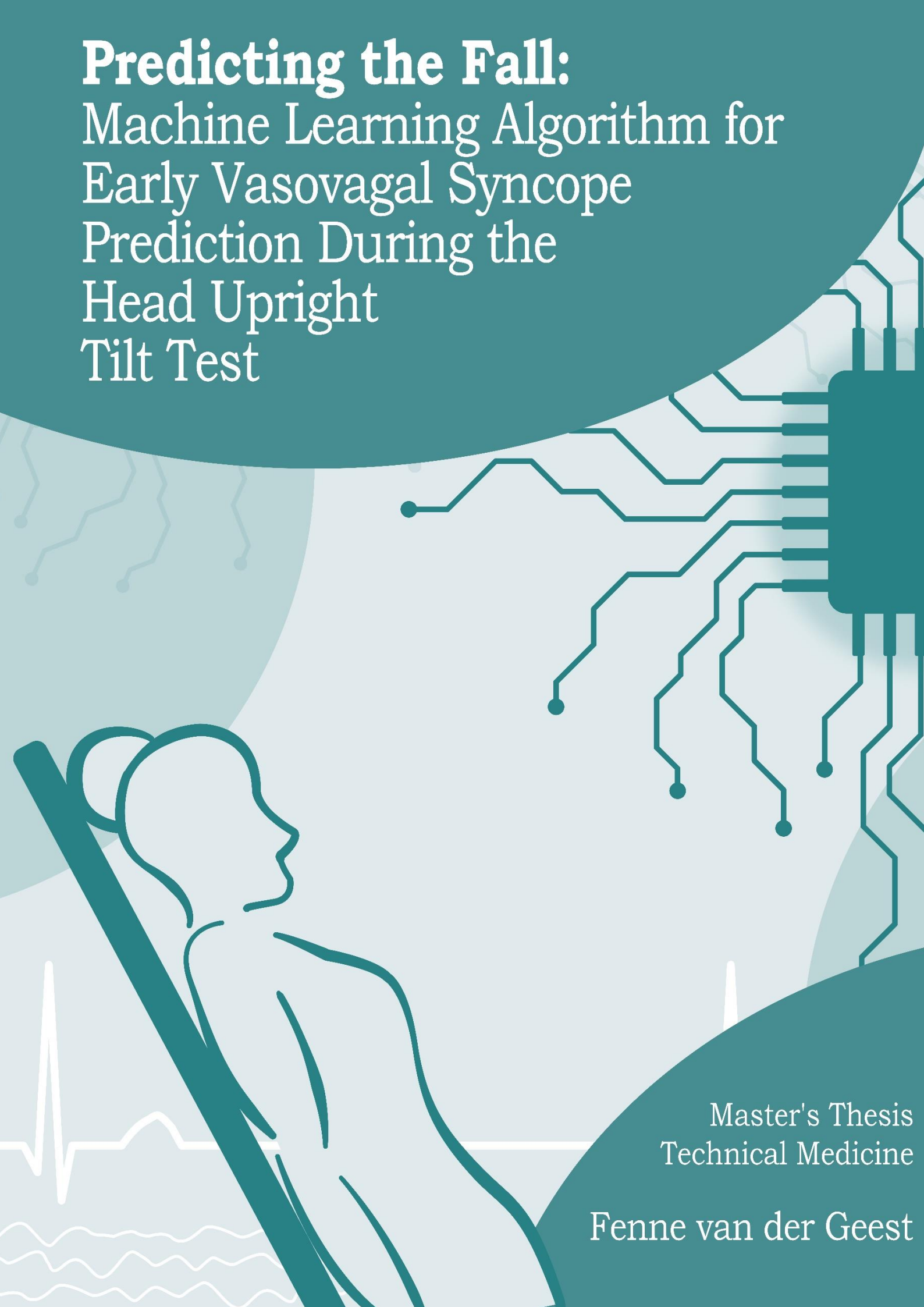


# Predicting the Fall: Machine Learning Algorithm for Early Vasovagal Syncope Prediction During the Head Upright Tilt Test

The background features a stylized illustration of a person sitting in a chair, tilted back, representing a head upright tilt test. The person is rendered in a dark teal outline. To the right, a complex circuit board pattern in a lighter teal color extends across the upper and middle sections of the image. In the lower-left corner, a white heart rate monitor (ECG) line is visible, with several peaks and troughs. The overall color palette consists of various shades of teal and light blue.

Master's Thesis  
Technical Medicine

Fenne van der Geest

*- This page was intentionally left blank -*

# **PREDICTING THE FALL: MACHINE LEARNING ALGORITHM FOR EARLY VASOVAGAL SYNCOPE PREDICTION DURING THE HEAD UPRIGHT TILT TEST**

Fenne van der Geest

Student number : 4654382

October 8, 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

*Technical Medicine*

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

**Master thesis project (TM30004 ; 35 ECTS)**

Department of Clinical Neurophysiology, LUMC

*February 2024 – October 2024*

## **Supervisors**

Dr. I.A. (Ineke) van Rossum

Dr. F. (Furong) Ye

Dr. M.R. (Martijn) Tannemaat

Dr. A.V. (Anna) Kononova

## **Thesis committee members**

Dr. I.A. (Ineke) van Rossum                      *LUMC*                      (*chair*)

Dr. F. (Furong) Ye                                      *LIACS*

Dr. Ir. R.M. (Roos) Oosting                      *TU Delft*

An electronic version of this thesis is available at <http://repository.tudelft.nl/>

*- This page was intentionally left blank -*

# Preface

---

This thesis marks the end of my student journey that started with my Bachelor's degree in Clinical Technology and continued with my Master's degree in Technical Medicine at Delft University of Technology, Leiden University Medical Center (LUMC) and Erasmus Medical Center (EMC). I chose this path because I wanted to be part of the latest innovations in healthcare and at the same time find a way to contribute my own creativity within this field.

Looking back over the past few years, I'm really grateful for all the opportunities I've had to grow, both personally and professionally. The combination of clinical and technical education has given me a unique perspective and I've gained a greater appreciation of how technology can make a real difference to health care. The various internships I have completed have been particularly educational, giving me the opportunity to gain more practical knowledge and a broader understanding of the different ways in which technology can be used in healthcare.

As I progressed through my studies, I became increasingly fascinated by the human brain and how much there is still to be discovered about it. At the same time, I started to learn about the possibilities of artificial intelligence in healthcare. This eventually led to my thesis in the Department of Clinical Neurophysiology at LUMC, which was the perfect opportunity to combine my interest in the brain and AI. During my internship, I was able to dive deeper into these topics and expand my knowledge.

There are a few people I would like to thank for their support during this project. Firstly, my medical supervisor, Ineke, your enthusiasm for the subject kept me inspired and motivated. You were always there to answer my questions and I learned so much from our discussions. I'm also very grateful to my technical supervisor, Furong, who guided me through the technical aspects of my thesis and helped lay the foundation for my analysis. Also a special thanks to Martijn for helping me combine the medical and technical sides of the project and the critical questions, and to Anna for the additional technical support.

I would also like to thank the team at the Department of Clinical Neurophysiology at LUMC who supported my clinical development during my internship, and of course my fellow interns and researchers who made coffee breaks and lunches something to look forward to every day.

Finally, to my friends, housemates, parents and sisters, thank you for your endless support and encouragement over the years. You have always been there to lift me up and I couldn't have done it without you.

To everyone who has been a part of this journey, thank you! I am incredibly proud of how far I have come and how much I have grown over the past years. When I think back to my seven year old self, she would hardly be able to believe where I am today. I am excited and curious about what the future holds and I look forward to applying the knowledge I have gained in the coming years.

