Automatic Prediction of Human Age based on Heart Rate Variability Analysis using Feature-Based Methods

Yusur Al-Mter

Supervisor: Krzysztof Bartoszek
Examiner: Anders Nordgaard
Upphovsrätt

Detta dokument hålls tillgängligt på Internet - eller dess framtida ersättare - under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/.

Copyright

The publishers will keep this document online on the Internet - or its possible replacement - for a period of 25 years starting from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/her own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

© Yusur Al-Mter
Abstract

Heart rate variability (HRV) is the time variation between adjacent heartbeats. This variation is regulated by the autonomic nervous system (ANS) and its two branches, the sympathetic and parasympathetic nervous system. HRV is considered as an essential clinical tool to estimate the imbalance between the two branches, hence as an indicator of age and cardiac-related events.

This thesis focuses on the ECG recordings during nocturnal rest to estimate the influence of HRV in predicting the age decade of healthy individuals. Time and frequency domains, as well as non-linear methods, are explored to extract the HRV features. Three feature-based methods (support vector machine (SVM), random forest, and extreme gradient boosting (XGBoost)) were employed, and the overall test accuracy achieved in capturing the actual class was relatively low (lower than 30%). SVM classifier had the lowest performance, while random forests and XGBoost performed slightly better. Although the difference is negligible, the random forest had the highest test accuracy, approximately 29%, using a subset of ten optimal HRV features. Furthermore, to validate the findings, the original dataset was shuffled and used as a test set and compared the performance to other related research outputs.
Acknowledgments

My appreciation and respect for my supervisor Krzysztof Bartoszek, for the valuable guidance, advice, and support he has provided. It was a real privilege for me to be his student and share his scientific knowledge throughout the entire Master program.

I would also like to thank my examiner Anders Nordgaard, for his priceless efforts and constructive feedback to ensure that this thesis meets the academic standards.

My gratitude to the program coordinator Oleg Sysoev, for the discipline he showed in organizing the Master program in general and this course in particular.

I would like to extend my sincere gratitude to my family for the unconditional support throughout my years of study in Sweden. And to my source of joy, nephews, and nieces.

Finally, I would like to thank the Computer and Information Science department, Linköping University, for the rich experience and knowledge gained in the Statistics and Machine Learning Master program.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Objective and research questions</td>
<td>3</td>
</tr>
<tr>
<td>2 Data</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Data Description</td>
<td>4</td>
</tr>
<tr>
<td>3 Method</td>
<td>8</td>
</tr>
<tr>
<td>3.1 Heart Rate Variability</td>
<td>8</td>
</tr>
<tr>
<td>3.2 Feature Extraction Methods</td>
<td>9</td>
</tr>
<tr>
<td>3.2.1 Time Domain Analysis</td>
<td>9</td>
</tr>
<tr>
<td>3.2.2 Frequency Domain Analysis</td>
<td>11</td>
</tr>
<tr>
<td>3.2.3 Non-Linear Analysis</td>
<td>12</td>
</tr>
<tr>
<td>3.3 Statistical Analysis</td>
<td>15</td>
</tr>
<tr>
<td>3.3.1 Pearson Correlation Coefficient</td>
<td>16</td>
</tr>
<tr>
<td>3.3.2 Kendall Rank Correlation Coefficient</td>
<td>16</td>
</tr>
<tr>
<td>3.4 Supervised Learning</td>
<td>16</td>
</tr>
<tr>
<td>3.4.1 Support Vector Machines</td>
<td>17</td>
</tr>
<tr>
<td>3.4.2 Random Forests</td>
<td>20</td>
</tr>
<tr>
<td>3.4.3 Extreme Gradient Boosting</td>
<td>22</td>
</tr>
<tr>
<td>3.5 Evaluation Methods</td>
<td>25</td>
</tr>
<tr>
<td>4 Results</td>
<td>26</td>
</tr>
<tr>
<td>4.1 HRV Metrics</td>
<td>26</td>
</tr>
<tr>
<td>4.2 Statistical Evaluations</td>
<td>28</td>
</tr>
<tr>
<td>4.3 Feature Importance</td>
<td>30</td>
</tr>
<tr>
<td>4.4 Model Performance and Evaluation</td>
<td>32</td>
</tr>
<tr>
<td>5 Discussion</td>
<td>37</td>
</tr>
<tr>
<td>5.1 Heart rate variability as an age indicator</td>
<td>37</td>
</tr>
<tr>
<td>6 Conclusion</td>
<td>40</td>
</tr>
<tr>
<td>Appendix</td>
<td>42</td>
</tr>
</tbody>
</table>
## List of Figures

2.1 Class Distribution .................................................. 5
2.2 Signal size .......................................................... 5
2.3 Distribution of RR time intervals for different genders ............... 5
2.4 Missing data sizes .................................................... 6
2.5 500 RRI records of Different age groups for both female and male .... 7
3.1 RR interval ............................................................ 9
3.2 Poincare plots derived from 4-hour RR recording belong to 3 females from different age groups .................................................. 14
3.3 Example plots of: (a) Hyperplane, (b) Regularization with small large value for C, and (c) Linearly separable and Linearly non-separable .................................................. 19
4.1 Bar plot of: (a) Pearson Correlation of HRV metrics with age decade, (b) Kendall rank Correlation of HRV metrics with age decade .................. 28
4.2 Heat maps of the correlated HRV features with P-values of the insignificantly correlated ones for two tests: (a) (Pearson r), (b) (Kendall rank) ................. 29
4.3 Correlation Difference between Pearson and Kendall methods .......... 30
4.4 HRV feature importance in: (a) Random Forest model, (b) XGBoost model .................................................. 31
4.5 Confusion matrices of SVM, Random Forests, and XGBoost classifiers with train, test, and shuffled data sets as validation sets. Using the original set of features (all features) as predictors. 0,1,...,6 represent 20th, 30th,..., 80th age decades, respectively. 34
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Description of subjects</td>
<td>4</td>
</tr>
<tr>
<td>2.2</td>
<td>Data Statistics</td>
<td>5</td>
</tr>
<tr>
<td>2.3</td>
<td>A sample of RRI records with gap</td>
<td>6</td>
</tr>
<tr>
<td>3.1</td>
<td>An example of confusion matrix for binary classification task</td>
<td>25</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary Statistics for Time domain, Frequency domain and Non-Linear Analysis</td>
<td>27</td>
</tr>
<tr>
<td>4.2</td>
<td>Subset of Insignificant Features with Pearson’s correlation &lt; 0.9</td>
<td>30</td>
</tr>
<tr>
<td>4.3</td>
<td>Subset of Random Forest ten most important features</td>
<td>32</td>
</tr>
<tr>
<td>4.4</td>
<td>Subset of XGBoost ten most important features</td>
<td>32</td>
</tr>
<tr>
<td>4.5</td>
<td>Feature-Based Classifiers’ Accuracy</td>
<td>33</td>
</tr>
<tr>
<td>4.6</td>
<td>SVM, Random Forest and XGBoost classifiers performance with both train and</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>test data as validation sets. Using different combinations of HRV set of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>features: original set of features (all features), Random Forest and XGBoost</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ten most important features, also insignificantly correlated set of features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>from Pearson’s correlation test</td>
<td>36</td>
</tr>
<tr>
<td>4.7</td>
<td>SVM, Random Forest, and XGBoost classifiers performance with shuffled data as</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a validation set. Using different combinations of HRV set of features as</td>
<td></td>
</tr>
<tr>
<td></td>
<td>predictors: original set of features (all features), Random Forest and XGBoost</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ten most important features, also insignificantly correlated set of features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>from Pearson’s correlation test</td>
<td></td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Background

The phenomenon of heart rate variability (HRV) is the variation in the time intervals between consecutive heartbeats (the so-called RR interval) and is measured in milliseconds. R is the peak of a QRS complex, where the latter is a combination of three electrical waves recorded using a typical electrocardiogram (ECG). QRS corresponds to the depolarization of the right and left ventricles of the human heart and contraction of the large ventricular muscles [1], see Figure 3.1. HRV has witnessed a growing scientific interest as an important tool for predicting the average rate of aging in the population, as well as is considered as an essential clinical tool in estimating the potential risk of medical events, in particular, age-related diseases and cardiovascular diseases [2]. Experimental evidence illustrates the influence of age on the cardiovascular system [3,4,5]. Thus, a recently growing interest is whether the age of a healthy human can be estimated from the cardiovascular states [6,7], in other words, analyzing the HRV and finding the parameters that affect such estimation.

The autonomic nervous system (ANS) and its two branches (the sympathetic and parasympathetic nervous systems) act unconsciously and are affected by several bodily functions, such as heart rate, respiratory rate, emotions, digestion, and many others. HRV is an emanating property among these interdependent regularity systems that operate on different time scales. Analyzing the phenomenon based on the non-linear interactions between those systems provides a significant interpretation of the cardiac state [8]. Certain factors cause the two ANS branches to produce hormones to accelerate and decelerate the heart rate (HR). For example, stress and exercise during the sympathetic tone and meditation and resting during the parasympathetic tone [1]. In other words, the reduced heart rate results from an increase in activity of the parasympathetic nervous system, and perhaps from a decrease in activity of the sympathetic nervous system. This natural interplay between the two systems allows the heart to cope with sudden physical and psychological changes rapidly, achieve homeostasis, and optimal performance.

In healthy biological systems, the dynamics of a cardiac RR interval vary with age. Experimental work has shown an association of increased age with the reduction in both the overall heart rate variability [9,10], and the complexity of the physiological dynamics [11]. This loss of complexity may be due to both structural and functional changes [12]. Many metrics are
involved in analyzing the dynamics of the cardiac RR interval. The most commonly used are from the time-domain, for which they quantify the RR interval variability using statistical measures. Frequency-domain measurements provide methods to quantify the various oscillations over any periodic interval in the RR time series. And finally, the non-linear measures which they quantify the unpredictability and the complexity of a time series [13, 14, 15].

Many efforts were employed recently to determine the optimal HRV features’ influence in estimating the age and cardiac-related events. Park et al. [16] used the support vector machine recursive feature elimination (SVM-RFE) to identify the optimal HRV feature set that can efficiently detect cumulative stress. The SVM classifier with a radial basis function (RBF) kernel increased performance to 93.11% accuracy by using time-domain features such as the mean of RR intervals, the standard deviation of RR intervals, and the coefficient of variance of RR intervals, as well as frequency domain features like the normalized low frequency and the normalized high frequency. Murukesan et al. [17] applied SVM and probabilistic neural networks (PNN), and they found a substantial improvement in the sudden cardiac arrest (SCA) prediction rate by using optimal HRV features derived by a sequential feature selection algorithm and tested by analysis of variance (ANOVA).

HRV features during nocturnal and supine rest have been used frequently by researchers, and the motivation of choosing this phase is the limitation of the artifacts. Also, it provides an ideal state for a better HRV features’ estimation [6]. False beats within RR interval signals occur due to technical and physiological artifacts. Missing or additional beats due to the movement of a subject or the computational methods are examples of technical artifacts. On the other hand, the physiological artifacts may include ectopic beats and arrhythmic events [18].

Recent studies analyzed the HRV series during the sleep stage. Poddar et al. [7] analyzed 34 linear and non-linear HRV features from ECG recordings in the supine state. They classified 60 healthy subjects to three different age groups, namely, young (18-30 years), middle (30-45 years) and old (45-60 years). SVM, the k-nearest neighbor (KNN), and the probabilistic neural network PNN algorithms are performed with and without principal component analysis (PCA). The PNN classifier with PCA outperformed the remaining classifiers with and without PCA by obtaining the highest accuracy around 70%. The study showed that in all models, the young age group was the most correctly classified. The middle and old age groups were overlapping and mainly due to the close values of most of their features (relatively low). While the young age group showed higher values, also, they claimed that the decrease in features’ values is an indication of aging.

