70] argued that after a stroke, blood pressure values already return normal, in two-thirds of the cases, after the first week. This means that although they are at higher risk due to their blood vessels being damaged, their blood values are not altered, and thus, the risk score is still correctly calculated if we assume they did not have a stroke. Additionally, given that cancer is also one of the main diseases in non-healthy participants, we observe that there also seems to be no studies relating the risk score's bio-markers to long-term effects of cancer treatment.

We have plotted the average risk line for the healthy and non-healthy participants, as displayed in Figures E.6a, and E.6b respectively. The latter Figure is used to plot if there are significant differences in terms of the points scattered between risk and age. However, we observe no strange deviations. Given that the risk-score does not significantly deviate from our average line, as depicted in Figure E.6c, we argue that the risk of the non-healthy group, in terms of bio-markers, is around the same as for the healthy participants. There are small variations; however, we reason they are not significant. We reason that the small variations can also be explained by the fact that each age group does not have enough participants to display the average risk correctly. Adding the non-healthy participants to the training data should also add more data per age group and make this more accurate.



(a) Healthy participants scatter plot

(b) Non-healthy participants scatter plot

(c) Difference healthy and non-healthy participants

Figure E.6: Difference in average risk between healthy and non-healthy participants.

The non-Healthy column in Table E.2 represents the data we are going to add to the training data. Although the average risk is higher for non-healthy participants, it is essential to note that this is due to an average higher age, and age is correlated with the risk score. We observed previously, per age group, that there was no significant difference in risk.

Both genders	Healthy	Non-Healthy
Amount	3554	998
low risk	2747	407
high risk	807	591
mean risk	6.95	16.67
mean age	53.06	68.22

Table E.2: Healthy Non-Healthy comparison

In addition to the non-healthy people, we are also curious if we can include males who had

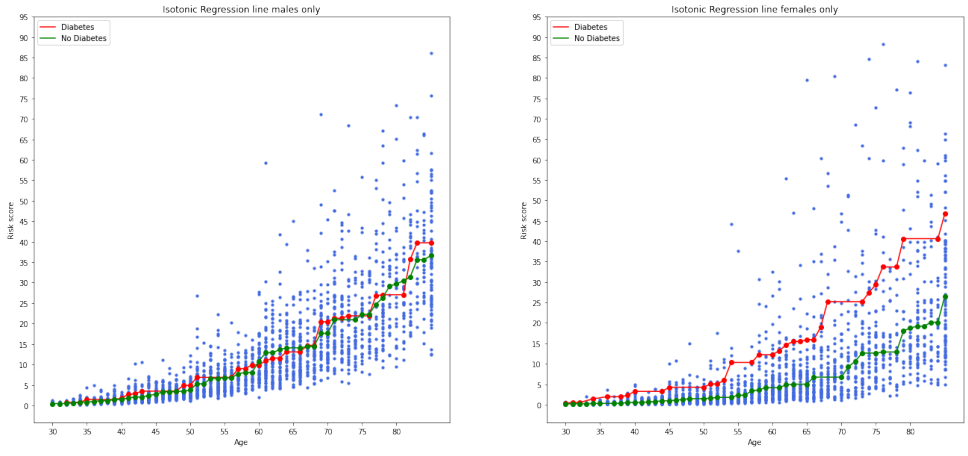
diabetes in our training data, given that the risk score was developed on non-diabetic men. We observe from our data that out of 2227 females, 295 have diabetes (13.2 %). Moreover, out of 2325 males, 335 have diabetes (14.4 %). This is also depicted in Table E.3 in more detail.

	Male		Female	
	with diabetes	without diabetes	with diabetes	without diabetes
Low risk	145	1277	121	1611
High risk	190	713	174	321
Mean risk	14,66	10.3	19,33	5.3
Mean age	62,69	55.06	63.97	55.5

Table E.3: Diabetes per gender.

Based on literature, diabetes in a woman increases the risk of cardiovascular events, with 44 % according to Gnuatiuc et al. [42], which is in line with the results of other papers [96], compared to men with diabetes. Although men generally have a more likely chance to be at high risk due to gender-specific attributes, the effect of diabetes is more noticeable in females. According to Gnuatiuc et al. [42], the risk of stroke roughly triples for women with diabetes and roughly doubles for men. From Table E.3 we do observe a more dominant increase in mean risk for females with diabetes compared to males with diabetes. This difference is explained because diabetes is considered for the RRS, for females, but not for males. Given that age is a dominant factor in the risk score, we will also investigate the increase in risk per age group for both males and females.

As seen in Figure E.7a, we can observe that there is no difference in the risk score for males, if you have diabetes or not. This difference can be explained because the RRS for males does not consider the effect of diabetes, and therefore is unable to increase the risk for those particular men. Although there are small deviations, these could come from the fact that we only have 335 males available, and spreading this out over all ages, leaves us with a very small amount of data.



(a) Males with and without diabetes

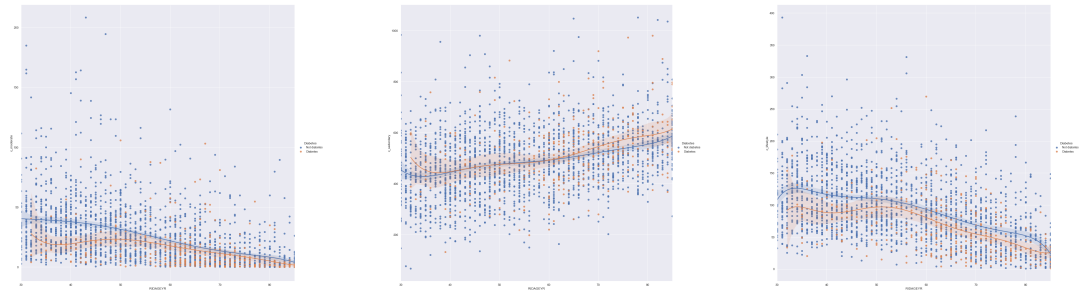
(b) Females with and without diabetes

Figure E.7: Difference in diabetes effect per age

For females, however, in Figure E.7b we observe a significant difference, which we could argue is comparable to the increase in risk argued by Gnuetiuc et al. [42]. We observe that the risk score is between two and three times higher for females with diabetes on average. Given these results, we have decided to exclude males with diabetes from the final analysis. Our analysis suggests that diabetes males are not at higher risk than males without diabetes. This conclusion, however, given the literature, is not the case. However, suppose these are included in the training data.