*Fenne van der Geest  
October 2024*

# Table of Contents

---

<b>PREFACE</b>	<b>5</b>
<b>LIST OF ABBREVIATIONS</b>	<b>7</b>
<b>SUMMARY</b>	<b>8</b>
Background	8
Methods	8
Results	8
Conclusion	8
<b>1. BACKGROUND</b>	<b>9</b>
1.1. <i>Vasovagal Syncope</i>	9
1.1.1. The Head-Upright Tilt Test	9
1.1.2. Reflex Response	10
1.1.3. Complexity of VVS	11
1.2. <i>Machine learning</i>	12
1.2.1. Supervised Learning	12
1.2.2. Machine Learning Pipeline	13
1.3. <i>Early Prediction</i>	15
1.4. <i>Thesis Objective</i>	15
<b>2. METHODS</b>	<b>16</b>
2.1. Data Selection	16
2.2. ML Pipeline	16
2.3. Feature Analysis	18
<b>3. RESULTS</b>	<b>19</b>
3.1. Data Selection	19
3.2. Model Performance	19
3.3. Feature Analysis	20
<b>4. DISCUSSION</b>	<b>28</b>
4.1. Model Performance	28
4.2. Feature Analysis	29
4.3. Limitations	30
4.4. Future Research	31
<b>5. CONCLUSION</b>	<b>32</b>
<b>6. REFERENCES</b>	<b>33</b>
<b>7. APPENDIX</b>	<b>37</b>
A. PRELIMINARY STUDY	37
B. VARIABLES	48
C. PYTHON PACKAGES AND VERSIONS	49
D. MACHINE LEARNING PIPELINE OVERVIEW	50
E. HYPERPARAMETER SEARCH SPACE	51
F. PERFORMANCE FORMULA'S	52
G. ROC-CURVES	53
H. SELECTED HYPERPARAMETERS	54
I. SHAP SUMMARY BAR PLOTS	55
J. SELECTED FEATURES IN RANDOM FOREST	57
K. OUTLIER AND ARTIFACT DETECTION	58

# List of Abbreviations

---

<b>AI</b>	Artificial Intelligence
<b>AUC</b>	Area Under the Curve
<b>BMI</b>	Body Mass Index
<b>BO</b>	Bayesian Optimization
<b>BP</b>	Blood Pressure
<b>CO</b>	Cardiac Output
<b>CV</b>	Cross-Validation
<b>CWT</b>	Continuous Wavelet Transform
<b>DT</b>	Decision Tree
<b>ECG</b>	Electrocardiogram
<b>EEG</b>	Electroencephalogram
<b>EMG</b>	Electromyography
<b>FFT</b>	Fast Fourier Transform
<b>HO</b>	Hyperparameter Optimization
<b>HR</b>	Heart Rate
<b>HUTT</b>	Head-Upright Tilt Test
<b>MAP</b>	Mean Arterial Pressure
<b>MIP-EGO</b>	Mixed-Integer Parallel Efficient Global Optimization
<b>ML</b>	Machine Learning
<b>RF</b>	Random Forest
<b>ROC</b>	Receiver Operating Characteristic
<b>SFS</b>	Sequential Forward Selection
<b>SHAP</b>	SHapley Additive exPlanations
<b>std</b>	Standard Deviation
<b>SV</b>	Stroke Volume
<b>SVI</b>	Stroke Volume Index
<b>SVM</b>	Support Vector Machine
<b>TPR</b>	Total Peripheral Resistance
<b>tsfresh</b>	Time Series Feature Extraction based on Scalable Hypothesis tests
<b>VVS</b>	Vasovagal Syncope
<b>XGBoost</b>	Extreme Gradient Boosting

# Summary

---

## Background

Vasovagal syncope (VVS) is the most common form of syncope, accounting for more than half of all syncope cases. At least 35% of people between the ages of 35 and 60 have experienced VVS at least once in their lives. The Head-Upright Tilt Test (HUTT) is commonly used to elicit VVS while monitoring clinical signs and changes in heart rate (HR) and blood pressure (BP). The current protocol is both uncomfortable and time-consuming with suboptimal diagnostic yield. This study aims to develop and evaluate a machine learning (ML) pipeline capable of providing an early prediction of whether a patient with suspected VVS will experience syncope during HUTT, while also identifying features that contribute to a better pathophysiological understanding of VVS.

## Methods

The study included 434 adult patients with suspected VVS who underwent HUTT. A ML pipeline was developed for the early prediction of VVS, classifying patients into syncope and no syncope groups. Based on the results of the preliminary study, raw continuous BP, HR and Electroencephalogram (EEG) data were separated into 3-minute epochs before and after the tilt. Linear interpolation was selected as the most effective method to fill in missing data points. The Python package 'tsfresh' was used for feature extraction and the Boruta algorithm for feature selection. Three classification models were tested: Random Forest (RF), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). Model performance was evaluated using 5-fold cross-validation with AUC as the primary metric. In addition, a SHapley Additive exPlanations (SHAP) analysis was performed to assess the importance of features, which provided insight into the features that contributed most to the model's predictions. These features were further analyzed to determine the significant difference between both the syncope and no syncope groups.

## Results

The best performing model was the RF model with 61% AUC, 70% sensitivity, and 45% specificity. All selected features contributed to the classification of the syncope and no syncope groups across all folds. Five features were selected repeatedly during cross-validation, including three stroke volume index (SVI) features and two HR features. The two most frequently selected features were the *after SVI minimum* and the *before HR partial autocorrelation* features. In particular, the *after SVI minimum* feature was identified as a valuable feature in every fold. Both features showed a significant difference, suggesting that higher values for these features were associated with a greater likelihood of not experiencing syncope during HUTT.

## Conclusion

This study introduced a novel approach to the early prediction of VVS during HUTT. By developing an automated ML pipeline, we demonstrated that with hemodynamic features from the 3 minutes before and after the tilt can provide predictive insight into the occurrence of syncope 20 to 30 minutes later, although with limited sensitivity and specificity. Future research should focus on methods for handling artifacts and outliers and improving model robustness and accuracy. Ultimately, early detection of syncope has the potential to make HUTT more efficient, reduce patient discomfort, and avoid unnecessary testing.



# 1. Background

---

## 1.1. Vasovagal Syncope

Vasovagal syncope (VVS) is the most common form of syncope, accounting for more than half of all syncope cases. At least 35% of people between the ages of 35 and 60 have experienced VVS at least once in their lives.<sup>1</sup> Syncope is defined as a sudden loss of consciousness resulting from cerebral hypoperfusion due to low systemic blood pressure. It is characterized by rapid onset, short duration, and spontaneous full recovery.<sup>2,3</sup> People can also experience pre-syncope, which refers to the sensation of almost fainting without progressing to full loss of consciousness.<sup>4</sup> VVS can be triggered by factors such as fear, pain, or prolonged standing and is often preceded by symptoms like sweating, pallor, and nausea.<sup>5,6</sup> Although VVS is not considered life-threatening, it can have a negative impact on quality of life by increasing anxiety and fear of physical activity.<sup>6,7</sup> Additionally, those who experience syncope are at a higher risk for injuries, including bruises, lacerations, and fractures, affecting approximately 33% of VVS cases.<sup>8</sup> This highlights the importance of diagnosing recurrent VVS. The Head-Upright Tilt Test (HUTT) is an effective technique for providing diagnostic evidence of VVS.<sup>4,9</sup>

### 1.1.1. The Head-Upright Tilt Test

HUTT is a diagnostic test during which a patient is tilted while continuous measurements, including blood pressure (BP) and heart rate (HR), are taken. 'The Italian protocol' is a commonly used HUTT protocol for the diagnosis of reflex syncope, which includes VVS.<sup>10</sup>

#### *The Italian Protocol*

'The Italian Protocol' consists of different phases. It starts with a 5-10 minute stabilization phase in the supine position, followed by a passive phase of 20 minutes at a tilt angle of 60 degrees, and a provocative phase of 15-20 minutes after sublingual administration of nitroglycerin (NTG). The patient is brought back to supine position after the provocative phase, or with (pre-)syncope.<sup>9,10</sup> For a schematic overview of the HUTT protocol, see *Figure 1*.

The full protocol takes approximately 45 minutes to complete, but this can be reduced to 25 minutes if the "fast Italian protocol" is used.<sup>11</sup> However, this time does not include the time required to apply and remove the equipment from the patient. Using the 'The Italian Protocol', sensitivity for reflex syncope is approximately 65%, indicating that around one third of VVS patients will be tilted for 30-40 minutes and have a false negative result.<sup>12</sup> When no (pre-)syncope occurs during the test, considerable time is lost without gaining diagnostic clarity. However, if (pre-)syncope does occur, it can be an unpleasant experience for the patient, as symptom recognition is a part of the diagnosis. Furthermore, the test may be stopped prematurely if, for example, the patient feels unwell. Consequently, the current protocol is both uncomfortable and time-consuming with suboptimal diagnostic yield.

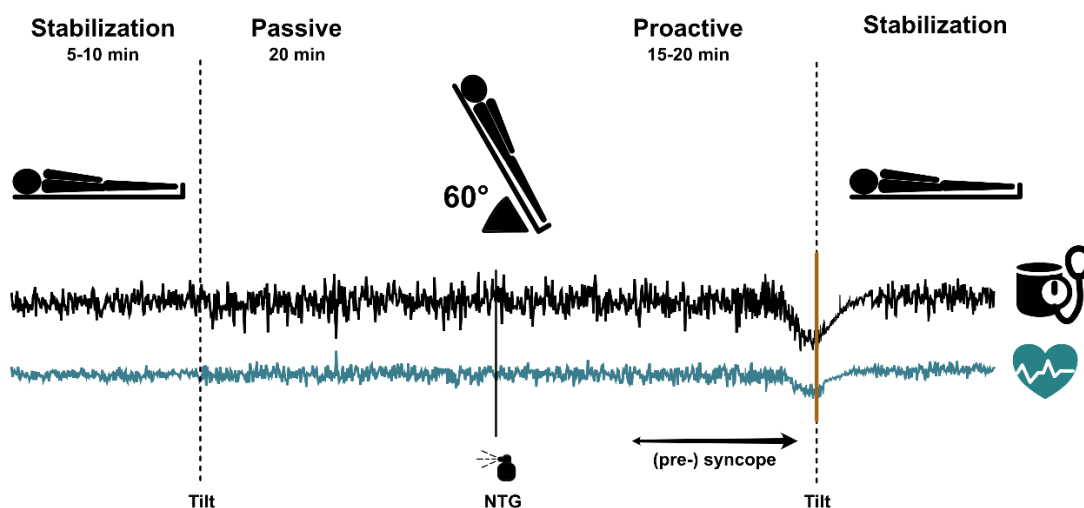


Figure 1. An overview of the full HUTT protocol in accordance with 'The Italian Protocol', consisting of the different phases and the administration of nitroglycerin (NTG).