Makowiec et al. [6] also experimented with an HRV series of 181 healthy subjects from different age decades in their nocturnal rest. They extracted linear and non-linear features to analyze the time series, and used linear and non-linear SVM with a Gaussian kernel to predict the age decade. They found improvement in prediction accuracy by fitting the classifiers for different sets of age groups and also by using different sets of features, for instance, entropic measures, best ten measures (with highest classification scores), 5 min segments with minimal stdRR and 5 min segments with minimal HR. The reported results showed higher accuracy for the set of ’20s, ’40s, ’60s, and 80’ s age decades using the non-linear SVM, achieving a score of around 93.6 ± 5.3%, and with a 98 ± 1% score for 5 min segments with minimal HR signals. They also claimed that shuffling the original signal was an appropriate way to validate their predictions.

Some previous studies tried to model the effect of the HRV measurements on age, but with an uncertainty in the prediction accuracy. For instance, using a wide range of features compared to relatively small data might mislead the analysis outcomes, and overfitting is usually caused. In this study, supervised machine learning algorithms are proposed to classify the human age by analyzing the HRV series. In particular, these methods are the support vector
machine (separating classes in the feature space by constructing hyperplanes and choosing
the ones that maximize the margin from all classes with minimum risk) [19]. Also, two en-
semble methods were employed, the random forests and extreme gradient boosting (they
generate many classifiers and aggregate their outcomes) [20]. Holdout method is used in this
analysis to provide an unbiased evaluation of the model, through splitting the dataset into
train-test sets.

1.2 Objective and research questions

The aim of the present study is to investigate the possibility to capture the influence of HRV
statistical features on age. In particular, we will use machine learning algorithms to find
features that give accurate predictions of age.

From the dependency between the biological age of a human and several HRV features, the
fundamental questions are:

1. Can we achieve an accurate classification of the human age?
2. How effective are these statistical features derived from the RR intervals to classify the
human age?
3. Which of these HRV features are optimal to classify the human age?
2 Data

2.1 Data Description

In this study, the dataset *i.e.*, ECG recordings was provided by the Institute of Theoretical Physics and Astrophysics and approved by the Ethics Committee of the Medical University of Gdańsk. The dataset consists of RR intervals records measured in milliseconds representing a 4-hour part from 24-hour ECG recordings during nocturnal rest. The measurements are from 181 healthy individuals of different genders.

The dataset includes signals for individuals subject to a specific criterion. Subjects who showed symptoms or diseases that significantly affect the heart rate rhythm in particular, and the cardiovascular system in general, were discarded. The sampling rate of ECG was 128 Hz, and the quality of the recordings were authenticated by specialists [6]. Information, such as gender, the time when the record started, also, the age decade of a subject, which is the response variable, is provided. The full description of the individuals in the dataset is shown in Table 2.1.

<table>
<thead>
<tr>
<th>Age groups</th>
<th>Total Subjects</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>30</td>
<td>17 female/13 male</td>
</tr>
<tr>
<td>30-39</td>
<td>21</td>
<td>11 female/10 male</td>
</tr>
<tr>
<td>40-49</td>
<td>33</td>
<td>13 female/20 male</td>
</tr>
<tr>
<td>50-59</td>
<td>31</td>
<td>13 female/18 male</td>
</tr>
<tr>
<td>60-69</td>
<td>27</td>
<td>12 female/15 male</td>
</tr>
<tr>
<td>70-79</td>
<td>22</td>
<td>10 female/12 male</td>
</tr>
<tr>
<td>80-85</td>
<td>17</td>
<td>11 female/6 male</td>
</tr>
</tbody>
</table>
2.1. Data Description

Figure 2.1 shows the distribution of classes (age decades), where 0 to 6 represents individuals in their 20th up till the 80th age decade, respectively. A histogram of signal size shown in Figure 2.2, where the shortest and the longest signal has 18888 and 28872 RRI records, respectively. The distribution is right-skewed with mean 23083.99 and median 22852. Figure 2.3 shows a distribution comparison of the RRI records for the female group and the male group with the mean and median shown in Table 2.2. The distribution of the female group is seemed to be left-skewed.

![Figure 2.1: Class Distribution](image1)

![Figure 2.2: Signal size](image2)

![Figure 2.3: Distribution of RR time intervals for different genders](image3)

### Table 2.2: Data Statistics

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Signal size</th>
<th>Female RRIs</th>
<th>Male RRIs</th>
<th>Missing data length</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>23083.99</td>
<td>944.2068</td>
<td>894.3494</td>
<td>8</td>
</tr>
<tr>
<td>median</td>
<td>22852</td>
<td>952</td>
<td>888</td>
<td>7</td>
</tr>
</tbody>
</table>
Table 2.3 shows two columns (RRI records and number of contractions) from a sample of a signal with missing RRI records. For instance, after the beat 54053, we have 54057, so the interval of 928 ms is not a time neighbor of 910 ms, and the size of the missing RRI is 3. Figure 2.4 presents a histogram of the missing data length, and it shows that small gaps with size 1 and 2 missing values are absent as beforehand were imputed with medians from the surroundings [-3,+3] neighbors [6]. The mean and the median of the missing data size are 8 and 7, respectively. Also, the percentage of the total missing data seems to be relatively low, around 0.045%. Larger sizes of missing data imputed using the linear interpolation method.

Table 2.3: A sample of RRI records with gap

<table>
<thead>
<tr>
<th>RRI records</th>
<th>Number of contractions</th>
</tr>
</thead>
<tbody>
<tr>
<td>944</td>
<td>54050</td>
</tr>
<tr>
<td>920</td>
<td>54051</td>
</tr>
<tr>
<td>922</td>
<td>54052</td>
</tr>
<tr>
<td><strong>928</strong></td>
<td><strong>54053</strong></td>
</tr>
<tr>
<td>910</td>
<td>54057</td>
</tr>
<tr>
<td>921</td>
<td>54058</td>
</tr>
<tr>
<td>921</td>
<td>54059</td>
</tr>
<tr>
<td>880</td>
<td>54060</td>
</tr>
<tr>
<td>888</td>
<td>54061</td>
</tr>
<tr>
<td>920</td>
<td>54062</td>
</tr>
<tr>
<td>944</td>
<td>54063</td>
</tr>
</tbody>
</table>

Figure 2.4: Missing data sizes
At last, Figure 2.5 shows ECG signals of 500 RRI recordings that belong to different age groups, namely, young, middle age, and old subjects, and for both genders female and male. A general fact is that young males tend to have a lower mean heart rate (HR) compared to elderly males and females. Also, variation in RR intervals decreases with age for both females and males [21]. In the medical field, if there is a pattern in RR variation (decrease and increase), specialists consider it normal and not pathological, and it is maybe due to sinus arrhythmia; see the young female in the upper right of the graph. On the other hand, no pattern in the variation is perhaps due to extra beats, see bottom left signal for the old male.
3 Method

The association between the heart rate variability analysis and the autonomic nervous system ANS is notable. Therefore, in this chapter, we first produce a brief explanation of the physiological HRV and its relationship with the ANS, followed by the different methods that have been employed to extract the necessary measures that reflect both short and long-term variability of the phenomenon. The physiological and mathematical correlation between HRV measures \[22\] will be explored using parametric and non-parametric tests. Also, these tests are employed to find an optimal subset of insignificantly correlated features and test their predictive ability. Finally, we present methods belonging to the supervised learning algorithms family \[23\] that are employed to achieve the thesis objective, followed by a description of the evaluation methods.

3.1 Heart Rate Variability

In this section, we summarize the important physiological background of the heart rate variability (HRV) and its association with the autonomic nervous system (ANS).

Heart rate (HR) is the number of heartbeats per minute. A healthy heart is not a metronome, and its beats do not have a regular rhythm. HR increases and decreases according to the current state of the body, for instance, eating, exercising, and sleeping, etc., and all these daily activities affect the heart rate regulations. Therefore, the fluctuation in HR, which is known by the heart rate variability (HRV), is normal and indicates the ability to adapt to environmental and psychological challenges. In general, HRV describes the variation in the time interval between consecutive heartbeats \[1\] (i.e., the time interval between two adjacent R peaks and called RR intervals), illustrated in Figure 3.1.

The autonomic nervous system (ANS), with both its branches sympathetic and parasympathetic nervous systems, regulates the heart rate. The respiratory sinus arrhythmia (RSA) is a normal physiological phenomenon that influences the heart rate as the latter increases and decreases with respiration via the sympathetic and parasympathetic systems causing gradual decrease and increase in the RR interval \[24\]. Specialists monitor the activity of the ANS, as it reflects the HRV, and any imbalance mark in ANS branches may indicate certain physiological conditions or events \[13\].
3.2 Feature Extraction Methods

The Task Force of the European Society of Cardiology and the North American Society for Pacing and Electrophysiology has determined the set of HRV standard measures that have been employed in this study. These measures are usually categorized into groups based on the duration of the analyzed RR interval records: long-term and short-term where measurements evaluated over several hours, typically 24-hours and 5-minutes recordings, respectively [13].

There are several available software packages to analyze the variation in heart rate. This analysis has used the open-source R software package RHRV [25]. Loading beat positions series of a particular ECG signal, RHRV creates a data structure and stores information belonging to the series in different fields, for instance, the date and time, RR intervals, non-interpolated heart rate, and beat positions. The mathematical expression of the inverse relationship between the instantaneous HR and the RR intervals, and all subsequent formulas in the following sections are adopted from [26]:

\[
(RR)_i = ((T)_i - (T)_{i-1}) \times 1000
\]

\[
(HR)_i = \frac{1000}{(RR)_i \times 60}
\]

\(T\) represents the time position of the \(i\)th beat. Millisecond and second are the unit measures for \(RR\) and \(T\), respectively, whereas, the HR is measured in beats per minute.

Several methods can be used to analyze the HRV. In the following sections descriptions of the methods that have been employed are given.

3.2.1 Time Domain Analysis

The statistical features in this analysis, either derived from the total record of RR intervals or 5-min segments of the recording may fall into two categories: features determined directly from the RR intervals, and others from differences between the RR intervals [27]. Following the notations for RR intervals from [26], RR and NN are used alternately. The features extracted from this analysis are:

**Mean RR**

The most straightforward feature is the mean of all RR intervals \(meanRR\) in the ECG recording, and can be defined as follows:

\[
meanRR = \frac{\sum_{i=1}^{N} (RR)_i}{N}
\]

where \(N\) is the number of RR intervals in a given series.
3.2. Feature Extraction Methods

**Mean niHR**

Another simple feature to determine is the mean of the non-interpolated heart rate $\text{meanniHR}$

$$\text{meanniHR} = \frac{\sum_{i=1}^{N} (HR)_i}{N}$$

**SDNN**

$SDNN$ described below is commonly used, which is the standard deviation of all RR intervals in a given ECG recording.

$$SDNN = \sqrt{\frac{\sum_{i=1}^{N} ((RR)_i - \text{mean RR})^2}{N - 1}}$$

Generally, SDNN depends on the length of the recording. Also, it reflects the overall variation of the series for both short-term and long-term variability [13].

**SDANN**

The standard deviation of means of RR intervals is denoted by $SDANN$ and is determined over a short period of recordings (5-min segments). It reflects the fluctuations in heart rate due to cycles longer than 5-min (i.e., estimates long-term variations) [13].

$$SDANN = \sqrt{\frac{\sum_{s=1}^{M} (\text{mean} (RR)_s - \text{mean} (RR)_{alls})^2}{M - 1}}$$

where $M$ is the total number of 5-min segments, $\text{mean} (RR)_s$ and $\text{mean} (RR)_{alls}$ are the mean of RR intervals in segment $s$ and the mean of $\text{mean} (RR)_s$ of all segments, respectively.

**SDNNIDX**

Another commonly used feature is the mean of the standard deviations of all RR intervals for each 5-min segment of the total recording $SDNNIDX$. This feature estimates the variability of the heart rate due to cycles shorter than 5-min [13].

$$SDNNIDX = \frac{\sum_{s=1}^{M} (SDNN)_s}{M}$$

where $(SDNN)_s$ is the standard deviation of RR intervals of a given segment $s$. For computing $SDANN$ and $SDNNIDX$, the RR signal is divided into successive windows of size equal to 300 seconds recording.

The following four measures that we obtained, were derived from the consecutive differences of RR intervals.