In that case, they are assumed not to have diabetes, where they would have similar risk scores given that the RRS does not differentiate between people with and without diabetes. Hence, we conclude that based on these facts, including diabetes males in the training data is valid due to the RRS not accounting for diabetes for males, and the difference in risk scores is negligible.

Lastly, we do not observe any major differences in physical activity patterns between males with or without diabetes, which we perhaps first expected. This can be seen in Figures E.8a, E.8b, and E.8c. We argue that given the amount of available data, the lines follow closely enough to be considered similar. Even if there were a slight deviation, we would argue that the impact would be minimal, and perhaps the additional data can be seen as a trade-off. We see a similar occurrence between the healthy and non healthy participants, visible in Figure E.9. Hence, we also conclude no significant difference between physical activity in healthy and non-healthy participants.



(a) Age moderate plot (b) Age lifestyle plot (c) Age sedentary plot

Figure E.8: Difference in PA between diabetic, and non diabetic men

We will test the addition of non-healthy participants in our training data on both the baseline features and the advanced feature set. We will observe if more data could be more beneficial in more advanced features, which perhaps needs more time to train. We want to maximize the AUC and PR scores and minimize the brier score (BS) as our evaluation metrics. We are interested in if more data will reduce the brier score, as more data should give the model more data and thus give the model more confidence in its prediction, especially for higher risk.

We observe from the tables and ROC and PR curves that the impact of adding the data is different per model used but beneficial. For the males, as depicted in Table E.4, we observe that the AUC values generally increase, similar to the auPR values. Only the decision tree using the advanced features decreases in performance. Moreover, from Figure E.10, we observe that, especially at earlier recall values, we observe a benefit in including the non-healthy participants in the training data. This benefit also occurs for both the baseline but also advanced feature set. Furthermore, we observe that the brier-score generally is better for our low-risk groups while being higher for the high-risk groups after including the non-healthy participants. This effect was not expected; given more data on high-risk people, we would expect more confidence in these people's predictions. However, given that we are using re-weighted classifiers, we observe that this is not the case. However, we did see an evident increase in brier score performance for the high-risk group when not using re-weighted classifiers. Hence, we believe this is an artifact of using the re-weighted classifiers. However, as previously discussed, we find the AUC and auPR more important. This allows us to classify more participants correctly; having a confident model would be a nice feature. Hence, for the males, we conclude that adding the non-healthy participants to our dataset is beneficial, given the significant increase in PR curves at earlier recall values.

	Logistic Regression		**Support Vector Classifier**		**Decision Tree**		**Random Forest**	
Baseline	Original	New	Original	New	Original	New	Original	New
AVG AUC	**0.814**	**0.814**	**0.832**	0.828	0.808	**0.824**	0.867	**0.873**
AVG BS 0	0.212	**0.155**	**0.066**	0.095	0.174	**0.124**	0.112	**0.1**
AVG BS 1	**0.181**	0.212	0.341	**0.268**	**0.194**	0.228	**0.194**	0.202
AVG auPR	0.596	**0.644**	0.643	**0.663**	0.623	**0.657**	0.722	**0.741**

	Logistic Regression		**Support Vector Classifier**		**Decision Tree**		**Random Forest**	
Advanced	Original	New	Original	New	Original	New	Original	New
AVG AUC	0.884	**0.885**	0.889	**0.891**	**0.811**	0.8	0.864	**0.871**
AVG BS 0	0.138	**0.101**	**0.057**	0.077	0.16	**0.135**	0.124	**0.103**
AVG BS 1	**0.137**	0.178	0.263	**0.207**	0.21	**0.245**	**0.182**	0.201
AVG auPR	0.744	**0.755**	0.753	**0.772**	**0.623**	0.606	0.717	**0.738**

Table E.4: Comparison of inclusion and exclusion of non-healthy participants to training data, males only, where original = without non-healthy participants in the training data and new = inclusion of non-healthy participants in the training data