### 1.1.2. Reflex Response

#### Normal Baroreflex

The baroreflex is continuously active in response to changes in BP. Under normal conditions when standing up, blood pools in the lower body, resulting in lower cardiac output (CO) and arterial BP, causing hypotension. Baroreceptors in the carotid sinus and aorta detect this decrease and send signals to the central nervous system. When triggered by the hypotension, the baroreflex decreases parasympathetic activity, thereby increasing HR. This in turn increases CO, which forces more blood into the arteries. In addition, the sympathetic response increases the HR and sends impulses to the arterioles, causing vasoconstriction and increasing total TPR. These combined effects drive more blood into the arteries and restrict outflow, increasing arterial BP. This process is displayed in Figure 2.<sup>13</sup>

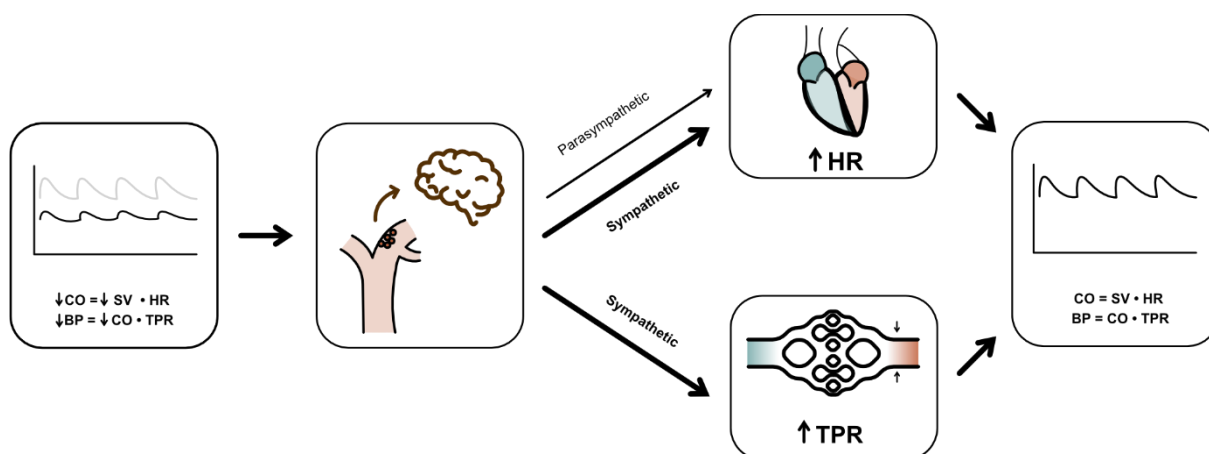


Figure 2. The normal baroreflex response to standing up. Blood pools in the lower body resulting in a decrease cardiac output (CO) caused by a decrease in stroke volume (SV). This hypotension is registered by the baroreceptors, which send a signal to the brain. This signal is then converted by the sympathetic and parasympathetic nervous systems into an increase in heart rate (HR) and total peripheral resistance (TPR) to increase blood pressure (BP) again.

### VVS Response

In VVS, if there is a decrease in BP, it is not sufficiently compensated by an increase in HR or TPR (Figure 3.) It is noteworthy that in VVS, just before the onset of syncope, there is actually a decrease in both HR and BP, as can be seen in the HUTT signal (Figure 4.) Hypothetically, there could be several reasons for this response. One possibility is that the reverse baroreflex response indicates that the normal response to the baroreceptors is lost or overridden by a stronger command. Another reason why a decrease in BP may not be met by an increase in HR or TPR is that it may be the result of an inappropriate reflex response to a trigger. It could also be an appropriate response to a problem, where the body's way of dealing with the problem is to cut off blood flow and brain activity, but it is not yet known what that problem might be.<sup>13</sup>

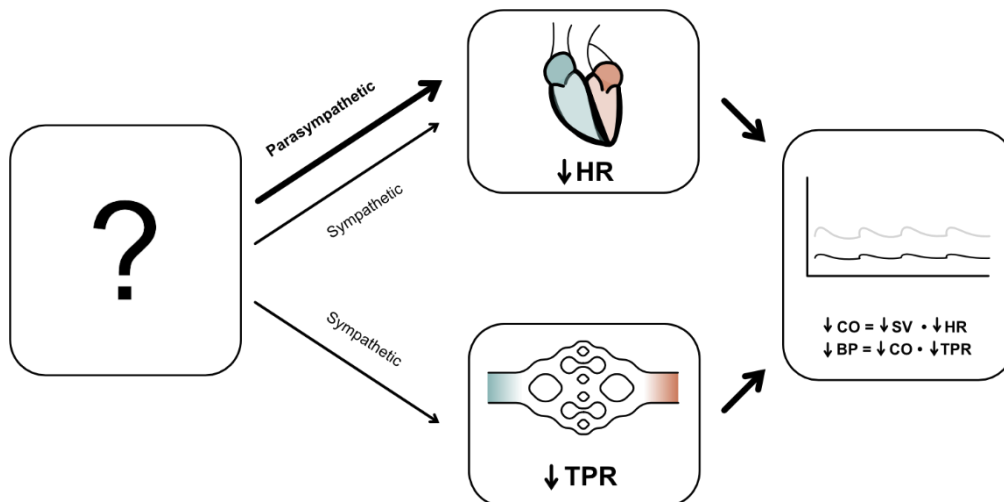


Figure 3. The VVS response is triggered by an unknown mechanism and causes a decrease in heart rate (HR) and total peripheral resistance (TPR), as well as a decrease in blood pressure (BP), which can lead to syncope.

### HUTT Measurements

In HUTT data of a patient with VVS, these BP and HR patterns can be seen as early as 9 minutes before syncope, see Figure 4. Initially, BP and stroke volume (SV) may start to decrease, while HR can increase slightly. Closer to syncope with 5 minutes before the event, a stronger drop in the BP can be seen. The HR and SV also follow with a strong drop around 30 seconds before VVS.<sup>14,15</sup>

### 1.1.3. Complexity of VVS

VVS presents a complex pathophysiologic pathway involving multiple interactions between the cardiovascular and autonomic nervous systems. For example, BP regulation by changes in HR and TPR. These interactions are further complicated by the interplay between the other physiological systems. For example, changes in respiratory rate can affect HR variability, which in turn affects BP regulation. In addition, individual variability plays a significant role in VVS. Variations in age and health status also contribute to different response patterns among individuals. Emotional and psychological factors, including stress and anxiety, can also trigger VVS episodes.

The examination of the HUTT signals are often limited to simple steady-state quantities including mean BP and HR, pulse pressure, and the HR variability during HUTT. This approach overlooks crucial details

within the dynamic interplay of these signals and significant temporal, structural, and spectral changes.<sup>16,17</sup> Exploring information within these diverse variables holds promise for early prediction and a deeper understanding of the mechanisms behind VVS.

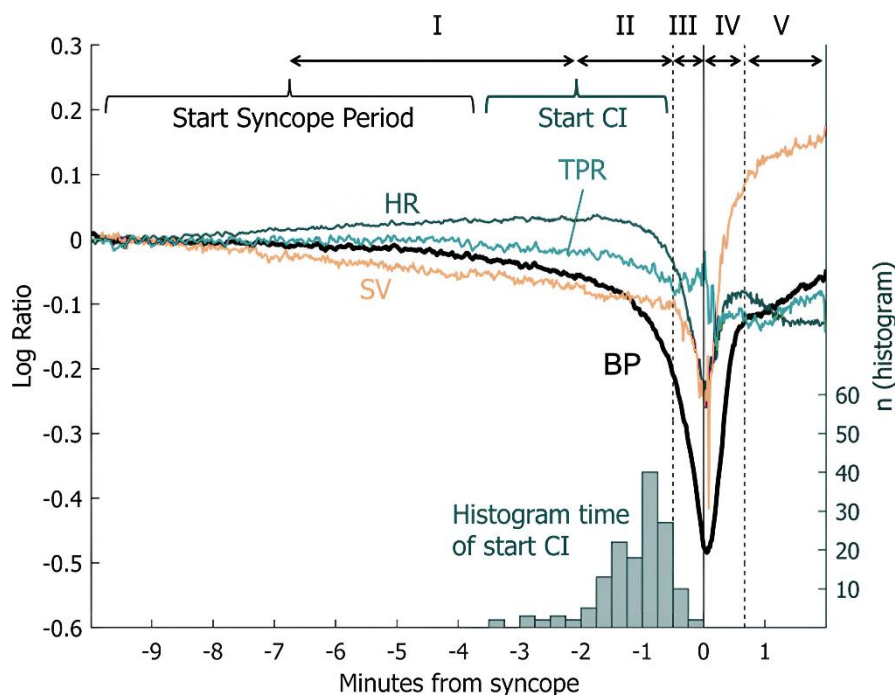


Figure 4. The Figure is adapted from Van Dijk et al. (2020b).<sup>14</sup> The hemodynamic parameters, such as blood pressure (BP), heart rate (HR), stroke volume (SV), and total peripheral resistance (TPR), are compared to a baseline taken shortly after the head-up tilt. The lines shown represent the average log-ratio values from vasovagal syncope (VVS) patients during HUTT.