**rMSSD**

The average change in the time interval between heartbeats $rMSSD$ is defined by the square root of the mean square of adjacent differences between RR intervals [28].

$$rMSSD = \sqrt{\frac{\sum_{j=1}^{N-1} ((\delta RR)_j)^2}{N - 1}}$$
where \((\delta RR)_j = (RR)_{j+1} - (RR)_j\), for \(j = 1, ..., N - 1\) are the differences between consecutive RR intervals.

pNN50

\(pNN50\), is the percentage of the differences of adjacent RR intervals greater than 50 milliseconds.

\[ pNN50 = \frac{NN50}{N - 1} * 100 \]

where \(NN50\) is the total number of \(\delta RR_j > 50\) milliseconds.

SDSD

Another measure derived from adjacent RR interval differences and have been used in this study is \(SDSD\), which is the standard deviation of those differences.

\[ SDSD = \sqrt{\frac{\sum_{j=1}^{N-1} ((\delta RR)_j - mean \delta RR)^2}{N - 2}} \]

MADRR

Martinez et al., in [26], define \(MADRR\) as "the median of the absolute differences between adjacent RR intervals."

\[ MADRR = median(|\delta RR|) \]

HRVi

\(HRVi\) is the integral of the density of the intervals’ histogram (i.e., the total number of RR intervals) divided by the maximum of the RR intervals’ histogram. In this study, the width of bins in the RR histogram was set to 7.8125 milliseconds to compute the geometric measure HRVi [13].

\[ HRVi = \frac{N}{maximum(RR \ histogram)} \]

3.2.2 Frequency Domain Analysis

The two branches of the ANS modulates the heart rate in response to changes in body functions. As mentioned previously in section 1.1, the influence of the two branches and their response time in modulating the heart rate differ inversely. Sympathetic activity tends to increase the heart rate, and the response is slow. In contrast, parasympathetic activity tends to decrease the heart rate, and the response is faster [29]. The variation in time response of the two branches is commonly evaluated by researchers using spectral analysis. This analysis is considered as an essential tool to evaluate the ANS modulation on the heart rate [30]. It provides information on how power (i.e., the variance) of the HRV distributes as a function of frequency. The HRV spectrum can be represented by different frequency bands, high frequency (HF), low frequency (LF), very-low frequency (VLF), and ultra-low frequency (ULF) [31]. These frequency bands are known as a reflection of a particular physiological phenomenon [14] that will be highlighted later in this section.

The RHRV package creates a data structure and stores information for the HRV spectral analysis. Many techniques can be adopted to compute the frequency domain features. In
3.2. Feature Extraction Methods

In this study, the employed methods are the short-time Fourier transform (STFT), and wavelets transform (WT).

STFT uses the shifted time-window concept to analyze the signal. By applying the Fourier transform (FT) for each segment (time window), the spectral contents are estimated. In contrast, the WT uses multiple windows obtained by scaling and shifting the mother wavelet. A wavelet is a small wave that grows and decays in a specific period and has zero mean. A function satisfying these conditions is called mother wavelet. The set of obtained wavelet functions after scaling and shifting is used to estimate the time-frequency contents. Compared to STFT, WT offers a better time-frequency resolution trade-off [32, 33].

Adopting two methods was based on the fact that accurate estimation of the spectral power bands depends on the length of the analyzed recording. For instance, one cannot capture longer RR interval variations by short-term spectral analysis of frequencies in the ultra-low frequency ULF power band [13].

The RR interval signal does not have equidistant samples [34], as the RR interval varies from beat to beat. In RHRV to perform the frequency analysis, a uniformly interpolated HR series is constructed using 4 Hz frequency sampling, and a default interpolating method. For each analysis, the mean of mostly used power bands, and the mean of low/high (LF/HF) frequency ratio are estimated and described below:

**mean HF**

Mean HF is the mean of high frequency (HF) spectral power in oscillation between 0.15-0.4 Hz in ms$^2$. HF power is known to reflect the modulation of efferent parasympathetic activity by the respiratory cycle, and it is usually recorded over 2-5 minutes [8].

**mean LF**

Mean LF is the mean of low frequency (LF) spectral power in oscillation between 0.05-0.15 Hz in ms$^2$. LF power is assumed to represent both sympathetic and parasympathetic activities, and requires a typical recording period minimum of 2 minutes [8].

**mean VLF**

Mean VLF is the mean of very-low frequency (VLF) spectral power in oscillation between 0.03-0.05 Hz in ms$^2$. Researchers suggested that in normal healthy subjects, VLF power reflects sympathetic activity, and requires a recording period of at least 5-10 minutes, but maybe better analyzed over 24-hour recording [8].

**mean ULF**

Mean ULF is the mean of ultra-low frequency spectral power in oscillation between 0-0.03 Hz in ms$^2$. ULF power seems to reflect the circadian oscillation in heart rate, and requires a recording period of at least 24-hour to be estimated [13].

**mean LF/HF**

Mean LF/HF is the mean of the ratio of low frequency to high frequency power. A high LF/HF ratio refers to a higher sympathetic activity compared to parasympathetic activity. Therefore, it can be an enhancement of the sympathetic activity estimation [14].

3.2.3 Non-Linear Analysis

The heart rate mechanism is a complex process regulated by the sympathetic and parasympathetic branches of the ANS, as we mentioned earlier. These two systems are affected by...
several factors, e.g., emotions, breathing, the hemodynamics of the blood circulation, medications, and many other factors. Analyzing the heart rate variability based on the non-linear interactions between these factors provides a significant interpretation of the phenomenon \[13\]. In this study, the measures used to analyze the non-linear properties of the HRV are as follows:

**SampleEn**

Sample entropy is a regularity measurement, quantifying the complexity or irregularity of the RR time series. Higher values for SampleEn indicate a high irregular (less predictable) series, and lower values of SampleEn indicate a more predictable series \[35\]. The SampleEn can be computed as follows:

Given a sequence of RR intervals, \( RR_1, RR_2, ..., RR_N \) a set of length \( m \) vectors \( X_1, X_2, ..., X_{N-m+1} \) is reconstructed from the RR intervals such that,

\[
X_i = [RR_i, RR_{i+\tau}, ..., RR_{i+(m-1)\tau}] \tag{3.1}
\]

where \( m \) and \( \tau \) are the embedding dimension and the time lag parameters, respectively. In our study, the package nonlinearTseries \[36\] was used. The average mutual information "ami" technique is employed to estimate a proper time lag \( \tau \) \[37\]. The embedding dimension \( m \) is determined by a method based on the idea of false neighbors \[38\]. The distance \( d(X_i, X_j) \) between vector \( X_i \) and vector \( X_j \) is defined as the maximum absolute difference between the corresponding components,

\[
d(X_i, X_j) = \max(|RR_i+l - RR_j+l|, l = 0, ..., m-1)
\]

Now, for each \( X_i \) the relative number of vectors \( X_j \) satisfying \( d(X_i, X_j) \leq r \), the index \( C^m_i(r) \), where \( r \) is the similarity criterion and in the range of \( 0 < r < 1 \), is found by:

\[
C^m_i(r) = \frac{\text{number of pairs } (X_i, X_j)|d(X_i, X_j) \leq r}{N-m}, \forall j \neq i \tag{3.2}
\]

where, \( 0 < C^m_i(r) < 1 \). However, if the two vectors in the sum are not statistically independent, the estimator tends to be biased (existence of temporal correlation), and to overcome this problem, we used the so-called Theiler window, which separates two vectors by a time steps to consider them as neighbors.

Averaging \( C^m_i(r) \) over \( i \), we get:

\[
C^m(r) = \frac{\sum_{i=1}^{N-m+1} C^m_i(r)}{N-m+1} \tag{3.3}
\]

and the sample entropy is determined by:

\[
\text{SampleEn}(m, r) = \log \left[ \frac{C^m(r)}{C^{m+1}(r)} \right] \tag{3.4}
\]

**Poincaré Plot**

The Poincaré plot is a useful tool to visualize and summarize the non-linear dynamics of the cardiac system and provides interpretation of the short and long-term heart rate variability. The plot represents the dependency of a consecutive RR intervals graphically by plotting \( RR_{i+\tau} \) as a function of \( RR_i \), where \( \tau \) is a positive integer and represent the time steps used to construct the dependency relationship \[39\].
The cloud of \((RR_i, RR_{i+1})\) points are fitted by an ellipse oriented along the line of identity \(y = x\). Two parameters, SD1 and SD2, characterize this cloud of points. SD1 determines the standard deviation of the points perpendicular to the line of identity, and describes the short-term variability of the RR time series. SD2 determines the standard deviation of the points along the line of identity, and describes the long-term variability of the RR time series. In addition, the identity line \(RR_i = RR_{i+1}\) in the Poincaré plot may be physiologically explained by the location of the points, equal successive RR intervals represented by the points on the line, whereas, the points above and below it reflects the decrease and increase in the heart rate, respectively [39]. Poincaré plots for three females of different age groups is presented in Figure 3.2.

The statistical formulae of SD1 and SD2 described below, illustrate the correlation with the linear time-domain measures [39]:

\[
(SD1)^2 = \frac{1}{2}(SDSD)^2
\]  

(3.5)

and,

\[
(SD2)^2 = 2 \ast (SDNN)^2 - \frac{1}{2}(SDSD)^2
\]  

(3.6)

where \((SDSD)^2\) and \((SDNN)^2\) as described in section 3.2.1 is the standard deviation of adjacent RR intervals differences and the standard deviation of all RR intervals in a given period, respectively.

Figure 3.2: Poincaré plots derived from 4-hour RR recording belong to 3 females from different age groups

(a) young female  
(b) middle age female  
(c) old female

The statistical formulae of SD1 and SD2 described below, illustrate the correlation with the linear time-domain measures [39]:

\[
(SD1)^2 = \frac{1}{2}(SDSD)^2
\]  

(3.5)

and,

\[
(SD2)^2 = 2 \ast (SDNN)^2 - \frac{1}{2}(SDSD)^2
\]  

(3.6)

where \((SDSD)^2\) and \((SDNN)^2\) as described in section 3.2.1 is the standard deviation of adjacent RR intervals differences and the standard deviation of all RR intervals in a given period, respectively.
3.3. Statistical Analysis

**Detrended Fluctuation Analysis (DFA)**

The fluctuations in the cardiac RR intervals reflect a non-stationary time series [40]. In this study, we adopted a suitable technique to detect the long-range correlations of a non-stationary time series. The DFA estimates the correlations for different time scales as follows [41]:

First, the original RR interval time series (of length N) is integrated:

\[ y(k) = \sum_{i=1}^{k} [RR_i - \overline{RR}] , \quad k = 1, ..., N \]  \hspace{1cm} (3.7)

where \( \overline{RR} \) is the average of the corresponding RR interval time series, and is subtracted to remove the monotonic behavior in the signal. Then, the integrated series is divided into non-overlapping windows or segments of equal numbers of \( n \) RR intervals. For each window, the local trend \( y_n(k) \) is estimated by performing least-square regression. The integrated series is detrended in each window, and finally, the root-mean-square of the fluctuation detrended time series is computed as follows:

\[ F(n) = \sqrt{\frac{\sum_{k=1}^{N} [y(k) - y_n(k)]^2}{N}} \]  \hspace{1cm} (3.8)

The previous procedure for all window sizes (time scales) is repeated, and a log-log graph is used to represent the relationship of \( F(n) \) and the window size \( n \). Two fluctuation scaling exponents are estimated using regression: short-term scaling exponent \( \alpha_1 \) and long-term scaling exponent \( \alpha_2 \), where they represent the slopes of the regression lines relating log(\( F(n) \)) to log(\( n \)) within specific \( n \) range.

The resulting scaling exponents have several indications. \( \alpha = 0.5 \): the signal reflects the correlation properties similar to those of white noise; an exponent \( \alpha = 1.5 \) indicates Brownian noise (the integration of white noise) and has smoother behavior than the former; and \( \alpha = 1 \) indicates long-range correlations (reflects 1/f noise behavior) [41].