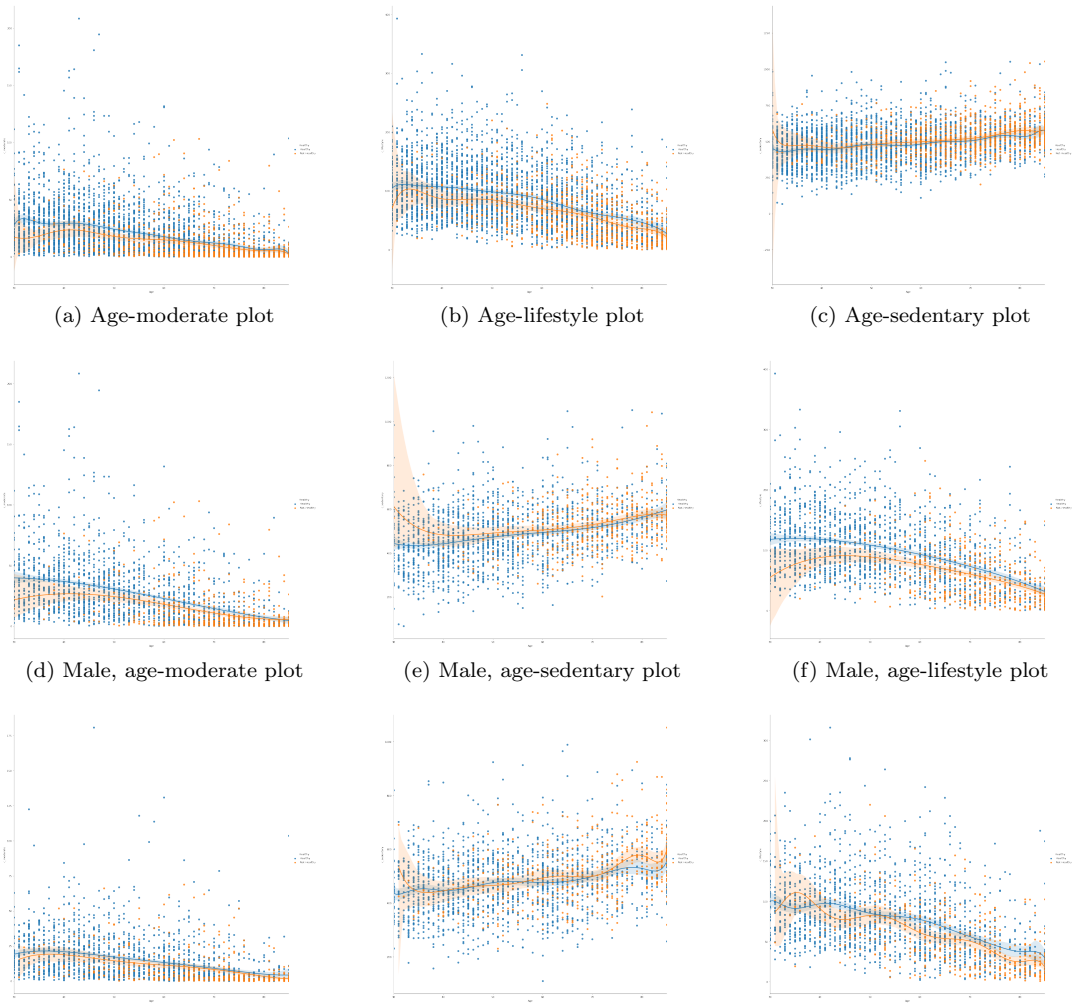
For the females, we observe that the results are slightly different, compared to the males, as displayed in Table E.5. First, it is crucial to observe that the auPR for the females is lower than for the males. We observe that the advanced features have less of an impact on the performance of the females compared to the males, and females have worse PR curves in general than males. Hence, we argue, given the increase in auPR values, that also for the females, we observe an increase in performance. Lastly, the ROC and PR curves can also be viewed for the females, as depicted in Figure E.11. We observe a performance increase when including non-healthy participants in the training data. It is, however, important to note for the PR curves for the females that we observe a drop at the start of the curve. This drop happens because the PR curves use a threshold value that decreases from left to right. Hence, the more we move to the right, the more lenient we become of what we deem as a positive sample, as our threshold of being positive decreases, and thus we classify more examples as positive. In this case, the model assigns a 0.9375 prediction probability to a female of the age 55, who to model believes to be high risk, but is low risk, with a risk of 3.9 %. However, the model was so confident it would predict this correct; this is the first value greater than the start of the threshold value in the PR curve, causing the drop, given that the prediction was wrong. We observe, when inspecting the female participants, this happens more often. However, after a close inspection of several cases, we conclude that the physical activity pattern could be seen as correct, as no abrupt or strange patterns can be observed. It is thus also not possible to remove those from the data as outliers, as their data seems normal. Appendix Section E.4 shows an example of two females.

	Logistic Regression		**Support Vector Classifier**		**Decision Tree**		**Random Forest**	
Baseline	Original	New	Original	New	Original	New	Original	New
AVG AUC	0.81	**0.816**	0.815	**0.82**	0.788	**0.807**	0.828	**0.832**
AVG BS 0	0.205	**0.171**	**0.034**	0.049	0.166	**0.134**	0.134	**0.099**
AVG BS 1	**0.171**	0.189	0.488	**0.406**	**0.232**	0.241	**0.214**	0.258
AVG auPR	0.446	**0.484**	0.468	**0.484**	0.462	**0.477**	0.498	**0.525**

	Logistic Regression		**Support Vector Classifier**		**Decision Tree**		**Random Forest**	
Advanced	Original	New	Original	New	Original	0.241	Original	New
AVG AUC	**0.83**	0.829	0.819	**0.832**	**0.781**	0.788	0.824	**0.83**
AVG BS 0	0.21	**0.177**	**0.034**	0.05	0.173	**0.13**	0.152	**0.122**
AVG BS 1	**0.166**	0.176	0.481	**0.395**	0.235	**0.257**	**0.196**	0.225
AVG auPR	0.46	**0.503**	0.477	**0.5**	**0.43**	0.476	0.468	**0.506**

Table E.5: Comparison of inclusion and exclusion of non-healthy participants to training data, females only, where original = without non-healthy participants in the training data and new = inclusion of non-healthy participants in the training data

Hence, we conclude that adding non-healthy participants to our training data is beneficial. It allows for more data and more information regarding people at high risk. We reason it is valid due to no significant differences in risk values per age or difference in PA between age groups.



(a) Age-moderate plot (b) Age-lifestyle plot (c) Age-sedentary plot

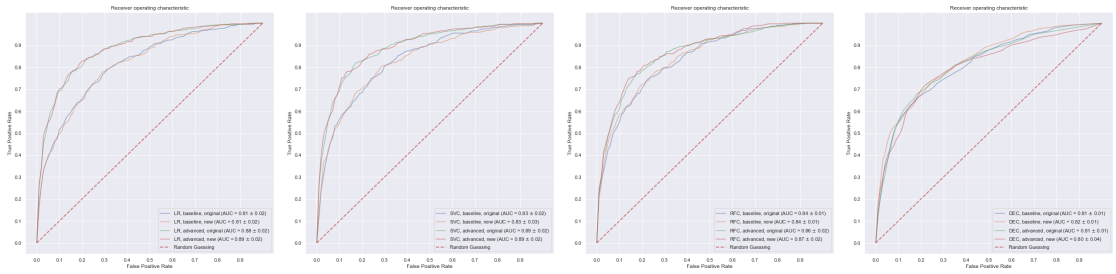
(d) Male, age-moderate plot (e) Male, age-sedentary plot (f) Male, age-lifestyle plot