## 1.2. Machine learning

In the era of digitization, the analysis of large amounts of complex data is becoming more common in healthcare.<sup>18,19</sup> Artificial intelligence (AI) is increasingly being utilized successfully in this process.<sup>19-21</sup> AI has the potential to offer deeper insights into the underlying mechanism and early prediction of VVS during HUTT.

### 1.2.1. Supervised Learning

Machine Learning (ML) is a field in AI that allows machines to learn from previous observations and experiences without human intervention. Different learning methods can be used, the four main methods are supervised, unsupervised, semi-supervised and reinforcement learning. In this study the supervised learning method was used. The main characteristic of supervised learning is that the dataset is labeled, meaning that each input comes with a corresponding known outcome. The goal of this method is to train an algorithm on labeled training data so that it can accurately predict outcomes for new, unseen data. This prediction can either be discrete or continuous. When the prediction is based on discrete information it is a classification algorithm, and when the prediction is continuous, it can also be referred to as a regression algorithm.<sup>22-24</sup> The ML pipeline consists of several steps that will be discussed below.

## 1.2.2. Machine Learning Pipeline

### *Pre-processing*

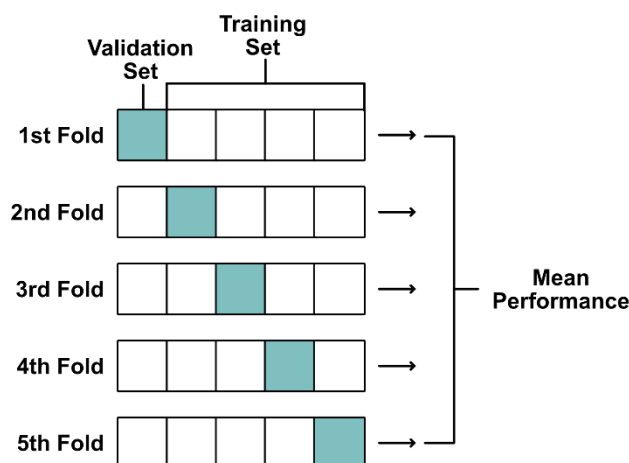
Cleaning the data consists of several parts, including selecting the right data and ensuring that the data is of good quality to enhance the performance of the algorithm.<sup>24</sup> It is important to ensure that the data is properly synchronized, if applicable, to determine if data transformations are needed, and to consider the impact of outliers on the model. It is also important to investigate what could have caused interference in the signal. In addition, it should be determined how to handle missing data and whether it should be filled in. All these factors and more will affect the quality of the data.<sup>22-24</sup>

### *Feature Extraction*

Feature extraction is the process of transforming raw data into numerical features that can be used in ML models. The goal of this process is to preserve the information from the original data while improving model performance compared to using the raw data alone. Which features should be extracted depends on the type of data, for example, images have different features than time series data. Feature extraction can be done manually or automatically. Manual feature extraction involves identifying and describing relevant features based on background knowledge. An example of a manual feature that can be extracted is the mean of a signal window.<sup>25,26</sup> Automatic feature extraction uses deep learning networks or algorithms, such as the Time Series Feature Extraction based on Scalable Hypothesis tests (tsfresh) python package that was applied in this project, automatically calculating a wide variety of features from time-series data.<sup>25,27</sup>

### *Data Splitting*

To ensure that the model is generalizable, it is important to train the model on one part of the dataset and test the model's performance on an unseen part of the dataset, also known as test data. This approach helps to detect overfitting, a situation where the model learns overly specific or complex patterns in the training data, which can lead to poor performance when applied to the test data. A commonly used technique to account for overfitting is cross-validation (CV). For example, in k-fold CV, the data is divided into k equal parts, and in each fold, a different part of the data is used to train and test the model (*Figure 5.*) In leave-one-out CV, the ML model is trained on all but one data point, and each point is iteratively left out to evaluate the model's generalization performance. The distribution of the data set between the training and test sets can also be done in different ways, for example, randomly or by taking into account the statistical distributions within the data set. Which data splitting method is best depends on the data set.<sup>24,28</sup>



*Figure 5. A schematic illustration of cross-validation, where a different part of the data is used as the training set and a different part as the validation set for each fold.*

### *Feature Selection*

Feature selection is a useful technique for handling high-dimensional data that would otherwise compromise the performance of ML algorithms. This process involves selecting a subset of relevant features to reduce dimensionality while conserving the essential information. There are different feature selection methods. For example, wrapper methods that evaluate the performance of different subsets of features, either by starting with an empty subset and adding features that reduce error (forward selection), or by starting with all features and removing those that reduce error (backward selection). In this study, the Boruta algorithm was used for feature selection, comparing the importance of original features against copies of the original features to determine their relevance.<sup>29</sup> Effective feature selection is key to maintaining the quality and accuracy of the learning process.

### *Classification Model*

There is a wide range of ML models that can be used to achieve optimal performance. Several supervised learning models have already been used in the literature to analyze HUTT data from syncope patients including Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) .

#### ***Support Vector Machine***

A SVM is a commonly used ML model on HUTT data from syncope patients<sup>17,30–33</sup>. SVM is a classifier that performs classification by creating a hyperplane in a higher dimensional space. SVM can be divided into linear and nonlinear models and works by mapping data into a feature space where each point in the feature space contains information about a data point. This feature space is created using a kernel function. The proper choice of a kernel can make inseparable classes separable by transforming data and creating the optimal decision boundary by maximizing the separation margin between classes.<sup>34</sup> The most commonly used kernels are: linear kernel, polynomial kernel, sigmoidal neural network kernel, and Gaussian radial basis kernel.<sup>33</sup>

#### ***Decision Tree and Random Forest***

DT and RF are also used in literature for the analysis of HUTT data.<sup>16,31</sup> A DT is a hierarchical classification technique that divides data into subsets, providing information to separate classes. It consists of internal nodes representing features, leaf nodes indicating outcomes, and branches dictating decision parameters. The computation of the decision parameters, including a feature and its split location, is determined by maximizing the information gain.<sup>34</sup> A RF uses a combination of DTs. The advantage of a RF is that a large number of DTs can be used, with each tree making a prediction and the algorithm selecting the best prediction through voting.<sup>16</sup>

#### ***Extreme Gradient Boosting***

Extreme Gradient Boosting (XGBoost) is a powerful gradient boosting framework that has become widely used in data analysis due to its efficiency and high performance. It builds on the principles of boosting to create a strong predictive model by adding DTs sequentially that correct errors made by previous trees. It also uses a regularized learning objective that helps to prevent overfitting and improves model generalization. This results in high prediction accuracy, which is specifically important in medical applications where decisions often depend on accurate predictions.<sup>35</sup>

### *Hyperparameter Optimization*

A hyperparameter is a variable that specifies details in the learning process of the ML model. Hyperparameter optimization (HO) is essential for building effective ML models because it refines the hyperparameters that shape the model's architecture that significantly affects its performance. However, manual hyperparameter tuning requires a deep understanding of the algorithm and can be time consuming due to the number and complexity of parameters. In contrast, automated HO

techniques, such as Grid Search and Bayesian optimization (BO), streamline this process. BO, for example, uses previous evaluation results to predict optimal hyperparameters, minimizing unnecessary trials and improving efficiency. This automation not only saves time, but it also enhances the model performance and ensures reproducibility.<sup>36</sup>

### *Performance Evaluation*

Evaluating a ML model involves identifying differences between predicted and actual outcomes to ensure accuracy and reliability. Model performance is evaluated using metrics such as accuracy, sensitivity, specificity, precision, F1 score, and Area under the Receiver Operating Characteristic Curve (AUC-ROC), which help measure the correctness of predictions by analyzing true positives, true negatives, false positives, and false negatives. These matrices are typically used in a simulation-based evaluation using the test data. However, to test the clinical relevance of the pipeline the performance of the model should also be tested when it is implemented, examine its generalizability to new data, user feedback, and whether medical experts trust the model.<sup>24</sup>

## 1.3. Early Prediction

As symptom recognition during tilt-induced VVS is crucial for diagnosis, the use of ML data cannot yet fully replace the current HUTT protocols for VVS. However, further exploration of the use of ML for early prediction of syncope during HUTT may aid in improving syncope care in four ways. Firstly, in about a third of patients with a clinical diagnosis of VVS, HUTT does not lead to syncope, even after the administration of NTG. Identifying those subjects during the first minutes of HUTT could save time for both patients and healthcare professionals, while also reducing the overall cost of the procedure.<sup>32,37</sup> Secondly, the use of ML for classification of syncope during HUTT may improve the understanding of hemodynamic changes prior to the actual syncope and thereby help to further unravel the pathophysiology of VVS. Thirdly, prediction of VVS during the early stages of HUTT can increase diagnostic accuracy in patients whose test must be stopped prematurely (e.g., due to feeling unwell). Fourthly, it could represent a step towards performing HUTT without requiring the patient to experience syncope for diagnosis, thus avoiding the discomfort and potential risks associated with a syncope episode.