3.3 Statistical Analysis

A statistical test is a tool to explore and analyze data of a certain phenomenon, identify the relationship between the variables, and allow us to draw conclusions and to validate prior beliefs about a certain procedure. Tests can be categorized into parametric and non-parametric tests. The former follows certain assumptions most commonly is the assumption of normal distribution of the population. A non-parametric test can be considered as an alternative procedure for the parametric when a certain assumption fails. Non-parametric tests are widely used specifically for ordinal or rank data, and can be determined without relying on the underlying population distribution. For that it can also be defined as distribution-free tests [42].

In this study, to present a statistical summary of HRV features during the nocturnal rest, we computed the mean, the standard deviation, and the median of these metrics, as well as the 25th and 75th percentile. Pearson’s correlation and Kendall rank correlation coefficients were used to quantify the dependencies between HRV features and between HRV features and the human age decade of the 181 healthy subjects. The significance level was set to 0.01. In the following subsections, we give a brief description of the used methods:
3.3.1 Pearson Correlation Coefficient

The sample Pearson correlation coefficient (or known as Pearson product-moment correlation PPMC) is a parametric test statistic that measures the linear association between pairs of random variables $X$ and $Y$ and is denoted by $r_{xy}$. The Pearson correlation coefficient is symmetric, that is $r_{xy} = r_{yx}$, and it is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \quad (3.9)$$

where, $\overline{x}$ and $\overline{y}$ are the sample mean of the random variables $x$ and $y$, respectively. Pearson correlation main property is that, $-1 \leq r \leq 1$. As $r$ describes the strength and direction of the linear association between the pairs. The closer the absolute value of $r$ to 1, the stronger the correlation between $X$ and $Y$. The trend (positive or negative correlation) is determined by the sign. A value of 0 is an indication of no linear relationship between the pairs. A significant test can quantify the correlation by testing the null hypothesis, that there is no linear correlation between the random variables,

$$H_0 : \text{correlation between } X \text{ and } Y \text{ is } = 0 \quad H_1 : \text{correlation between } X \text{ and } Y \text{ is different from } \neq 0$$

For further implementation purposes, we used this test to extract a subset of insignificantly correlated features with correlation value $r < |0.9|$.

3.3.2 Kendall Rank Correlation Coefficient

Kendall Rank Correlation Coefficient is a non-parametric ranked based statistical test, and it reflects the strength of the ordinal association between pairs of variables say $X$ and $Y$ [43].

Let $(x_i, y_i)$ and $(x_j, y_j)$ where $i < j$, be a pair of observations of the joint random variables $X$ and $Y$ respectively. The pair is said to be concordant if the sign of $x_i - x_j$ and $y_i - y_j$ is the same (i.e., $(x_i - x_j)(y_i - y_j) > 0$); otherwise the pair are said be discordant. For a sample size $n$, there are $\binom{n}{2} = \frac{n(n-1)}{2}$ different pairs in the set of observations. Hence, Kendall’s $\tau$ can be defined by:

$$Kendall’s \, \tau = \frac{2 \times (\text{total no. of concordant pairs} - \text{total no. of discordant pairs})}{n(n-1)} \quad (3.10)$$

where, $-1 \leq \tau \leq 1$. A positive value in the numerator reflects a dominance of the concordant pairs, and it signifies a positive correlation between the variables $X$ and $Y$. Otherwise, a negative value reflects the dominance of discordant pairs, which signifies a negative correlation between the variables $X$ and $Y$. And a value of 0 is an indication of independency between $X$ and $Y$ [43].

3.4 Supervised Learning

One of the most significant applications for Machine Learning is the supervised classification, where the class label is known. The aim is to build a model that uses labeled instances and assign new input data to its correct class [23]. The main reasons for using supervised classification in this thesis are the richness of literature and the popularity of its algorithms in terms of complexity, interpretability, and efficacy. Many algorithms are appropriate for
analyzing the HRV time series. In this work, we adopted: support vector machine (SVM), random forests, and extreme gradient boosting (XGBoost). The last two algorithms belong to the ensemble learning family, which increases model performance by aggregating multiple models (classifiers) and produces the class of instances by majority voting for a classification problem or by averaging over all individual classifiers output for a regression problem [44]. The following sections will present a more detailed description of the algorithms.

3.4.1 Support Vector Machines

In machine learning classification, support vector machine is considered as being the state-of-the-art method. The idea of SVM is to search for an optimal hyperplane (a decision boundary) that best separates the classes by maximizing the margin with minimum risk. The margin concept proposed by this technique represents the distance between the closest input data that belongs to different classes and the hyperplanes. The data points closest to the hyperplane are called support vectors [19]. Chapelle et al., [45] described the method as follows:

For a binary classification problem, let \((x_i, y_i)\) be a set of training labeled instances where \(x_i \in \mathbb{R}^n\) and the labels \(y_i \in \{-1, +1\}, \forall i = 1, ..., N\). If the training set is “linearly separable” in input space, we want to find \(w\) and \(b\) such that,

\[
y_i(w \cdot x_i + b) \geq 1, \forall i
\]

holds. The hyperplane can be determined by,

\[
w \cdot x + b = 0
\]

see (a) in Figure 3.3. The margin which is the distance between the closest point and the hyperplane is equal to \(\frac{1}{||w||}\). The optimal hyperplane that maximizes the margin can be determined by minimizing the risk as described below:

\[
H(w) = \min_w \frac{1}{2} ||w||^2
\]

and is subject to the constraint in equation (3.11). In practical applications, there are cases where such a hyperplane can lead to “overfitting” the training data. The solution for this problem is by producing a set of non-negative slack variables \(\xi_i \geq 0, i = 1, ..., N\). These provides a soft margin that allows misclassification errors to some extent. Thus, the linearity constraint and the cost function in equations (3.11) and (3.12), respectively, can be reformulated with respect to this violation as follows:

\[
y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall i
\]

\[
H(w, \xi) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i
\]

where \(\sum \xi_i\) is an upper bound on the number of misclassified training errors. The cost function (3.14), which is subject to the constraint in (3.13) represents a trade-off between model complexity and misclassification errors. \(C\) is a constant chosen by the user. Choosing larger values for \(C\) means a higher penalty for misclassified errors, see (b) in Figure 3.3.

The solution for the optimization problem in (3.14) associated with constraint (3.13) can be achieved by introducing Lagrange multipliers \(\alpha_i, i = 1, ..., N\) (i.e., each of the constraints has
one multiplier $a_i$). Hence, solving the Lagrangian function with respect to $w$ and $b$ gives a simplified form of the optimization problem, which we want to maximize, represented by:

$$\max_{a_i} \left[ \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} a_i y_i a_k y_k x_i \cdot x_k \right]$$ (3.15)

subject to the constraints $a_i \geq 0, \forall i$ and $\sum_{i=1}^{N} a_i y_i = 0$. And since it is a quadratic function in $a_i$, hence can be efficiently solved by using quadratic programming techniques [46]. The training vector $x_i$ which corresponds to non-zero $a_i$ is called a "support vector". The vector $w$ can be defined by a linear combination of the training instances, $w = \sum_{i=1}^{N} a_i y_i x_i$ and $b$ can be found by solving $y_i (w \cdot x_i + b) = 1$.

To classify new data points, we examine the sign of the constraint in equation (3.11). And can be expressed as follows:

$$f(x) = \text{sign} \left[ \sum_{i=1}^{N} (a_i y_i x_i \cdot x) + b \right]$$ (3.16)

Some real-life applications are "not separable by a linear classifier", see (c) in Figure 3.3. Thus a more complex decision function known as the kernel-based SVM classifier [47] is proposed. In this case, the input data $x_i$ is mapped into a higher dimensional feature space through non-linear transformation $\Phi(x_i)$, then finding the optimal hyperplane that separates the data in this new feature space. Practically, following the previous expressions in the linear case and given a kernel that satisfies $K(x_i, x_k) = \Phi(x_i) \cdot \Phi(x_k)$ the kernel-based classifier becomes,

$$f(x) = \text{sign} \left[ \sum_{i=1}^{N} (a_i y_i K(x_i, x) + b \right]$$ (3.17)

In this work the binary classifier is not suitable, since predicting the age decade of humans is a multi-class problem. **Multi-Class Classifiers** are extensions of the binary classifiers, most applied approaches are: "one-vs-the others" and "one-vs-one". In "one-vs-the others", $l$ binary classifiers in $l$-class problem are trained to discriminate one class from all other classes. The highest output from the decision function in (3.16) appoints the class for the new input data. While "one-vs-one" has $\frac{l(l-1)}{2}$ classifiers and follows a voting strategy by comparing between classes and the class with the highest vote is the class appointed for the input data.

In this study, we used the SVM function from the e1071 package [48] with a radial basis kernel to classify the human age. An extensive search with 10-fold cross-validation is applied to find the cost $C$ hyper-parameter with the highest accuracy.
3.4. Supervised Learning

(a) Hyperplane example

(b) Regularization example

(c) Linearly separable and Non-separable Example

Figure 3.3: Example plots of: (a) Hyperplane, (b) Regularization with small large value for C, and (c) Linearly separable and Linearly non-separable
3.4.2 Random Forests

Random forest is an ensemble algorithm and proposed by Breiman, 2001 [49]. As we mentioned before, factors like the applicability of this method for both classification and regression problems, simplicity, and interpretability made it of great interest among researchers, as well as an essential machine learning tool for real-life applications. Random forest is a significant development of the bagging concept by Breiman, 2001, which creates a combination of uncorrelated decision trees grown by independently sampled random vectors. Then, averaging their outcome to enhance model performance in terms of variance reduction [49]. The mathematical expressions and description of the algorithm is adopted from Zhang et al., [50]:

The general structure of Random Forests is based on combined trees. Each tree depends on multiple random vectors of size \( p \), let \( X = \{ X_1, X_2, ..., X_p \} \) represents the input vector of random variables, and \( Y \) is the corresponding response. The joint distribution of \( X \) and \( Y \) is assumed to be unknown. The method intends to estimate an approximate function \( f(X) \) of the actual \( Y \) by minimizing the expected value of the loss function \( L(Y, f(X)) \). Loss function is a measure of how close \( f(X) \) to \( Y \) and the most commonly used for regression and classification problems are the squared loss function [51] and the zero-one-loss [52], respectively. The functions \( f \) that minimize the loss function in respective case are:

For the regression problem:

\[
f(X) = E(Y|X = x)
\]

and

For the classification problem:

\[
f(X) = \arg \max_{y \in Y} P(Y = y|X = x)
\]

In general, ensembles algorithms construct \( f \) in terms of a collection of base learners \( h_1(x), h_2(x), ..., h_J(x) \) which are combined and averaged in case of regression to estimate the ensemble predictor, while in classification the ensemble predictor is estimated by the majority voting, respectively defined as follows:

\[
f(X) = \frac{1}{J} \sum_{j=1}^{J} h_j(x)
\]

and,

\[
f(X) = \arg \max_{y \in Y} \sum_{j=1}^{J} I(y = h_j(x))
\]

where \( I(.) \) is the indicator function.

In random forest, the \( j \)th base learner \( h_j(X, \Theta_j) \) is a tree, where \( \Theta_j \) is a vector of random variables and the \( \Theta_j/s \) are independent. These trees partition the predictor space using a sequential binary splitting form on individual variables. Hence, the tree structure is represented by a root node, which includes the entire predictor space. Terminal nodes (i.e., the leaves), where they cannot split further reflect the final partition of the predictor space. Non-terminal nodes are nodes, for which each is split into two descendant nodes by considering all possible predictor values in every possible splits, and choosing the optimal based on a splitting criterion. Different splitting criteria are applied for regression and classification trees. For \( y_1, y_2, ..., y_n \) response values at the node, the splitting measure for regression trees is the mean squared residuals,

\[
Q = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]
where $\bar{y}$ is the predicted value at the node. The Gini index is the measure of node purity for $M$ classes in classification trees,

$$Q = \sum_{m \neq m'} (\hat{p}_m \hat{p}_{m'})$$ (3.21)

where $\hat{p}_m$ is the proportion of class $m$ observations in the node,

$$\hat{p}_m = \frac{1}{n} \sum_{i=1}^{n} I(y_i = m).$$ (3.22)

Once the split is performed for a node, all the subsequent descendant nodes are treated the same as the parent node in a recursive way. The split halts when a particular criterion is met, and a predicted value is determined for all observations in the terminal nodes by averaging the response or counting the majority votes for regression and classification problems, respectively.