(g) Female, age-moderate plot (h) Female, age-sedentary plot (i) Female, age-lifestyle plot

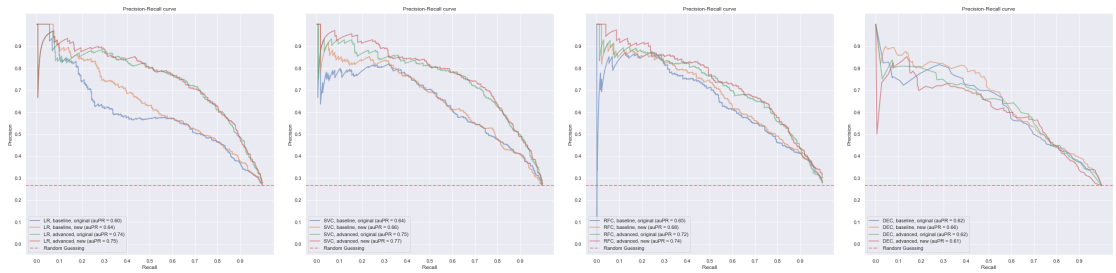
Figure E.9: Difference in PA between healthy and non healthy people per age.

	Male			Female		
	No diabetes	Diabetes	Pre-Diabetes	No diabetes	Diabetes	Pre-Diabetes
Amount	2604	327	46	2702	315	48
Average age	50.47 (0.36)	62.49 (0.73)	62.52 (2.23)	49.53 (0.36)	62.77 (0.77)	60.17 (2.46)
Mean activity count	349.56 (3.57)	244.32 (7.99)	250.35 (19.25)	281.66 (2.40)	202.01 (6.04)	211.10 (15.86)
Mean sedentary counts	479.99 (2.53)	520.21 (7.24)	515.80 (20.97)	473.85 (2.10)	515.34 (6.49)	527.31 (19.59)
Mean light counts	253.12 (1.36)	234.95 (3.87)	247.80 (9.86)	268.62 (1.28)	253.32 (4.29)	247.46 (10.29)
Mean lifestyle counts	95.71 (1.08)	66.16 (2.61)	68.54 (6.69)	76.44 (0.84)	49.67 (2.19)	55.54 (6.27)
Mean moderate counts	27.46 (0.52)	13.83 (1.01)	14.46 (2.84)	14.93 (0.30)	7.14 (0.60)	8.12 (1.57)
Mean vigorous counts	0.92 (0.07)	0.12 (0.05)	0.00 (0.00)	0.45 (0.05)	0.15 (0.10)	0.04 (0.03)

Table E.6: Diabetes statistics

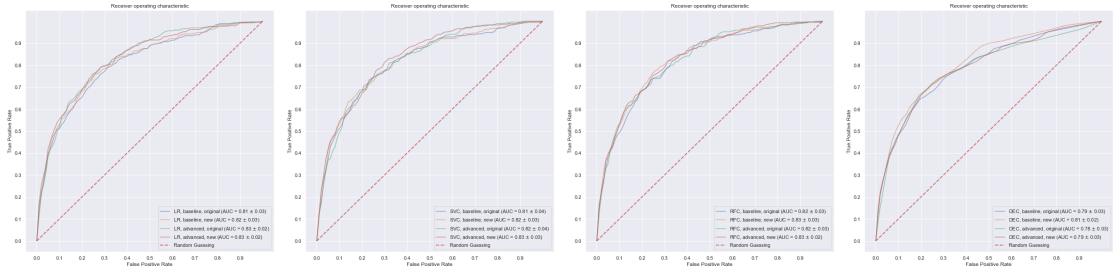


(a) LR ROC curve (b) SVC ROC curve (c) DEC ROC curve (d) DEC ROC curve

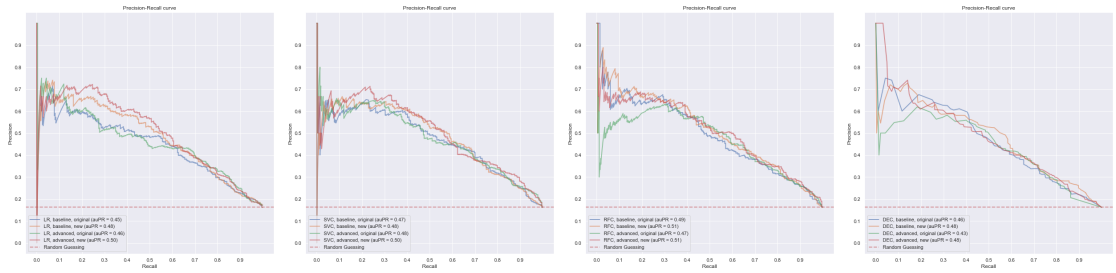


(e) LR PR curve (f) SVC PR curve (g) DEC PR curve (h) DEC PR curve

Figure E.10: Male ROC and PR curves with and without non-healthy participants in training data.