The use of ML for early prediction of VVS during HUTT has been the subject of several studies in the literature.<sup>17,32,33,38</sup> However, the current study has the added value of using the largest database to date, extending the comparison of different ML models, and working with a wider range of calculated features of both common HR and BP variables and electroencephalogram (EEG) variables, which will be examined for relevance and possible impact on the underlying pathophysiology of VVS.

## 1.4. Thesis Objective

The aim of this study is to develop and evaluate a ML pipeline capable of providing an early prediction of whether a patient with suspected VVS will experience syncope during HUTT, while also identifying features that contribute to a better pathophysiological understanding of VVS. The dataset identified in the preliminary study (*Appendix A*) will be applied to various classification algorithms, and the selected features will be further analyzed. The objectives are:

- 1) To develop an automated ML pipeline for the early prediction of VVS during HUTT.
- 2) To perform an in-depth analysis of the selected features.

## 2. Methods

---

### 2.1. Data Selection

#### 2.1.1. Patient Selection

HUTT data were gathered from the Syncope Unit of the Leiden University Medical Centre between January 2019 and November 2023. Inclusion criteria were adult patients, with a clinical history suspected of VVS who underwent a HUTT test with a modified Italian protocol.<sup>10</sup> Patients were excluded when a different protocol had been used, when another form of syncope had been diagnosed, or when there were technical problems during HUTT.

A modification of the Italian protocol was used, with 10 minutes of supine rest, 20 minutes of head-up tilt, administration of 0.4 mg nitroglycerin (NTG) sublingually, and 20 minutes of tilt. Patients were tilted back before the allotted time when (pre-)syncope, asystole, or EEG slowing occurred.<sup>14</sup> The patient was diagnosed with syncope at the time of loss of consciousness or in the presence of (pre-)syncope, accompanied by a corresponding VVS BP and HR pattern. Patients who experienced (pre-)syncope before or after NTG administration were included in the 'syncope' group, and patients who did not experience (pre-)syncope during HUTT were included in the 'no syncope' group.

#### 2.1.2. Data Acquisition

Data were obtained by recording continuous finger BP (Finapres Nova or BMEye Nexfin), using one electrocardiogram (ECG) channel, two EEG channels c3-o1 and c4-o2 according to the 10-20 system (Nihon Kohden Neurofax EEG-1200), and video recording. In addition, age and sex were also registered. The continuous BP and ECG signals were recorded at a rate of 1 Hz, and the EEG signals were recorded at a rate of 200 Hz. During the test, or afterward using video recordings, the times of upward tilt (approximately 12 seconds), backward tilt, and the onset of syncope were documented. The five minutes before and after the tilt had to be complete and accessible in the database, without the occurrence of syncope during this period of time. Patients were excluded if not all of the measurement files were available, if more than 100 consecutive data points had been lost, or if it was not possible to synchronize the different measurements.

Based on the results of the preliminary study (*Appendix A*), two 3-minute epochs, one before and one after the tilt, were used for the analysis. Both basic and extra variables were included, as described in *Appendix B*. A schematic overview of the selected data can be found in *Figure 6*.

### 2.2. ML Pipeline

This ML pipeline was based on a previous study for the classification of normal and abnormal electromyography (EMG) data, with the exception of some modifications that include pre-processing, the testing of different ML classifiers and the addition of feature analysis.<sup>39</sup> Python version 3.9.19 was used for developing this pipeline. An overview of the used packages can be found in *Appendix C*. The details of the pipeline are described in the following section and a schematic overview can be found in *Appendix D*.

#### 2.2.1. Pre-processing

The necessary pre-processing steps were examined in the preliminary study (*Appendix A*). Linear interpolation was used to fill in the missing data and the 20 seconds before and after the tilt were removed from the data to minimize tilt artifacts. The EEG data were filtered with a 50 Hz notch filter.



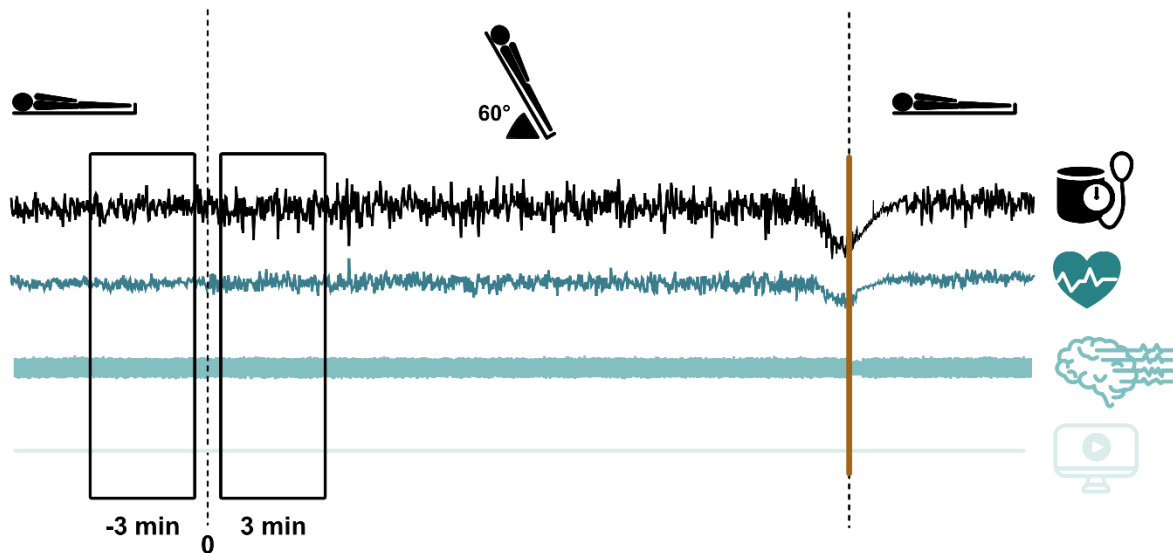


Figure 6. A schematic overview of the data selection, where two 3-minute epochs are created, one from before the tilt and one from after the tilt. Each epoch contains the time of the tilt, blood pressure, heart rate and electroencephalogram data.

### 2.2.2. Feature Extraction

To streamline the process of automatically extracting features, the tsfresh package was used.<sup>27</sup> The calculated features describe the time-series data and include statistical features, time and frequency domain features, distribution features, trend features, shape-based features, and more. This pre-defined package calculates over 750 different features per variable for each time-series.<sup>40</sup> In this study, tsfresh was used with its default settings.

### 2.2.3. Data Splitting

To prevent overfitting, a 5-fold CV was used. The data were divided into stratified folds, where the percentage of samples for each class in each fold was preserved. This was done using Sklearn's StratifiedKFold and Numpy's random to randomly select the patients per group in a stratified manner.<sup>41,42</sup>

### 2.2.4. Feature Selection

To select the most relevant features from the thousands of extracted features, the Boruta feature selection algorithm was used.<sup>29</sup> Boruta aims to find all relevant features by using a RF model with original features and shadow features. Shadow features are copies of the original features, but with randomly mixed values. This keeps their distribution the same, but cancels out their importance. The goal is to find all features that outperform the best shadow features, while rejecting those that underperform. The randomness helps to determine which features are really important and which are not. Boruta's default settings were combined with Sklearn's RandomForestClassifier with the number of trees set to 1000.<sup>41</sup>

### 2.2.5. Classification Model

Three different classifiers were used in this study: RF, SVM, and XGBoost. For RF and SVM, the RandomForestClassifier and SVC from the sklearn library were used, respectively.<sup>41</sup> For XGBoost, the XGBClassifier from the xgboost library was used.<sup>35</sup> Default parameters were applied for each classifier unless otherwise specified during HO.