To comprehend Random Forest methodology, it is essential to understand the structure of the decision tree for both regression and classification problems. As we stated earlier random forest is a tree-based learner $h_j(X, \Theta_j)$. Considering a training set of pairs $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ where $x_i$ is a $p$-dimensional predictor, and $y_i$ is the corresponding response, and a particular realization $\theta_j$ of $\Theta_j$, then the fitted tree is denoted by $\hat{h}_j(x, \theta_j, D)$ and this is the original formulation of Breiman [49].

Practically, the random element $\theta_j$ is considered implicitly, as we introduce two key concepts that gives the ensemble method this “randomness” feature. First concept, as with bagging, independent bootstrap sampling of the training data is done when building each tree. Moreover, when a bootstrap sampling is performed, some of the observations are left out, and they are called out-of-bag (OOB) data. Second, random subsets of $m$ predictors are considered when splitting each node in order to find the best split. That means random forest trains each tree on a slightly different set of predictors. The randomization assigned in bootstrap sampling and used to sample the predictors refers to the two parts of $\theta_j$. The out-of-bag data seems to be useful for estimating the generalization error and measuring the variable importance [49].

In terms of tuning parameters in a random forest, the most frequent ones are:

• "m", the number of predictor variables randomly sampled as candidates at each split. The default values are different, for the classification problem $m = \sqrt{p}$ where $p$ is the total number of variables, and for the regression problem $m = p/3$.

• As for the parameter "J", the number of trees to grow, Breiman [49] illustrated that the increment in J’s number would not affect the generalization error behavior, and it will converge to a certain limit. However, it should not be too small, to guarantee that every observation gets predicted at least a few times.

• Finally, there is the parameter "tree size", which is measured by the minimum size of the terminal nodes, with larger values causing smaller trees to be grown, or by the maximum number of terminal nodes.

Variable importance makes random forest an attractive and robust method to use for both classification and regression problems. As we mentioned earlier, the characteristic of out-of-bag data is useful in measuring variable importance; as a consequence, it is useful for variable selection and for interpreting the fitted trees. Variable importance aims to quantify the contribution of each variable predictor to the model performance [49].
In this study, we used the package caret \cite{53} in R for tuning, as it offers a great facility to tune ML algorithm parameters. Then the random forest is fitted using the Random Forest package \cite{54} and the determined values of the parameters. Also, variable importance is computed, and the ten most important features are extracted as a subset for further implementations.

### 3.4.3 Extreme Gradient Boosting

Extreme Gradient Boosting algorithm (XGBoost), developed by Chen and Guestrin, 2016 \cite{55} as a research project, has become one of the most popular machine learning algorithms recently for its fast computation and enhanced model performance.

XGBoost is a decision tree-based ensemble method, more precisely an ensemble of weak learners that uses a gradient boosting framework. It can be used for both regression and classification problems. **Boosting** is a very effective process for increasing the efficiency of the model. It implements multiple sequential steps, and in each step, a new learner (function) is added to correct the error estimated from the preceding learner until no further improvement can be obtained. **Gradient boosting** uses the concept of gradient descent to optimize the loss function. It constructs a prediction model in the form of an ensemble of weak prediction learners. **XGBoost** follows the same concepts but with minor enhancement in the regularization term \cite{55}. The mathematical expressions and description of the algorithm is adopted from Chen and Guestrin original paper \cite{55} and presented below:

For a given data set \( D = \{(x_i, y_i)\} (|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}) \) where \( n \) and \( m \) are number of examples and features, respectively, a tree ensemble model using \( K \) additive functions to predict the outcome can be expressed as,

\[
\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F
\]  

where \( K \) is the number of trees, \( f_k \) is a function corresponds to an independent tree, \( F \) is the space of all tree functions, and finally \( \hat{y}_i \) is the prediction for the corresponding input data \( x_i \). The aim now is to learn the set of functions (trees) used in the model, and it is done by minimizing the “regularized” objective,

\[
L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k), \quad f_k \in F
\]  

where the differentiable loss function \( l \) measures the difference between the predicted output \( \hat{y}_i \) and the actual value \( y_i \) and can be computed depending on the nature of the problem. The regularization term \( \Omega \) penalizes the complexity of the model and lowers the risk of over-fitting \cite{55}, and can be written as,

\[
\Omega(f) = \gamma T + \frac{1}{2} \lambda ||w||^2
\]

where \( T \) is the number of leaves in the tree. \( \lambda \) and \( \gamma \) are hyper-parameters, \( \gamma \) penalizes the number of terminal nodes, and \( \lambda \) represents \( L2 \) regularization, the higher the values, the more conservative the algorithm will be. The ensemble model in \ref{3.24} contains functions \( f \) (trees) as parameters and cannot be optimized directly. Hence, the model is trained in an additive approach (**Boosting**), starting from constant predictions \( \hat{y}_i^{(0)} = 0 \) and adding functions in an iterative way \cite{56},

\[
\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)
\]
3.4. Supervised Learning

\[ \hat{y}^{(2)}_i = f_1(x_i) + f_2(x_i) = \hat{y}^{(1)}_i + f_2(x_i) \]

\[ \hat{y}^{(t)}_i = \sum_{k=1}^l f_k(x_i) = \hat{y}^{(t-1)}_i + f_t(x_i) \]  

(3.25)

where \( \hat{y}^{(t)}_i \) is the model prediction of i-th instance at iteration t. Next is to minimize the objective function in equation (3.24) which can be re-written as,

\[ L^{(t)} = \sum_{i=1}^n \left[ l(y_i, \hat{y}^{(t-1)}_i + f_t(x_i)) + \frac{1}{2} h_t f_t^2(x_i) \right] + \Omega(f_t) + \text{constant} \]  

(3.26)

and is done by finding and adding \( f_t \) that improves the model performance the most. Applying the second order Taylor expansion for equation (3.26) quickly optimizes the objective function (faster computations) \[57\], then equation (3.26) can be rewritten as \[55\],

\[ L^{(t)} \approx \sum_{i=1}^n \left[ l(y_i, \hat{y}^{(t-1)}_i + g_t f_t(x_i)) + \frac{1}{2} h_t f_t^2(x_i) \right] + \Omega(f_t) + \text{constant} \]  

(3.27)

where \( g_t \) and \( f_t \) are the 1st and 2nd order gradient statistics on the loss function, and defined as follows,

\[ g_t = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}_i) \]

and,

\[ h_t = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}_i) \]

By removing the constants from (3.27) we get the following simplified objective at step t,

\[ \bar{L}^{(t)} = \sum_{i=1}^n \left[ g_t f_t(x_i) + \frac{1}{2} h_t f_t^2(x_i) \right] + \Omega(f_t) \]  

(3.28)

Defining \( I_j = \{ i | q(x_i) = j \} \) as the instance set of leaf j and expanding the regularization term, equation (3.28) can be reformulated as follows:

\[ \bar{L}^{(t)} = \sum_{j=1}^T \left[ G_j w^*_j + \frac{1}{2} (H_j + \lambda) w^2_j \right] + \gamma T \]  

(3.29)

which can be further simplified as,

\[ \bar{L}^{(t)} = \sum_{j=1}^T \left[ G_j w^*_j + \frac{1}{2} (H_j + \lambda) w^2_j \right] + \gamma T \]  

(3.30)

where \( G_j = \sum_{i \in I_j} g_i \) and \( H_j = \sum_{i \in I_j} h_i \). The optimal weight \( w^*_j \) of leaf j, assuming fixed structure of tree \( q(x) \) can be obtained by,

\[ w^*_j = \frac{G_j}{H_j + \lambda} \]
and the corresponding optimal loss value by,

$$L^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{(H_j + \lambda)} + \gamma T$$

(3.31)

The score function (3.31) measures the tree structure quality, and with lower values for the score, the better the structure [56]. Hence, to find the optimal tree structure \( q \), a greedy approach is applied to reduce the burden of enumerating all possible tree structures, which starts from a tree with a single leaf then adding splits iteratively. The loss reduction after adding the split is defined by,

$$L_{split}(\text{Gain}) = \frac{1}{2} \left[ \frac{G_L^2}{(H_L + \lambda)} + \frac{G_R^2}{(H_R + \lambda)} - \frac{(G_L + G_R)^2}{(H_L + H_R + \lambda)} \right] - \gamma$$

(3.32)

where the first and second term is the score for the left and right split, respectively, and the third term represents the score if the split is not performed. When the value of the first three terms (and known by the training loss reduction) in (3.32) is smaller than the penalizing term (the complexity penalty when adding an extra leaf to the structure) then the overall Gain score is negative, and it is an indication to stop splitting (since by adding further leaves will cost a reduction in the tree performance [56]). Another approach for pruning by XGBoost, is to grow the tree to the maximum depth and then prune backward until the improvement in loss function is below a certain threshold.

In terms of tuning parameters, the XGBoost package [58] in R provides different types of parameters. General parameters, defining the booster type in the model, specific parameters to enhance the performance of the selected booster (in this study the tree booster is applied), and the learning task parameters that evaluate the learning process of the booster [59]. The mlr package [60] provides a function called `makeParamSet` and allows us to make a grid search on a set of parameters to find the optimal ones for the given model.

Booster parameters contain several parameters that correspond to the tree structure. The ones used for this study are [59]:

- **max-depth**, which controls the maximum depth of the tree, with higher values, the complexity of the model increases, the higher possibility of overfitting.

- **nrounds**, which is the maximum number of trees to grow.

- The parameter \( \eta \) is the learning rate, which penalizes the instance weight after each boosting round, the lower the value, the slower the computations.

- The \( \gamma \) parameter sets a limit for minimum loss reduction needed in order to make a further split on a leaf, the larger the value, the more conservative the model will be.

- The building process stops further splitting if the leaf node has the minimum sum of instance weight (hessian) less than \( \text{min-child-weight} \).

- The parameter \( \text{subsample} \), is a ratio for randomly selecting the training instances to grow the tree, and it will avoid overfitting.

- The last parameter tuned is \( \text{colsample-bytree} \), the ratio of the selected features for each tree to grow.

XGBoost library provides evaluations for each feature contribution in the overall model performance. Gain is one of the metrics available, and as we mentioned earlier in this section, it computes the relative contribution of each feature to the model based on the total gain of
its feature splits. With higher values, the more significant the feature is for model prediction. A subset of the ten most important features is considered for further implementations.

3.5 Evaluation Methods

This section will focus on the methods that are used to evaluate the performance of the implemented models. The reason why we need to evaluate the models is to examine how well the model generalizes for an unseen data. Some well known metrics in the classification tasks such as, accuracy, precision, recall and F1-score are used in this study. In a binary classification task with the class labels Positive and Negative a confusion matrix can be represented as shown in Table 3.1. The correctly classified labels are represented by true positive (TP) and true negative (TN), and the misclassified labels are defined by false positive (FP) and false negative (FN).

Table 3.1: An example of confusion matrix for binary classification task

<table>
<thead>
<tr>
<th>Actual value</th>
<th>Predicted value</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positive (TP)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td>Negative</td>
<td>False Positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

Accuracy is one of the metrics applied to assess the performance of the models, and it is calculated by the number of correctly classified samples divided by the total number of classified samples as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.33)
\]

Precision is computed in addition to the accuracy metric where it measures the proportion of correctly classified positive samples to the total number of the positive classified samples, as described below:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (3.34)
\]

Recall and also known as sensitivity, is the proportion of actual positive cases which are correctly classified and is represented by the following formula:

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (3.35)
\]

F1-score considers both precision and recall metrics and is defined by the harmonic mean of these metrics as follows:

\[
\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.36)
\]

F1-score is considered as an effective metric for evaluating the model performance since it considers both false positive and false negative, and ranges between 0 and 1. The higher the score (close to 1), the higher the performance.
4 Results

In this chapter, the most important results of the analysis are presented.

4.1 HRV Metrics

The open-source package RHRV in the R environment is used to estimate the set of necessary HRV measurements for this analysis. The estimation is done by first loading the beat positions series of the ECG signal using the LoadBeatVector function to the HRV data structure. The data structure is created by the CreateHRVData function and stores information belonging to the series in different fields such as the estimated instantaneous heart rate signal and the RR intervals.