(a) LR ROC curve (b) SVC ROC curve (c) DEC ROC curve (d) DEC ROC curve

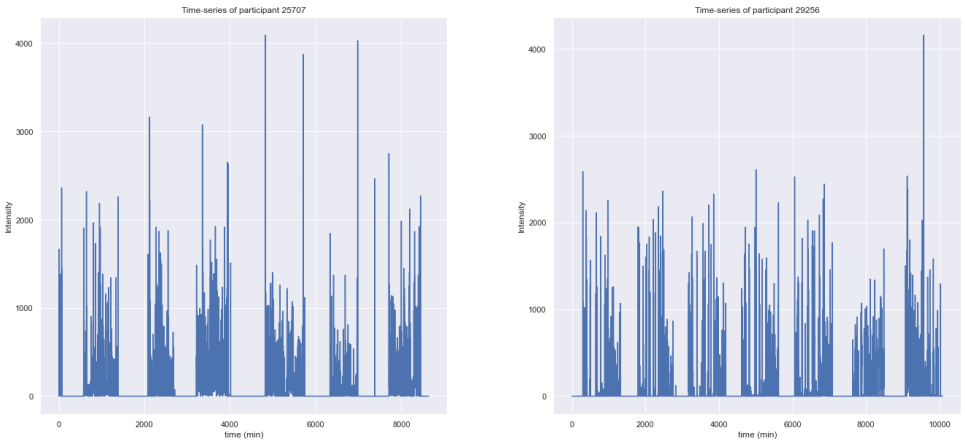


(e) LR PR curve (f) SVC PR curve (g) DEC PR curve (h) DEC PR curve

Figure E.11: Female ROC and PR curves with and without non-healthy participants in training data.

E.4 Outliers

In this section, we elaborate further on 'outliers' based on the model output. For the female data, we observed that the PR curves, as depicted in Appendix E.11, that there is a steep drop at the start of the PR curves. An outlier could explain this drop, for instance, a participant who has an invalid physical activity pattern or a very physically inactive participant who is still at low risk. We conclude that the 'outlier' PA time series can be regarded as expected, and no abnormal patterns are observed. We conclude, as can be observed from Figures E.12a and E.12b, that the time-series appear normal for two of the outliers inspected.



(a) Participant 25707 (b) Participant 29256

Figure E.12: Original Time Series of outliers

Moreover, even after a more close inspection for participant 25707, as depicted in Figure E.13, we observe no strange patterns or abnormal activity patterns. Hence, this is a data point we cannot remove, and we have to deal with, given that this is a case where the physical activity paints a dire picture of the actual risk of a participant. Additionally, given that the biomarkers are considered normal and lead to a low-risk score, we deem these points not to be outliers, simply participants who perhaps do not have to be physically active be at low risk for CVD.

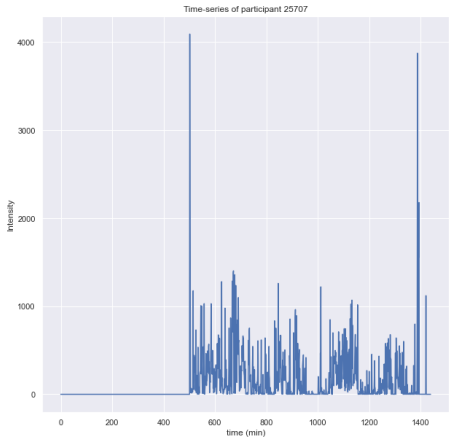
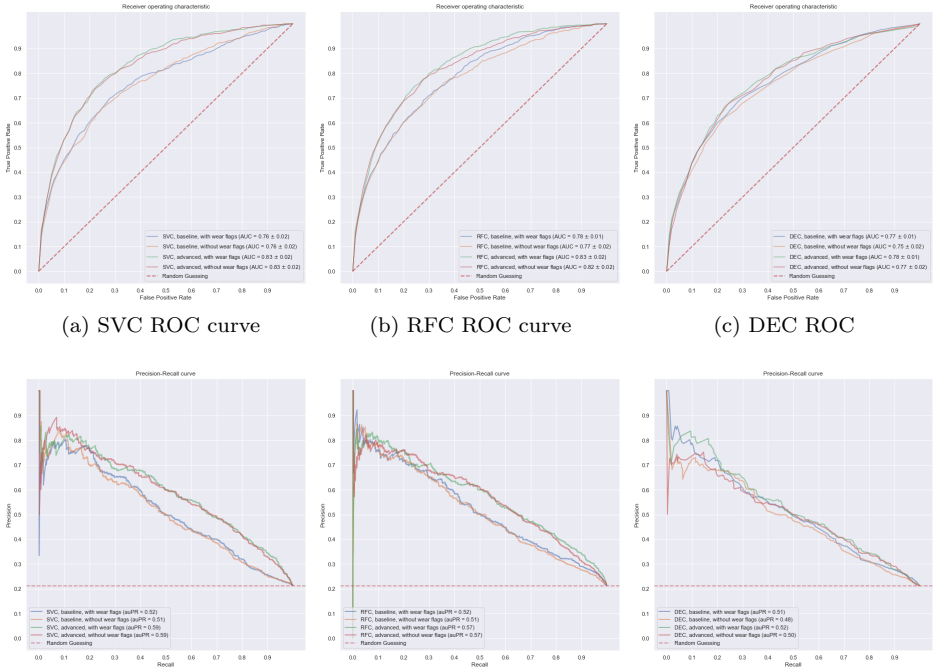