### 2.2.6. Hyperparameter Optimization

In this study, the Mixed-Integer Parallel Efficient Global Optimization (MIP-EGO) algorithm was selected for HO.<sup>43</sup> MIP-EGO is an optimization framework that integrates mixed-integer programming with BO techniques to efficiently find the optimal solution. MIP-EGO is effective for both continuous and discrete variable problems. It builds a surrogate model that was tested in a 10-fold CV to approximate the objective function. It uses this model to select the most promising hyperparameters to try next. This approach saves time and computational resources by reducing the number of evaluations required. BO is used to optimize these expensive black-box functions. In each iteration, MIP-EGO proposes a new set of hyperparameters. These hyperparameters are then tested and evaluated based on the performance of the model on a validation dataset. The AUC was used as the optimization parameter over a maximum of 100 iterations. *Appendix E* shows the search space and the hyperparameters used for all three models.

### 2.2.7. Performance Evaluation

Each fold included training the model for 100 iterations. The highest AUC was used to determine overall model performance. Accuracy, precision, sensitivity, specificity, F1 and the ROC curve were also used as performance measurements. Formulas for these measurements can be found in *Appendix F*. The mean and standard deviation (std) of these values were calculated over the 5-folds.

## 2.3. Feature Analysis

SHapley Additive exPlanations (SHAP) was used to identify which features had the biggest impact on the model and to better understand why a model made a particular prediction.<sup>44</sup> We used the default settings of SHAP, and applied the trained model in combination with the test data. For each fold, a summary bar plot and a beeswarm plot based on the no syncope group were created.

In addition, the frequency of feature occurrence was calculated for the best performing model. This was done to determine if certain features were selected more often than others. The features that were selected more than once in the model were further examined. Welch's t-test was used to test feature significance across the entire dataset, with a p-value < 0.05 considered statistically significant. Furthermore, the two most frequently used features were plotted on the entire dataset to show their distribution.

### 3. Results

#### 3.1 Data Selection

From a total of 577 patients meeting the inclusion criteria, 434 patients were included in accordance with the data criteria (Figure 7). Out of these patients, there were 175 patients who did not experience syncope during HUTT and 259 patients who experienced (pre-)syncope during HUTT. The number of patients who experienced syncope before NTG administration was the smallest group among those of the syncope group. Proportionally, men were more likely than women to experience syncope before NTG administration (Table 1).

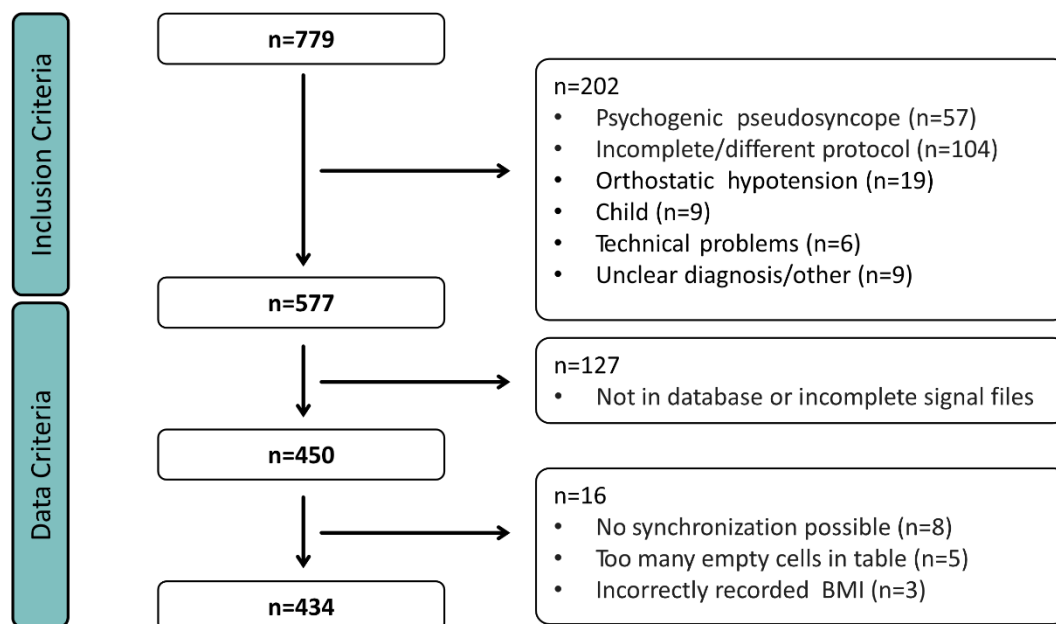


Figure 7. Flow diagram of the patient inclusion and exclusion.

Table 1. Baseline Characteristics.

Diagnosis	Mean BMI (std)	Mean age (std)	Group size (men)
<b>Syncope before NTG</b>	23.4 (4.1)	39,7 (22,5)	51 (19)
<b>After NTG</b>	24.9 (4.3)	47,7(21,9)	208(72)
<b>No syncope</b>	25.3 (4.7)	44,4 (18,7)	175 (76)

BMI: Body mass index, NTG: Nitroglycerin, std: standard deviation

#### 3.2. Model Performance

As shown in Table 2, the RF was the best performing model with a sensitivity of 70%, a specificity of 45%, and an average AUC of 61% with a std of 5% as depicted in Figure 8. The ROC curves for the XGBoost and SVM models are presented in Appendix G. The XGBoost model had an AUC of 54% and

the SVM model had an AUC of 45%. The hyperparameters for each classification model and fold can be found in *Appendix H*.

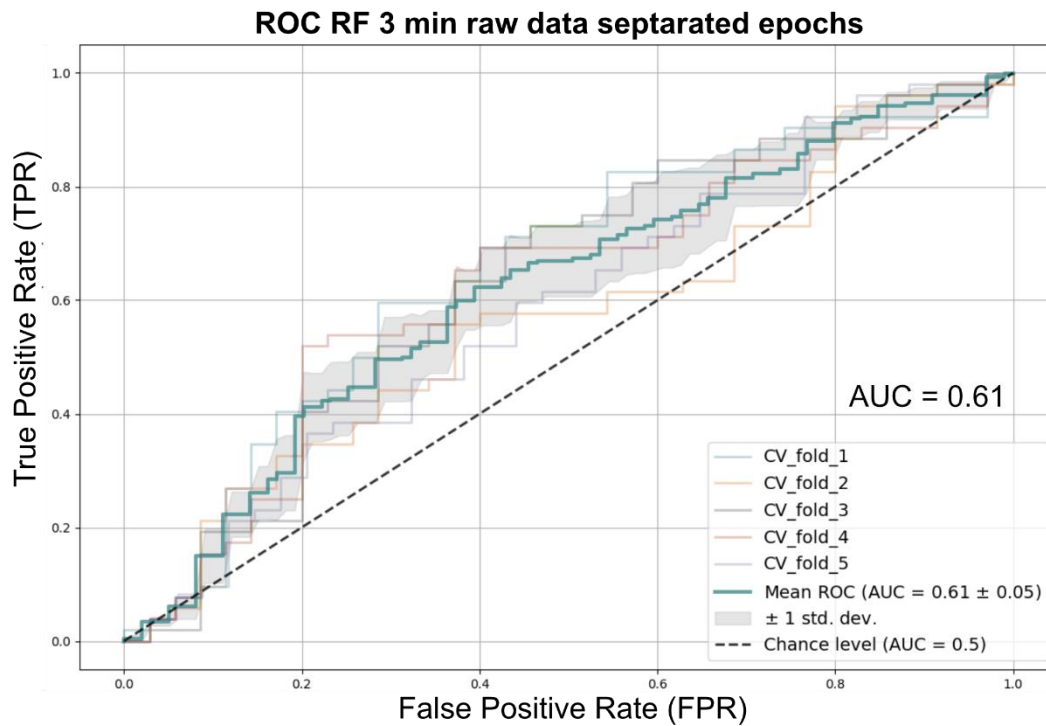


Figure 8. The ROC-curve of the final RF model with an AUC of 61% with a standard deviation of 5%.

Table 2. The performance measurements of the Random Forest, Extreme Gradient Boosting(XGBoost), and Support Vector Machine model.

	Random Forest	XGBoost	Support Vector Machine
<b>AUC</b>	61	54	45
<b>Accuracy</b>	60	57	49
<b>Precision</b>	58	54	49
<b>Sensitivity</b>	70	67	45
<b>Specificity</b>	45	41	53
<b>F1</b>	57	54	48

### 3.3. Feature Analysis

The RF model was used for further analysis of the features. SHAP summary bar plots and beeswarm plots for each fold are shown in *Appendix I* and *Figure 10*, respectively. The summary plots show that all features contribute to the classification of both the syncope and no syncope groups across each fold.