To extract the time analysis measures, the CreateTimeAnalysis is used by setting the size of the analyzed window to 300 seconds. Also, the width of bins in the histogram is set to 7.8125 milliseconds to compute the geometric measure $HRV_i^{[13]}$.

A preprocessing analysis, like filtering and interpolating the heart rate, is applied using FilterNIHR and InterpolateNIHR functions with default values. As mentioned earlier in section 3.2.2, RR signals do no have equidistant samples since they vary from beat to beat. Such analysis is needed in the RHRV package to perform the frequency analysis. Thus, a uniformly interpolated HR series is constructed using 4 Hz frequency sampling with a default interpolating method.

Next, spectral HRV measures were calculated to evaluate the cardiovascular autonomic function. A frequency analysis data structure is created by utilizing the CreateFreqAnalysis function, and corresponding spectral power bands are found by using CalculatePowerBand. The short-time fourier transform STFT and wavelets transform WT are the two methods adopted in this work. The size of the window used to compute the RR interval series spectrogram for STFT is 300 seconds (i.e., 5 minutes), and 30 seconds for the displacement of the window. The WT method is performed with maximum error allowed equal to 0.01 Hz. We used “la8” to represent the family and the length of the wavelet.

Finally, from Non-Linear Analysis, some measurements are estimated for further interpretation of the HRV phenomenon. A data structure is created; then, the sample entropy is estimated using CalculateSampleEntropy and EstimateSampleEntropy functions. As mentioned
in section 3.2.3 to avoid the possible existence of temporal correlation, the parameter Theiler Window is set to 30. The dependency of consecutive RR intervals is shown graphically by Poincaré plot. The time step (τ) is used to construct this dependency and is set to 1. The measures that describe the short and long-term variability of the signal are estimated. The last two non-linear measures are obtained from the detrended fluctuation analysis. For the CalculateDFA function, a range of window sizes (10-300) is used to estimate the fluctuation function. The behavior of the fluctuation functions for several signals is tested visually. The plots indicate two different regions, roughly, n < 35 and n > 35. Therefore, (10,35) and (35,200) regression ranges are used for the EstimateDFA function in order to estimate the two scaling exponents. Table 4.1 we presents the obtained set of HRV metrics of different analysis along with summary statistics for each metric. The results are rounded to two decimal digits.

Table 4.1: Summary Statistics for Time domain, Frequency domain and Non-Linear Analysis

<table>
<thead>
<tr>
<th>HRV metrics</th>
<th>Time Domain Analysis</th>
<th>Frequency Domain Analysis</th>
<th>Non-Linear Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>Mean</td>
<td>SD</td>
<td>25th%</td>
</tr>
<tr>
<td>MeanniHR</td>
<td>64.48</td>
<td>7.14</td>
<td>59.21</td>
</tr>
<tr>
<td>meanRR</td>
<td>952.51</td>
<td>108.62</td>
<td>871.67</td>
</tr>
<tr>
<td>SDNN</td>
<td>90.52</td>
<td>33.03</td>
<td>67.37</td>
</tr>
<tr>
<td>SDANN</td>
<td>58.42</td>
<td>25.78</td>
<td>41.83</td>
</tr>
<tr>
<td>pNN50</td>
<td>37.91</td>
<td>19.23</td>
<td>23.32</td>
</tr>
<tr>
<td>SDSI</td>
<td>21.52</td>
<td>12.26</td>
<td>16.00</td>
</tr>
<tr>
<td>LFmean-FT</td>
<td>853.97</td>
<td>601.81</td>
<td>430.47</td>
</tr>
<tr>
<td>VLFmean-FT</td>
<td>174.24</td>
<td>129.10</td>
<td>80.63</td>
</tr>
<tr>
<td>HFmean-FT</td>
<td>276.86</td>
<td>245.96</td>
<td>90.60</td>
</tr>
<tr>
<td>LFHFmean-FT</td>
<td>0.34</td>
<td>3.48</td>
<td>1.64</td>
</tr>
<tr>
<td>ULFmean-Wav</td>
<td>6409.30</td>
<td>5230.73</td>
<td>3272.48</td>
</tr>
<tr>
<td>VLFWave-Wav</td>
<td>420.73</td>
<td>294.19</td>
<td>177.28</td>
</tr>
<tr>
<td>LFHFmean-Wav</td>
<td>8.90</td>
<td>4.64</td>
<td>5.60</td>
</tr>
<tr>
<td>SampleEN</td>
<td>0.37</td>
<td>0.13</td>
<td>0.28</td>
</tr>
<tr>
<td>SD1</td>
<td>26.36</td>
<td>13.59</td>
<td>15.94</td>
</tr>
<tr>
<td>SD2</td>
<td>124.60</td>
<td>45.78</td>
<td>92.31</td>
</tr>
<tr>
<td>alpha1</td>
<td>1.26</td>
<td>0.14</td>
<td>1.16</td>
</tr>
<tr>
<td>alpha2</td>
<td>0.98</td>
<td>0.10</td>
<td>0.92</td>
</tr>
</tbody>
</table>
4.2 Statistical Evaluations

Two tests are conducted to determine the correlation within HRV features, and between the features and the human age decade. The tests are, Pearson’s and Kendall rank correlation. The level of significance is set to 0.01 to quantify this dependency. In Figure 4.1, bar plots of the two tests visualize the relationship of HRV features with the human age decade. There is a clear indication of both positive and negative behavior between the corresponding components, where the negative is more dominant. Also, strong correlations are observed for frequency analysis measures: LFmean-FT, LFmean-Wav, HFmean-Wav, and HFmean-FT, as well as SDNNIDX from the time-domain analysis.

In Figure 4.2, heat maps show the correlation between the HRV features, also the p-value of insignificantly correlated features for two tests: (a) (Pearson r), (b) (Kendall rank). The dark red regions with correlation values close to -1 indicate the highest negative correlation between the variables. Dark blue regions with values close to 1 indicate the highest positive correlation, whereas, white regions with values close to 0 indicate insignificant correlation (no correlation between HRV features). The difference between the two tests is illustrated in Figure 4.3. Also, in Table 4.2, The subset of insignificant features with Pearson correlation < 0.9 are presented. This subset will be used later to examine and compare the performance of the implemented models.

---

**Figure 4.1:** Bar plot of: (a) Pearson Correlation of HRV metrics with age decade, (b) Kendall rank Correlation of HRV metrics with age decade
4.2. Statistical Evaluations

(a) Pearson correlation

(b) Kendall rank correlation

Figure 4.2: Heat maps of the correlated HRV features with P-values of the insignificantly correlated ones for two tests: (a) (Pearson r), (b) (Kendall rank)
4.3 Feature Importance

For further interpretation of the fitted model, ensemble methods have a useful characteristic as they evaluate each feature contribution in model performance. Features with high scores are considered as most important and have more impact on model prediction. Random Forest and XGBoost implementations compute the variable importance during training of the model by using importance and xgb.importance functions, respectively.

Figure 4.4 shows feature importance for both random forest and XGBoost classifiers. For random forest, mean decrease accuracy was used to evaluate importance. On the other hand, the gain metric was used for XGBoost. Both approaches show clusters of features where the frequency domain’s HRV features have the highest values. Noticeable variation in the two approaches, illustrated in the dominance of HFmean-Wav feature as it has been given weight with XGBoost less than the random forests where it is the most important. This variation is mostly due to the different approaches of each method in terms of selecting the feature to further growing the tree, so we should expect them to behave slightly differently. The
Table 4.3 and 4.4 are presenting the two extracted subsets with ten most important features of random forest and XGBoost models.

Figure 4.4: HRV feature importance in: (a) Random Forest model, (b) XGBoost model
4.4 Model Performance and Evaluation

In this section, the performance of the fitted classifiers described in section 3.4 for predicting the age decade is presented. For the purpose of the analysis, the data is preprocessed to consist of 181 observations across 26 variables, where 25 variables are most frequently used HRV features extracted using methods described in section 3.2 with statistical summary in Table 4.1. The variable age served as the dependent variable.

First, the dataset is randomly split into train and test datasets, with a ratio of (60% : 40%). In other words, approximately 109 of the observations are devoted to training the models and 72 for testing. This procedure is applied to the original obtained set of HRV features and for a different combination of feature subsets described in Table 4.2, 4.3 and 4.4. Next, for each classifier used in this analysis, and as there is no previous knowledge on which optimal combination of parameters is best to build a good classifier, an extensive search is performed for all parameters.

The SVM is fitted using svm function from e1071 package with a radial basis kernel to classify the human age using parameter $\gamma$ in the range of 0.1 to 1 with step size equal to 0.1. 10-fold cross-validation is applied to find the optimal cost hyper-parameter $C$. Large values for $C$ will commit overfitting, and to avoid this problem, the built-in function tune and a range of values for $C$ from 0.1 to 5.1 with step size 0.5 were used.

Again, we fitted the predictors against the dependent variable with randomForest function from the Random Forest package. To find the optimal parameters described in section 3.4.2, we used the train function from the caret package and set a range of values for each. The number of predictor variables at each split $mtry$ is set to a range between 1 to 15. The range for the maximum number of terminal nodes in a tree $maxnodes$ is between 5 and 20. Finally, the range for the number of trees to grow $ntree$ is from 25 to 550, with a step size of 25.

Finally, an XGBoost classifier is fitted using the XGBoost package. A grid search using the makeParamSet function from MLR package is applied to find the optimal parameters described in section 3.4.3. The maximum depth of tree $max-depth$ range is set between 3 and 10, the maximum number of trees to grow $nrounds$ is between 20 to 100. The range for the learn-
ing rate $\eta$ is between 0.1 and 0.3, while $\gamma$ is in range of 0.1 to 3. \textit{min-child-weight} threshold is between 1 to 10. The ratio range of a random selection for the training samples \textit{subsample} is between 0.5 and 1. Lastly, the range of the subsample ratio for selecting the predictors \textit{colsample-bytree} to grow each tree is between 0.5 and 1.

Table 4.5 shows the accuracy of each fitted classifier using train, test and a shuffled version of the dataset as validation sets, and for each set of obtained features. Values with bold highlights represent the highest accuracy achieved for a specific classifier with a particular validation set and a particular set of features. In general, the overall result indicates low accuracy for all cases in predicting the age class.

SVM classifier had the lowest accuracy among all while others performed slightly better with test accuracy between 20\% and 30\%. The main idea of the XGBoost classifier is to reduce the overall error, and thus one would expect that it would perform better. Nevertheless, the test accuracy for the classifier with the set of ten best features from random forest is less than the Random Forest classifier with the same set. Also, the accuracy of XGBoost for the train and shuffled data is approximately equal to 79\% and 80\%, respectively. These values indicate a tendency of overfitting and mainly due to the small dataset. Although, the differences between the two classifiers are minor and negligible, but the winning combination seems to be the Random Forest with its ten most important features set.

Table 4.5: Feature-Based Classifiers’ Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Validation set</th>
<th>All Features</th>
<th>RF Best10</th>
<th>XGBoost Best10</th>
<th>Insignificant features</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>Train data %</td>
<td>73.14815</td>
<td>63.88889</td>
<td>65.74074</td>
<td>64.51481</td>
</tr>
<tr>
<td></td>
<td>Shuffled Data %</td>
<td>70.71823</td>
<td>47.51381</td>
<td>53.59116</td>
<td>47.51381</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Train data %</td>
<td>75.92593</td>
<td>72.22222</td>
<td>72.22222</td>
<td>78.70370</td>
</tr>
<tr>
<td></td>
<td>Test data %</td>
<td>23.28767</td>
<td>28.76712</td>
<td>23.28767</td>
<td>26.02740</td>
</tr>
<tr>
<td></td>
<td>Shuffled Data %</td>
<td>61.32597</td>
<td>67.40331</td>
<td>53.59116</td>
<td>70.71823</td>
</tr>
<tr>
<td>XGBoost</td>
<td>Train data %</td>
<td>78.70370</td>
<td>67.59259</td>
<td>66.66667</td>
<td>71.29630</td>
</tr>
<tr>
<td></td>
<td>Test data %</td>
<td>26.02740</td>
<td>23.28767</td>
<td>23.28767</td>
<td>26.02740</td>
</tr>
<tr>
<td></td>
<td>Shuffled Data %</td>
<td>70.71823</td>
<td>49.17127</td>
<td>61.87845</td>
<td>79.55801</td>
</tr>
</tbody>
</table>

Figure 4.5 presents confusion matrices of SVM, Random Forests, and XGBoost classifiers with train, test and shuffled data sets as validation sets. Using the original set of features (all features). The age classes 20\textsuperscript{th}, 30\textsuperscript{th}, ..., 80\textsuperscript{th} are represented with 0,1,...,6, respectively. Regions with a light blue color indicate a higher number of classified data, while regions with darker color reflect the failure of the classifier on predicting the age.