Figure E.13: Day 4 of participant 25707

E.5 Wear Time Flags

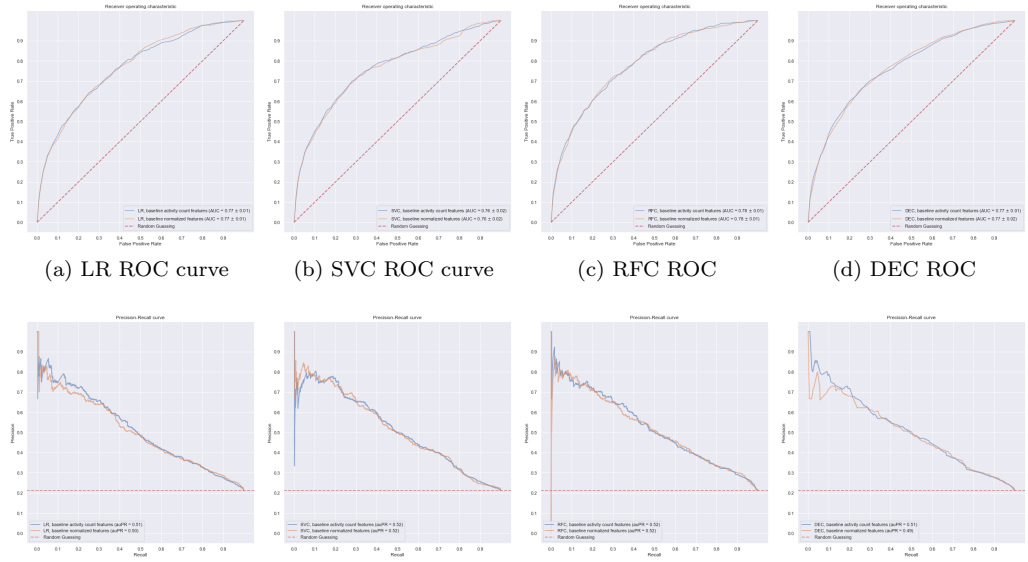


(a) SVC ROC curve (b) RFC ROC curve (c) DEC ROC

(d) SVC PR curve (e) RFC PR curve (f) DEC PR curve

Figure E.14: ROC and PR curves comparison with and without wear flags

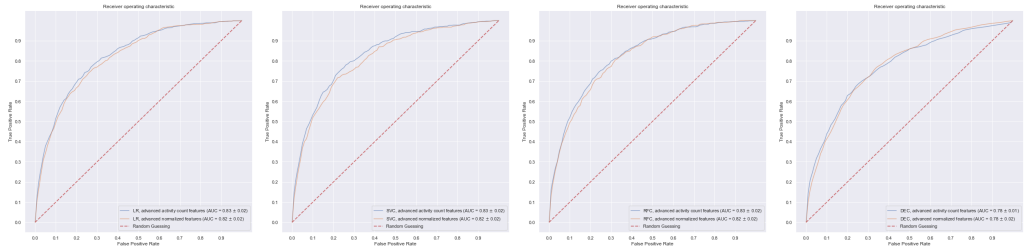
E.6 Normalized Features Results



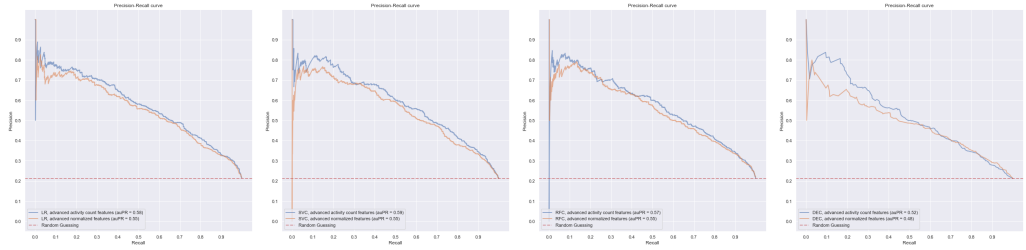
(a) LR ROC curve (b) SVC ROC curve (c) RFC ROC (d) DEC ROC

(e) LR PR curve (f) SVC PR curve (g) RFC PR curve (h) DEC PR curve

Figure E.15: ROC and PR-curves comparison baseline model, normalized features



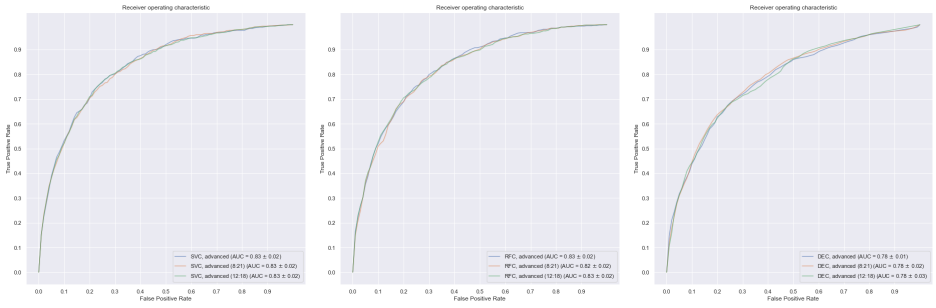
(a) LR ROC curve (b) SVC ROC curve (c) RFC ROC curve (d) DEC ROC curve



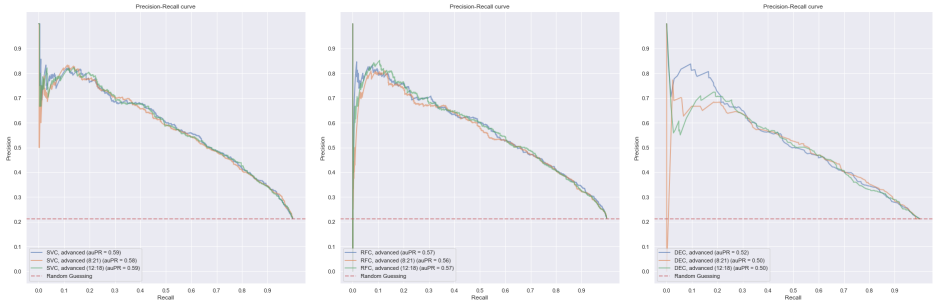
(e) LR PR curve (f) SVC PR curve (g) RFC PR curve (h) DEC PR curve

Figure E.16: ROC and PR-curves comparison advanced model, normalized features

E.7 Imputation Results



(a) SVC ROC curve (b) RFC ROC curve (c) DEC ROC curve



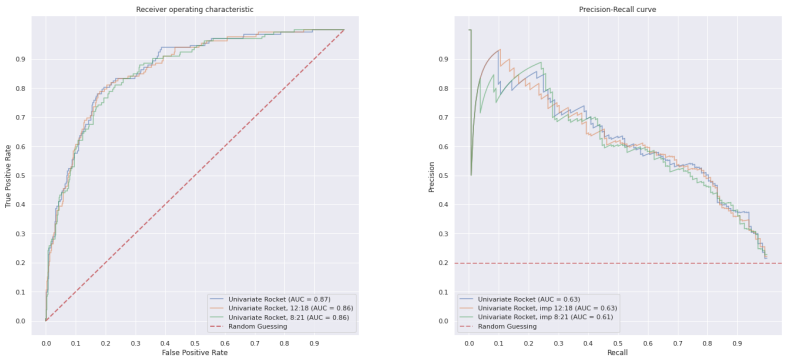
(d) SVC PR curve (e) RFC PR curve (f) DEC PR curve

Figure E.17: ROC and PR curves comparison main imputation settings.