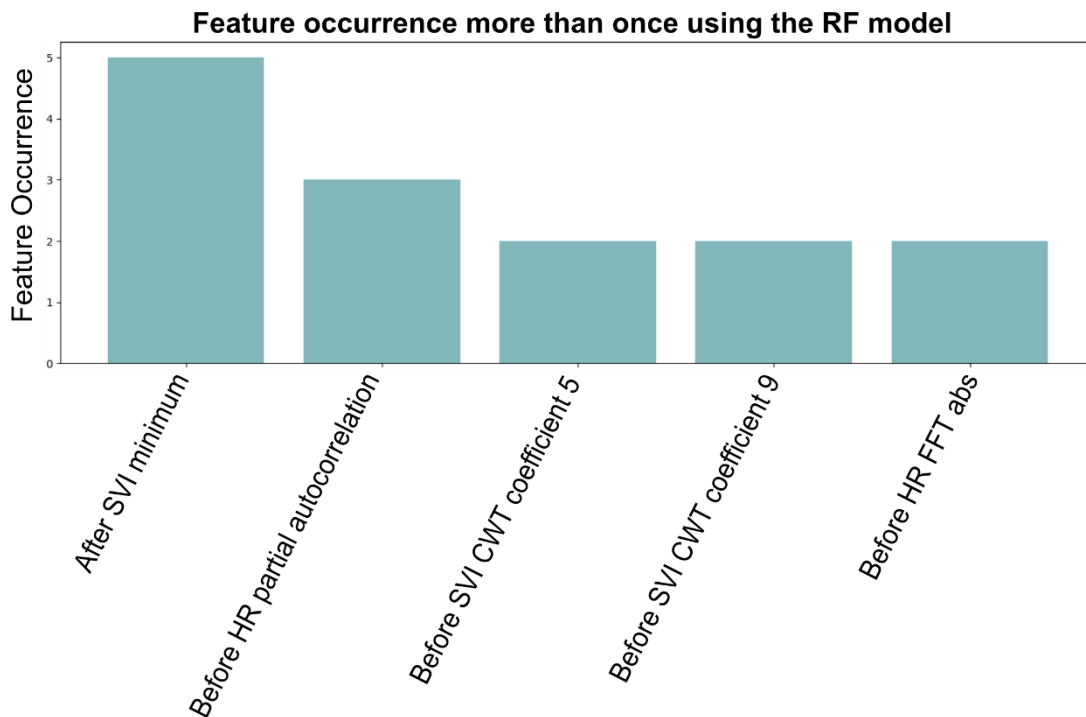


Figure 9. An overview of the features that were selected more than once in the 5-fold cross-validation.

An overview of all the selected features and their names can be found in *Appendix J*. Examining the number of features selected more than once across the five folds, we identified five features (*Figure 9*):

- *After SVI minimum,*
- *Before HR partial autocorrelation,*
- *Before SVI CWT coefficient 5,*
- *Before SVI CWT coefficient 9*
- *Before HR FFT abs*

The *after SVI minimum* and *before HR partial autocorrelation* features were the two most frequently selected features, with the *after SVI minimum* feature appearing in all five folds. In addition, *after SVI minimum* was the only feature that was consistently selected from the data after tilt. It is also noteworthy that the most frequently selected features were three SVI and two HR features. None of the EEG features or the age or sex were selected, with the majority of selected features coming from the HR and BP data, highlighting their importance in classification.

### 3.3.1 Explanation of Features

#### **After SVI minimum**

The minimum value of the stroke volume index (SVI) in the 3 minutes after the tilt

#### **Before HR partial autocorrelation**

Partial autocorrelation quantifies how much one value in the time series is related to a value at a previous time, leaving out the influence from the values in between. A *lag* of 4, shows how much the current HR depends on the HR from 4 time points ago, where a certain time in the timeseries  $t$

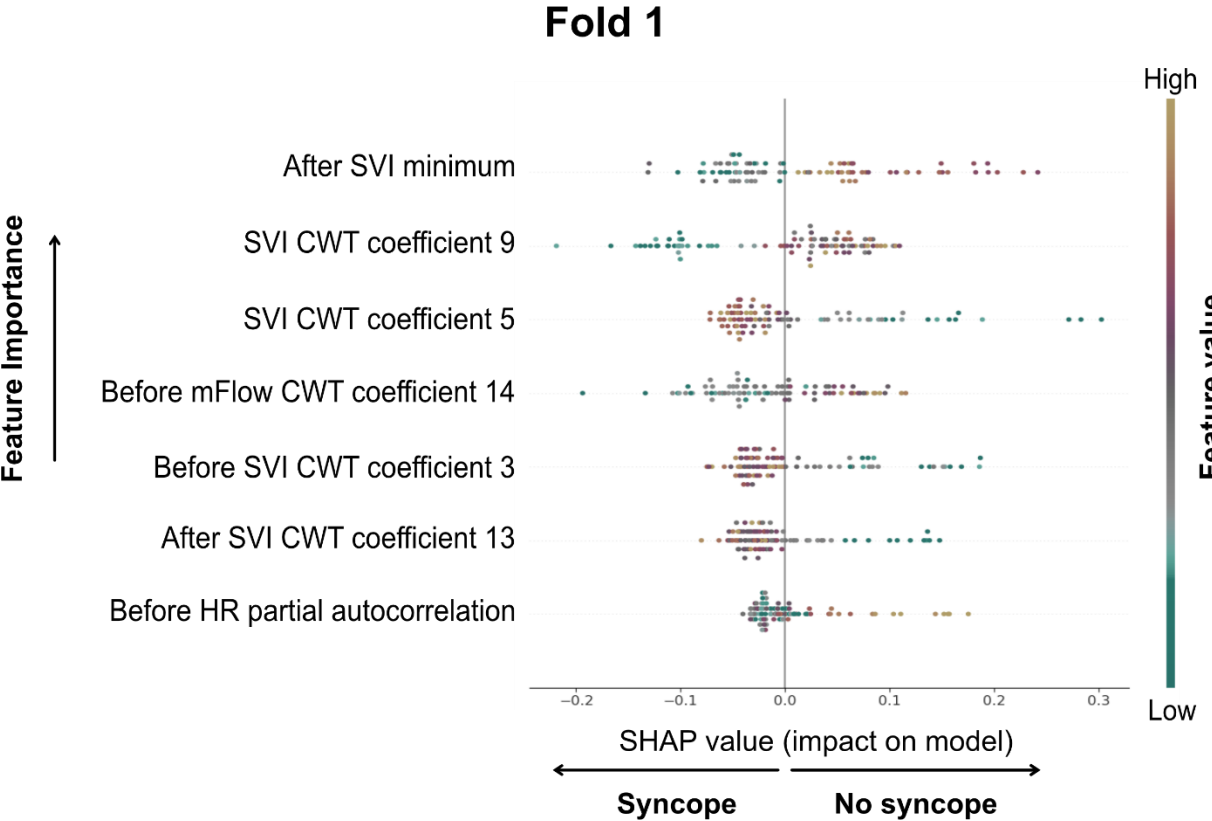
corresponds to  $t-4$  that can be summarized in a correlation. This correlation of the HR with a lag of 4 of the 3 minutes before the tilt was given in this feature.

**Before SVI CWT coefficients**

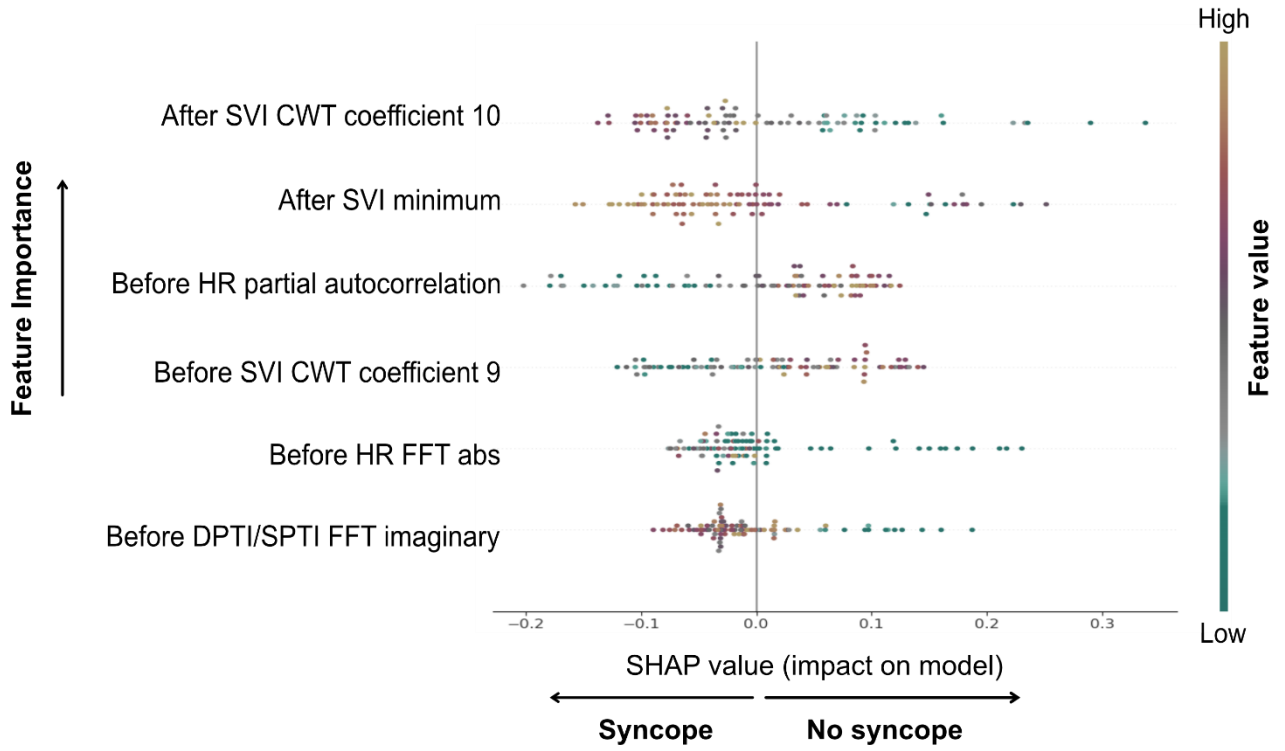
A Continuous wavelet transform (CWT) for the Ricker wavelet was used for the calculation of this feature.<sup>45</sup> The CWT allows for time-localized frequency analysis at different scales, where low frequencies are calculated over a longer period of time and the higher frequencies capture short-term variations. The CWT was used to calculate features from the SVI 3 minutes before the tilt with different coefficients.