Observing the performance of the classifiers with the test data, and despite the number of correctly classified classes, the best performance was for predicting the middle age (i.e., 40\textsuperscript{th} and 50\textsuperscript{th} age decades); while the performance declines for the young and old age. Shuffled data in all classifiers seems to have the same behaviour of the train data, with a slight variation in the number of correctly classified data.

All confusion matrices for different classifiers with different combination set of features can be found in the Appendix. More interpretation of the confusion matrices are presented and summarized in Table 4.6 for train and test data, and in Table 4.7 for the shuffled data. Since the accuracy is not a preferable estimator when it comes to an uneven class distribution, we
adopted three metrics that are widely used in classification problems to evaluate the model performance. The metrics are the precision, recall, and F1-score. The values show cases when the classifiers fail in identifying the true positive cases of the age.

Figure 4.5: Confusion matrices of SVM, Random Forests, and XGBoost classifiers with train, test, and shuffled data sets as validation sets. Using the original set of features (all features) as predictors. 0,1,...,6 represent 20th, 30th, ..., 80th age decades, respectively.
### 4.4. Model Performance and Evaluation

Table 4.6: SVM, Random Forest and XGBoost classifiers performance with both train and test data as validation sets. Using different combinations of HRV set of features as predictors: original set of features (all features), Random Forest and XGBoost ten most important features, also insignificantly correlated set of features from Pearson’s correlation test.

<table>
<thead>
<tr>
<th>Class</th>
<th>Train Precision</th>
<th>Train Recall</th>
<th>Train F1</th>
<th>Test Precision</th>
<th>Test Recall</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>0.774</td>
<td>0.923</td>
<td>0.842</td>
<td>0.167</td>
<td>0.750</td>
<td>0.237</td>
</tr>
<tr>
<td>30-39</td>
<td>1.000</td>
<td>0.364</td>
<td>0.353</td>
<td>NA</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>40-49</td>
<td>0.692</td>
<td>0.900</td>
<td>0.793</td>
<td>0.182</td>
<td>0.500</td>
<td>0.229</td>
</tr>
<tr>
<td>50-59</td>
<td>0.452</td>
<td>0.813</td>
<td>0.650</td>
<td>0.222</td>
<td>0.267</td>
<td>0.242</td>
</tr>
<tr>
<td>60-69</td>
<td>1.000</td>
<td>0.500</td>
<td>0.667</td>
<td>0.250</td>
<td>0.077</td>
<td>0.118</td>
</tr>
<tr>
<td>70-79</td>
<td>0.700</td>
<td>0.636</td>
<td>0.647</td>
<td>0.273</td>
<td>0.273</td>
<td>0.273</td>
</tr>
<tr>
<td>80-85</td>
<td>1.000</td>
<td>0.600</td>
<td>0.750</td>
<td>NA</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>20-29</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
</tr>
<tr>
<td>30-39</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>40-49</td>
<td>0.375</td>
<td>0.375</td>
<td>0.375</td>
<td>0.375</td>
<td>0.375</td>
<td>0.375</td>
</tr>
<tr>
<td>50-59</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
</tr>
<tr>
<td>60-69</td>
<td>0.313</td>
<td>0.313</td>
<td>0.313</td>
<td>0.313</td>
<td>0.313</td>
<td>0.313</td>
</tr>
<tr>
<td>70-79</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
</tr>
<tr>
<td>80-85</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
</tr>
</tbody>
</table>

#### (a) SVM Model

<table>
<thead>
<tr>
<th>Class</th>
<th>Train Precision</th>
<th>Train Recall</th>
<th>Train F1</th>
<th>Test Precision</th>
<th>Test Recall</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>0.686</td>
<td>0.923</td>
<td>0.797</td>
<td>0.188</td>
<td>0.750</td>
<td>0.300</td>
</tr>
<tr>
<td>30-39</td>
<td>1.000</td>
<td>0.182</td>
<td>0.308</td>
<td>NA</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>40-49</td>
<td>0.938</td>
<td>0.750</td>
<td>0.833</td>
<td>0.286</td>
<td>0.308</td>
<td>0.296</td>
</tr>
<tr>
<td>50-59</td>
<td>0.536</td>
<td>0.938</td>
<td>0.682</td>
<td>0.130</td>
<td>0.200</td>
<td>0.158</td>
</tr>
<tr>
<td>60-69</td>
<td>1.000</td>
<td>0.643</td>
<td>0.793</td>
<td>0.333</td>
<td>0.077</td>
<td>0.125</td>
</tr>
<tr>
<td>70-79</td>
<td>0.900</td>
<td>0.818</td>
<td>0.857</td>
<td>0.357</td>
<td>0.455</td>
<td>0.400</td>
</tr>
<tr>
<td>80-85</td>
<td>1.000</td>
<td>0.800</td>
<td>0.899</td>
<td>0.333</td>
<td>0.143</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>20-29</td>
<td>0.722</td>
<td>0.722</td>
<td>0.722</td>
<td>0.722</td>
<td>0.722</td>
<td>0.722</td>
</tr>
<tr>
<td>30-39</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>40-49</td>
<td>0.889</td>
<td>0.889</td>
<td>0.889</td>
<td>0.889</td>
<td>0.889</td>
<td>0.889</td>
</tr>
<tr>
<td>50-59</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
</tr>
<tr>
<td>60-69</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
</tr>
<tr>
<td>70-79</td>
<td>0.818</td>
<td>0.818</td>
<td>0.818</td>
<td>0.818</td>
<td>0.818</td>
<td>0.818</td>
</tr>
<tr>
<td>80-85</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

#### (b) Random Forest Model

<table>
<thead>
<tr>
<th>Class</th>
<th>Train Precision</th>
<th>Train Recall</th>
<th>Train F1</th>
<th>Test Precision</th>
<th>Test Recall</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>0.774</td>
<td>0.923</td>
<td>0.842</td>
<td>0.250</td>
<td>0.750</td>
<td>0.375</td>
</tr>
<tr>
<td>30-39</td>
<td>0.889</td>
<td>0.727</td>
<td>0.800</td>
<td>0.500</td>
<td>0.100</td>
<td>0.167</td>
</tr>
<tr>
<td>40-49</td>
<td>0.888</td>
<td>0.800</td>
<td>0.842</td>
<td>0.353</td>
<td>0.456</td>
<td>0.400</td>
</tr>
<tr>
<td>50-59</td>
<td>0.722</td>
<td>0.813</td>
<td>0.765</td>
<td>0.130</td>
<td>0.200</td>
<td>0.158</td>
</tr>
<tr>
<td>60-69</td>
<td>0.600</td>
<td>0.571</td>
<td>0.667</td>
<td>0.500</td>
<td>0.077</td>
<td>0.133</td>
</tr>
<tr>
<td>70-79</td>
<td>0.600</td>
<td>0.818</td>
<td>0.692</td>
<td>0.333</td>
<td>0.364</td>
<td>0.348</td>
</tr>
<tr>
<td>80-85</td>
<td>1.000</td>
<td>0.700</td>
<td>0.824</td>
<td>0.200</td>
<td>0.143</td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>20-29</td>
<td>0.727</td>
<td>0.727</td>
<td>0.727</td>
<td>0.727</td>
<td>0.727</td>
<td>0.727</td>
</tr>
<tr>
<td>30-39</td>
<td>0.875</td>
<td>0.875</td>
<td>0.875</td>
<td>0.875</td>
<td>0.875</td>
<td>0.875</td>
</tr>
<tr>
<td>40-49</td>
<td>0.857</td>
<td>0.857</td>
<td>0.857</td>
<td>0.857</td>
<td>0.857</td>
<td>0.857</td>
</tr>
<tr>
<td>50-59</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
</tr>
<tr>
<td>60-69</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>70-79</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>80-85</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
</tbody>
</table>

#### (c) XGBoost Model
Table 4.7: SVM, Random Forest, and XGBoost classifiers performance with shuffled data as a validation set. Using different combinations of HRV set of features as predictors: original set of features (all features), Random Forest and XGBoost ten most important features, also insignificantly correlated set of features from Pearson’s correlation test.

<table>
<thead>
<tr>
<th>Class</th>
<th>All features</th>
<th>RF Best10 features</th>
<th>XGBoost Best10 features</th>
<th>Insignificantly correlated features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
</tr>
<tr>
<td>20-29</td>
<td>0.730</td>
<td>0.900</td>
<td>0.806</td>
<td>0.632</td>
</tr>
<tr>
<td>30-39</td>
<td>1.000</td>
<td>0.333</td>
<td>0.500</td>
<td>1.000</td>
</tr>
<tr>
<td>40-49</td>
<td>0.839</td>
<td>0.788</td>
<td>0.813</td>
<td>0.500</td>
</tr>
<tr>
<td>50-59</td>
<td>0.350</td>
<td>0.710</td>
<td>0.620</td>
<td>0.341</td>
</tr>
<tr>
<td>60-69</td>
<td>0.680</td>
<td>0.630</td>
<td>0.654</td>
<td>0.364</td>
</tr>
<tr>
<td>70-79</td>
<td>0.600</td>
<td>0.818</td>
<td>0.692</td>
<td>0.464</td>
</tr>
<tr>
<td>80-85</td>
<td>1.000</td>
<td>0.647</td>
<td>0.786</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(a) SVM Model

<table>
<thead>
<tr>
<th>Class</th>
<th>All features</th>
<th>RF Best10 features</th>
<th>XGBoost Best10 features</th>
<th>Insignificantly correlated features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
</tr>
<tr>
<td>20-29</td>
<td>0.558</td>
<td>0.800</td>
<td>0.656</td>
<td>0.632</td>
</tr>
<tr>
<td>30-39</td>
<td>1.000</td>
<td>0.095</td>
<td>0.134</td>
<td>1.000</td>
</tr>
<tr>
<td>40-49</td>
<td>0.575</td>
<td>0.697</td>
<td>0.630</td>
<td>0.634</td>
</tr>
<tr>
<td>50-59</td>
<td>0.477</td>
<td>0.677</td>
<td>0.560</td>
<td>0.634</td>
</tr>
<tr>
<td>60-69</td>
<td>0.833</td>
<td>0.556</td>
<td>0.667</td>
<td>0.727</td>
</tr>
<tr>
<td>70-79</td>
<td>0.680</td>
<td>0.773</td>
<td>0.723</td>
<td>0.708</td>
</tr>
<tr>
<td>80-85</td>
<td>1.000</td>
<td>0.529</td>
<td>0.692</td>
<td>1.000</td>
</tr>
</tbody>
</table>

(b) Random Forest Model

<table>
<thead>
<tr>
<th>Class</th>
<th>All features</th>
<th>RF Best10 features</th>
<th>XGBoost Best10 features</th>
<th>Insignificantly correlated features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
</tr>
<tr>
<td>20-29</td>
<td>0.684</td>
<td>0.867</td>
<td>0.765</td>
<td>0.571</td>
</tr>
<tr>
<td>30-39</td>
<td>0.583</td>
<td>0.333</td>
<td>0.424</td>
<td>0.250</td>
</tr>
<tr>
<td>40-49</td>
<td>0.735</td>
<td>0.758</td>
<td>0.746</td>
<td>0.538</td>
</tr>
<tr>
<td>50-59</td>
<td>0.750</td>
<td>0.677</td>
<td>0.712</td>
<td>0.389</td>
</tr>
<tr>
<td>60-69</td>
<td>0.010</td>
<td>0.815</td>
<td>0.759</td>
<td>0.480</td>
</tr>
<tr>
<td>70-79</td>
<td>0.667</td>
<td>0.727</td>
<td>0.696</td>
<td>0.517</td>
</tr>
<tr>
<td>80-85</td>
<td>0.786</td>
<td>0.647</td>
<td>0.710</td>
<td>0.333</td>
</tr>
</tbody>
</table>

(c) XGBoost Model
5 Discussion

5.1 Heart rate variability as an age indicator

In a normal healthy heart, the RR interval continuously fluctuates between decelerations and accelerations to adapt to environmental and physiological challenges. This process is regulated by the autonomic nervous system ANS and its branches the sympathetic and parasympathetic. The sympathetic is known to increase the heart rate and the parasympathetic tends to decrease the heart rate [24]. Several metrics analyzing the HRV with an attempt to quantify the ANS activity have been proposed [13]. In this study, the adopted measures belonging to different analyses are summarized in Table 4.1.