E.8 MiniRocket

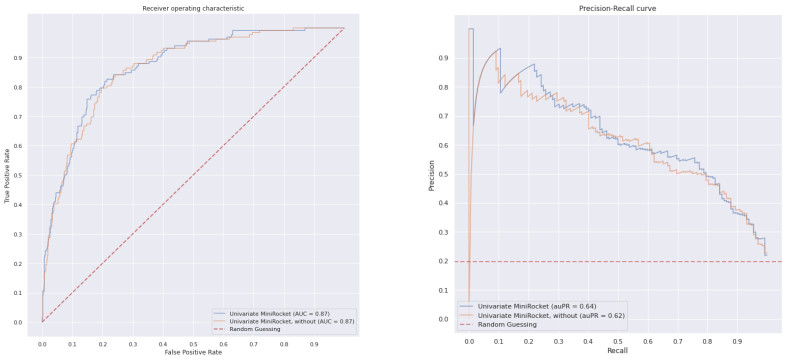
(a) BA-diff MiniRocket (b) BA-diff MiniRocket log

Figure E.18: MiniRocket PA only model comparison, BA-diff plots log



(a) ROC curve (b) PR curve

Figure E.19: Univariate MiniRocket imputation comparison



(a) ROC curve (b) PR curve

Figure E.20: Univariate MiniRocket inclusion of non-healthy participants comparison