As shown in Figure 4.2, the time domain measures correlate with each other such as, rMSSD, pNN50, MADRR, SDSD, and SDNNIDX where they reflect an estimate of the short-term variability in the RR interval signal. The SDANN is an estimate of the long-term variability, and SDNN and HRV are estimates of the overall heart rate variability. For choosing the measures, one should consider the analysis type, also consider the role of each measure. For instance, the measures that estimate the overall HRV, long-term, or short-term variability are not reasonable to be replaced with each other [13]. On the other hand, the power bands that have been employed in this study to estimate frequency domain measures are widely used in HRV analysis by researchers, especially for adults in their supine rest or during performing a low-level activity. However, it might not be suitable for a high-level activity or pathological conditions. For instance, HF power will not be properly estimated. Hence, it cannot accurately reflect the parasympathetic activity since the HF will expand and exceed the adopted upper band limit in the case of high-level activity. So, further investigations are needed in terms of the proper selection of the power bands by considering the analysis type.

In agreement with previous studies [22], the Pearson’s and Kendall rank correlation tests in this analysis illustrate the relationship between the measures of the time and frequency domains in terms of the short-term and long-term heart rate variability, see Figure 4.2. The overall and the long-term variability measures SDNN and SDANN show a strong correlation with the ultra-low frequency (ULF) power as well as the long-term nonlinear measure SD2. SDNN also showed a relationship with the very-low (VLF) and low (LF) frequency powers. The short-term measures rMSSD and pNN50 were correlated with high frequency (HF) power as well as the short-term nonlinear measure SD1.
As we mentioned previously in section 1.1, the dynamics of a cardiac RR interval vary with age in a healthy system. Aging leads to a reduction in both the overall HRV and the complexity of the physiological dynamics. The results of the correlation analysis regardless of the correlation strength confirm this association, showing a negative relationship between the age and most of the HRV parameters, see Figure 4.1.

Considering the previous paragraph, the possibility of answering our research question of which of these statistical HRV features are optimal to classify the human age has been explored in different approaches. These approaches are the Pearson correlation test and the variable importance characteristic of the random forest and XGBoost classifiers. For Pearson correlation, we set a threshold and obtained a subset of features by removing the irrelevant and redundant ones. The second approach is used in an attempt to reduce the random search for the optimal features. Subsets of highly scored features were extracted, see Figure 4.4 using different methods and different assessments of the feature contribution in model prediction for both Random Forests and XGBoost. However, alternative approaches can be adopted, and further investigations can be performed to find the optimal features.

Tables 4.2, 4.3 and 4.4 show the extracted subsets of features. The classifiers showed a slight improvement when considering the set of optimal features obtained from the random forest model, achieving higher accuracy among all the extracted subsets of features. This subset of features also showed the strongest relationship with the age in Figure 4.1. The measures obtained from the spectral analysis are the most dominant in the set. These measures are known as effective estimates for reflecting the dynamic fluctuations of the cardiac autonomic function in the short and long-term variability. Also, they can provide information on cardiac stress or diseases [16].

In this analysis, we aimed to examine the predictive ability in defining the age class using different classifiers. As first intuition, from the correlation analysis and studying the bar plots, almost all of the HRV features have shown some dependency (positively and negatively) with the class (i.e., the age decade). One can assume that these features hold some useful information to classify the age of each healthy individual. But observing the performance of different adopted classifiers using different set of features, the overall results indicates poor performance in capturing the true class. The highest test accuracy obtained was for the random forest classifier with approximately 29% using the subset of its ten best features.

So, the primary research question of whether the classification task of the human age can be achieved accurately and based on this analysis outcome remains unanswered. Several reasons might lead to this poor result, and it cannot be presented in a solid context; thus, further investigations are needed:

- having only the RR interval series or having inaccurate labels as age decades instead of the precise age of the subject, or the data is rather small, might not be enough to extract the relevant information necessary to predict the desired output.

- Based on the non-stationary behavior of the RR intervals time series, which is affected by several previously mentioned factors, one might think of adopting extra features related to cardiovascular mechanism. In other words, to consider features in terms of acceleration and deceleration of the heart rate (i.e., non-linear features) to interpret more the fluctuations in the cardiac RR intervals, and give extra measures of the time series irregularity.

- Also, the four-hour RR intervals series recorded during the nocturnal rest might not be enough to capture the most important information of the cardiovascular regulation, especially the ones that are embedded in longer periods. So, longer recording periods (24-hour) may be needed for HRV metrics to estimate a better reflection of the long-term variability of the heart rate.
Another factor, such as the gender, might be taken into consideration when analyzing the HRV. Many researchers illustrated the association of the gender with the cardiac regulations, as the adult females tend to have a higher average heart rate than males, between (78% – 82%) and (70% – 72%) beats per minute, for female and male respectively. This difference is mainly due to the smaller size of the female heart compared to the male; the female’s heart needs to beat at a faster rate to pump the required amount of blood. Also, studies showed that both this gender difference and heart rate pace tend to decrease after the age of (40 – 50), $^{21}$. Thus, the gender factor is considered significant, and it is usually reflected in the HRV measures from the frequency domain analysis. In general, the females tend to have a higher HF power band and total power compared to the males $^{30}$.

Despite the fact that the feature-based methods we used are robust and efficient in terms of the clear correlation existence, which is defined earlier in section 4.2 between the extracted HRV features and the output we aimed to quantify. ML provides a wide range of powerful predictive tools that can serve as an alternative approach for the adopted methods in this analysis. The neural network methods may be suitable to analyze the HRV for further either confirmation or refutation for the obtained results.

To validate the findings in this study, and in comparison with Makowiec et al. $^{6}$ research mentioned in section 1.1, where they performed an exhaustive analysis in many aspects of the phenomenon and obtained some profitable findings. However, they conducted a classification task using the linear and non-linear SVM with Gaussian kernel to classify the age of 181 healthy subjects. They claimed that, fitting the classifiers for different sets of age groups with different sets of features would enhance the model prediction. Also, assumed random shuffling of the original signal is a suitable approach to validate the model outcome. A higher score achieved using this approach was around 93.6 ± 5.3% with the non-linear SVM classifier and a specific set of age decades. This type of validation might overfit the training. Instead, they could have adopted the simple hold-out approach used in this study to produce an unbiased estimate of the model performance or any other alternative approach. Comparing the results obtained in this study, when considering the shuffled dataset to validate the models, the accuracies with all combinations of feature subsets showed higher values than the test data, see Table 4.5. Moreover, they used C=1, and it is known that the higher values for the cost parameter C, the higher penalty for the misclassified errors, the more likely to overfit. In this study, the SVM with the set of features belong to RF best10, and XGBoost best10 had C values higher than 1 with test accuracies approximately to 14% and 19%, respectively. Hence, the results obtained in their study cannot be generalized, and the findings in this study seemed to be acceptable.
Conclusion

In this study, we aimed to answer the original research questions, we explored the effect of the statistical features derived from the RR interval series, and which of these features are the most optimal ones. We found different subsets from the original set of features obtained by different methods. The predictive ability of the classifiers showed a slight improvement when a certain set of features are considered. More precisely, the subset of ten most optimal features obtained from the random forest model, showed higher accuracy among all the extracted subsets of features with approximately 29%. And it is noteworthy that the measures obtained from the spectral analysis are the most dominant in the set. The reason might be due to the significant role of frequency power bands in assessing the dynamic fluctuations in cardiac autonomic function over any given period (short and long-term period). They can provide information about both the frequency and amplitude of certain fluctuations in the signal. Thus, it can be used as a tool to reflect cardiac events. Hence, for this analysis the set of features consisting of SDNNIDX, MADRR, LFmean-FT, HFmean-FT, LFHFmean-FT, VLFmean-Wav, LFmean-Wav, HFmean-Wav, LFHFmean-Wav and alpha1 considered as an optimal set of features.

Methods used to analyze and extract HRV measures are studied thoughtfully and thoroughly by researchers and showed a significant interpretation of the phenomenon in many aspects of cardiac-related events. Employing HRV measures as signs of predicting the age might enhance age-related events, such as reducing mortality and increasing life cycle. In this study, multiple approaches conducted to analyze the HRV phenomenon and different feature-based classifiers performed to define the age class of a new (unseen) observation. The results for predicting the age showed a constant output, which is significantly low. The overall test accuracy for all models was lower than 30%. SVM classifier had the lowest performance, while the random forests and XGBoost performed slightly better. Although the difference is negligible between the last two, but the random forest, as we mentioned earlier, had a higher accuracy with 29%. XGBoost showed the highest tendency of overfitting with 79% for training data and 80% for shuffled data, and that might be due to the smaller size of the given dataset, and also, depends on the type of the studied problem.

Findings in this analysis are significantly lower than other previous studies conducted for the same purpose. It is hard to elicit conclusions other than the fact that the dataset has a significant role in prediction output. The main drawback to the Gdańsk dataset was possibly
the size, 181 volunteers where 109 are used for training. Models will not generalize well, and risks of overfitting might highly occur, having a small number of observations to train the model. The age decade might not be a proper age class rather than having the real age, particularly with such a small size of data. Another outcome is that ECG recordings with longer periods of RR intervals of more than four-hour during nocturnal rest is probably needed to capture the longer variability of the cardiac autonomic function. Thus, we recommend when further analysis is conducted:

- To consider longer periods (24-hour recordings) to provide more information of the heart rate variability, since the patients’ cardiac will be subject to many events during the day.
- Test with different larger datasets, considering the true age of the human.
- More features for extra interpretation of the irregularity of the time series can be extracted.
- Also, the models adopted in this study are feature-based and required feature engineering. As an alternative, models that learn features automatically can be adopted, such as a recurrent neural network (RNN).
Appendix

Figure: Confusion matrices of the SVM classifier with both train and test sets as validation sets. Using different combinations of features as predictors: Random Forest and XGBoost ten most important features, also insignificantly correlated set of features from Pearson’s correlation test. 0,1,...,6 represent 20th, 30th,..., 80th age decades, respectively.
Figure: Confusion matrices of the Random Forests classifier with both train and test sets as validation sets. Using different combinations of features as predictors: Random Forest and XGBoost ten most important features, also insignificantly correlated set of features from Pearson’s correlation test. 0,1,...,6 represent 20th, 30th,..., 80th age decades, respectively.

Figure: Confusion matrices of the XGBoost classifier with both train and test sets as validation sets. Using different combinations of HRV set of features as predictors: Random Forest and XGBoost ten most important features, also insignificantly correlated set of features from Pearson’s correlation test. 0,1,...,6 represent 20th, 30th,..., 80th age decades, respectively.
Figure: Confusion matrices of the SVM (1st row), Random Forests (2nd row), and XGBoost (last row) classifiers with shuffled data as a validation set. Using different combinations of HRV set of features as predictors: Random Forest and XGBoost ten most important features, also insignificantly correlated set of features from Pearson’s correlation test. 0,1,...,6 represent 20th, 30th,..., 80th age decades, respectively.


[58] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. `xgboost`: Extreme Gradient Boosting. R package version 0.90.0.2. 2019. URL: [https://CRAN.R-project.org/package=xgboost](https://CRAN.R-project.org/package=xgboost).