E.9 Feature Importance RFC

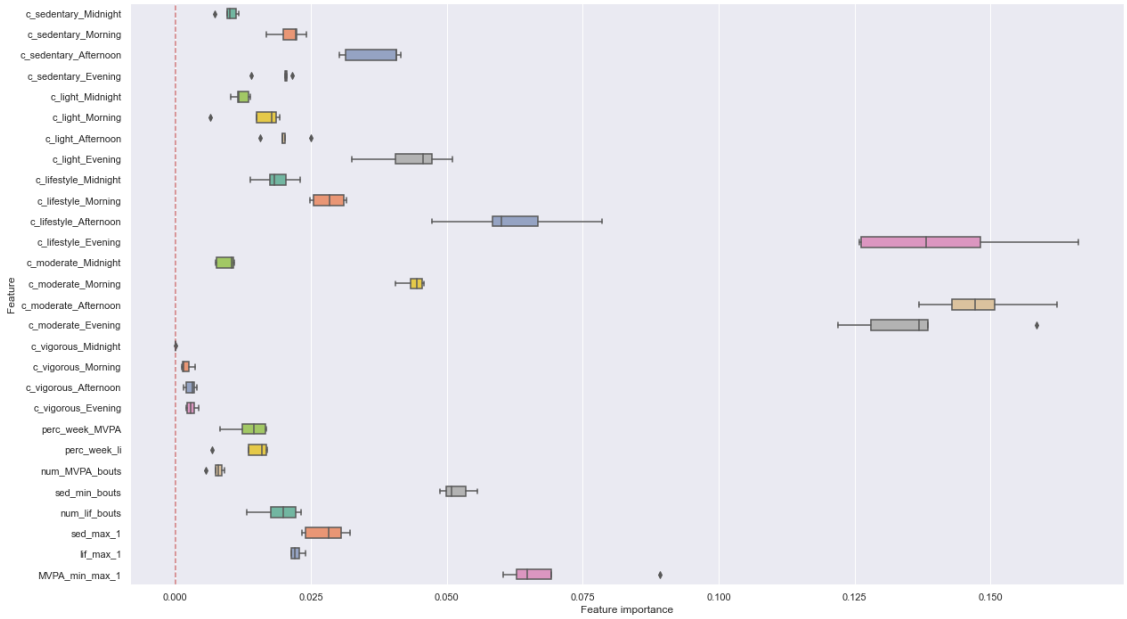


Figure E.21: Feature importance RFC

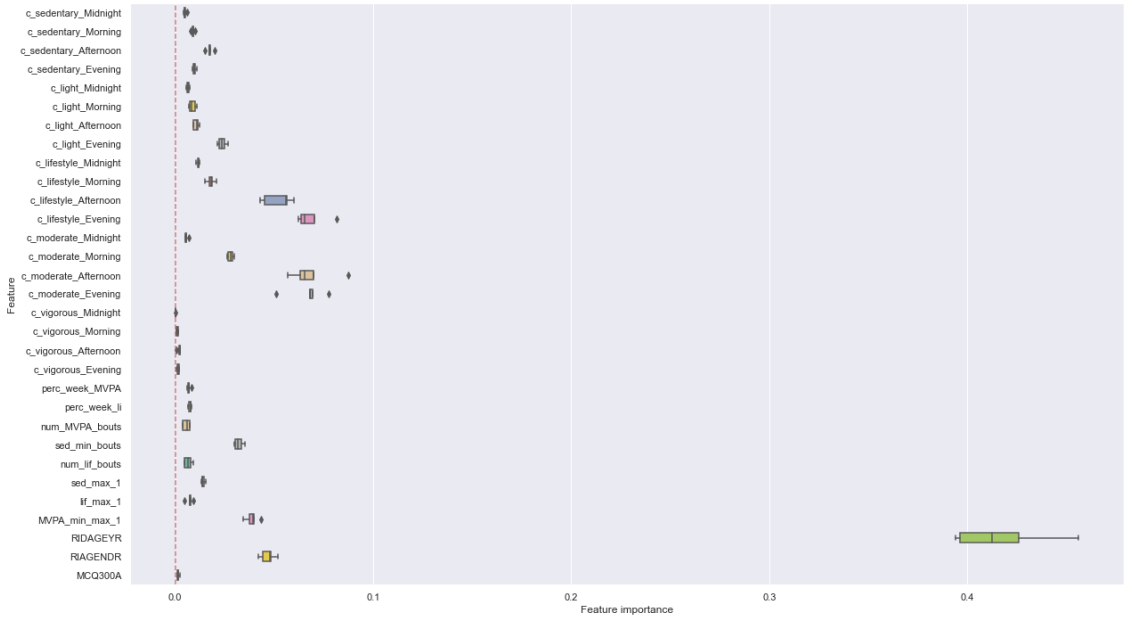


Figure E.22: Feature importance RFC including NMF

Appendix F

Parameters and Architectures

F.1 Machine Learning Parameters

It is important to note that Figure F.1 list the parameters used in all the experiments, however, a more extensive parameters list has been tested. To improve the running time of these models, especially for the SVC model, we have limited the available parameters based on which parameters were most often chosen based on the validation data. For instance, we observed that the SVC model always used the rbf kernel, hence why the option rbf is the only available kernel.

```
param_grid_svc_b = {'C': [100, 1000],  'gamma': [0.01, 0.001, 0.0001], 'kernel': ['rbf'],
                    'random_state' : [200], 'probability' : [1]}

param_grid_lr_b = {'C': [0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1, 5, 10, 25, 100],
                   'penalty': ['l1', 'l2'], 'solver' : ['liblinear'],
                  'random_state' : [200], 'class_weight': ['balanced'], 'max_iter' : [2500]}

param_grid_rfc_b = {'criterion': ['gini'], 'max_depth' : range(1,10), 'n_estimators' : [50, 100, 150],
                    'min_samples_leaf' : [10,25,50], 'random_state' : [200], 'class_weight': ['balanced']}

param_grid_dec_b = {'criterion': ['gini'], 'max_depth' : range(1,10),
                    'min_samples_leaf' : [10,25,50], 'random_state' : [200]}

param_grid_dec_a = {'criterion': ['gini'], 'max_depth' : range(1,10),
                    'min_samples_leaf' : [10,25,50], 'random_state' : [200], 'class_weight': ['balanced']}
param_grid_ada = [{'n_estimators': [100, 150, 200, 250, 300], 'learning_rate' : [0.01, 0.05, 0.1, 0.5],
                    'random_state': [200], 'loss' : ['linear', 'square', 'exponential']}]

param_grid_rfr = [{'n_estimators': [100, 150, 200, 250, 300], 'max_depth': [1,2,3,4,5,6,7,8],
                    'random_state': [200], 'max_features' : ['auto', 'sqrt', 'log2']}]

start = np.linspace(0,1,500)
start2 = [1,2,3,4,5,6,7,8,9,10,25,50,100,250,500,1000]
total_values = np.append(start, start2)

param_grid_ridge = [{'normalize' : [False], 'alpha': total_values, 'random_state': [200]}]

param_grid_rocket = [{'normalize' : [True], 'alpha': total_values, 'random_state': [200]}]
```

Figure F.1: GridSearch parameters for all models.

F.2 Deep Learning

In Table F.1 are the specific details about the architectures of the models used. As mentioned in the literature review, we have taken inspiration from several models and used those as our baseline models for our problem. These include the FCN and CNN of Fawaz et al. [38], the ResNet of Wang et al. [121] and the InceptionTime network of Fawaz et al. [39]. These models have been optimized for our specific use-case, and thus are slightly altered based on their original structure.

BatchNormalization happens before the Relu activation layers, and dropout happens after the Relu activation layer. Lastly, if dropout is added to the InceptionTime model, it is added after the concatenation layer in the InceptionTime module, which is the top layer as displayed in Figure 2.6.

Methods	# Conv/LSTM/GRU Layers	Normalize	Filters/Units	Kernel	Dropout	Activation	Regularization (L2)
CNN	3	Batch	$[64, 32, 16]$	$[5, 3, 3]$	$[0.5, 0.3, 0]$	ReLu/Sigmoid	0.01
FCN	5	Batch	$[64, 128, 32, 16^2]$	$[8, 5, 3^3]$	$[0.5^3, 0.25^2, 0^2]$	ReLu/Sigmoid	0.01
ResNet	9	Batch	$[32^3, 64^6]$	$[8, 5, 3]$	$[(0.5, 0.3, 0.1)^3]$	ReLu/Sigmoid	0.01
InceptionTime	31	Batch	$[32^{31}]$	$[40, 20, 10]$	$[0.2]$	ReLu/Sigmoid	[39]
LSTM	2	-	$[100, 10]$	-	$[0.35, 0.2]$	Sigmoid	-
GRU	2	-	$[100, 10]$	-	$[0.35, 0.2]$	Sigmoid	-
FCN + NMF	7	Batch	$[256, 128, 64, 32, 16^2, 8]$	$[8, 5^2, 3^4]$	$[0.5^3, 0.25^2, 0^2]$	ReLu/Sigmoid	0.01
InceptionTime + NMF	31	Batch	$[32^{31}]$	$[40, 20, 10]$	-	ReLu/Sigmoid	[39]
FCN Regr	5	Batch	$[64, 128, 32, 16^2]$	$[8, 5, 3^3]$	$[0.5^3, 0.25^2, 0^2]$	ReLu	0.01
InceptionTime Regr	31	Batch	$[32^{31}]$	$[40, 20, 10]$	-	ReLu	[39]
FCN Regr + NMF	3	Batch	$[64, 128, 32]$	$[8, 5, 3]$	$[0.5, 0.25, 0]$	ReLu	0.01
InceptionTime Regr + NMF	31	Batch	$[32^{31}]$	$[40, 20, 10]$	-	ReLu	[39]

Table F.1: Deep Learning Architectures