**Before HR FFT abs**

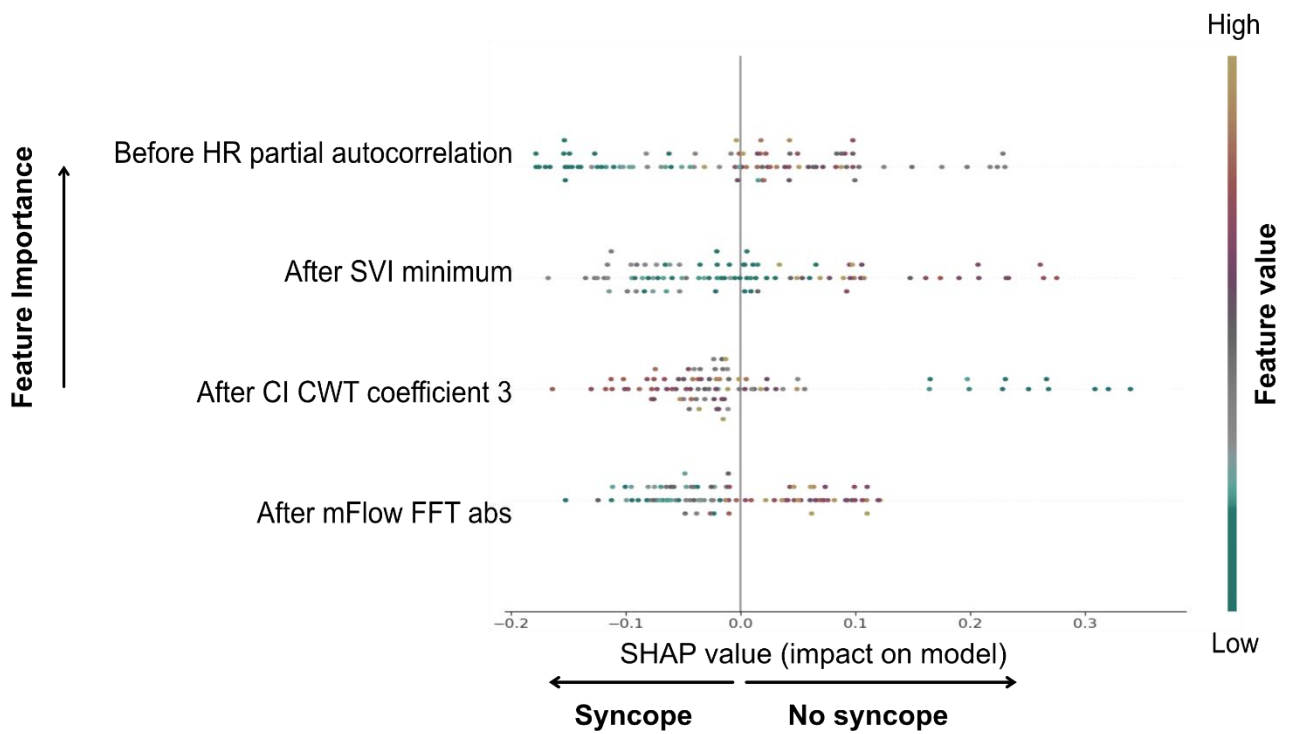
Fast Fourier Transform (FFT), is used to transform a time series into its frequency components. In this specific feature the absolute value (abs) was used to measure the magnitude of the frequency component, disregarding its phase or direction. This feature showed the strength or intensity of a certain frequency component in the HR signal 3 minutes before the tilt.



## Fold 2



## Fold 3



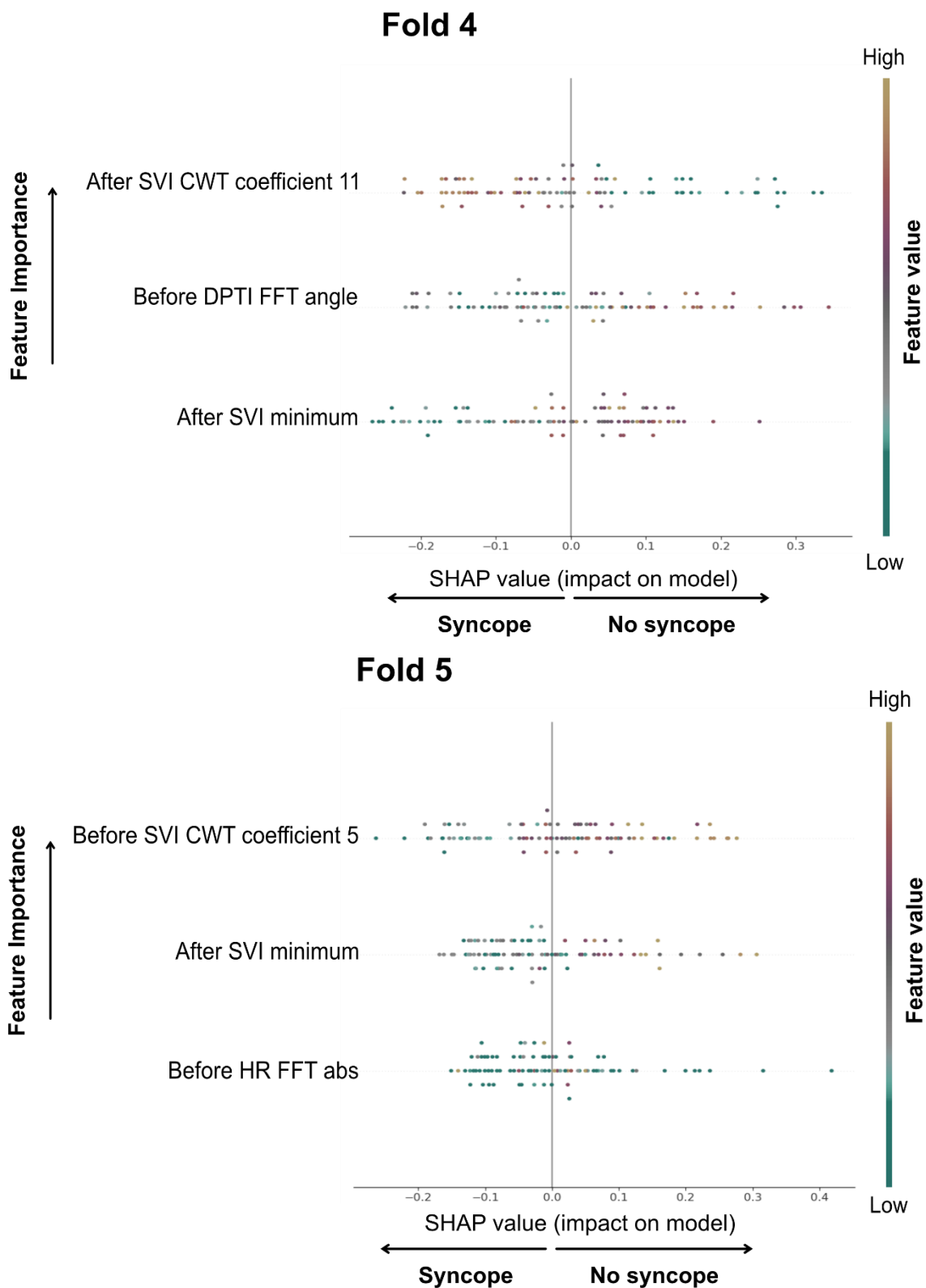


Figure 10. Beeswarm plots of the selected features for each of the five folds. These plots illustrate the feature importance, with the highest-ranked feature having the greatest influence on the model. The SHAP value indicates the probability of syncope or no syncope, with higher SHAP values corresponding to a higher probability of no syncope. The effect of a low or high feature value can be visualized with the probability of syncope or no syncope to determine the effect on the model. For example, in Fold 1, after SVI minimum is the most important feature, and higher values of this feature increase the likelihood of no syncope. Full feature names can be found in Appendix J.



### 3.3.2 Feature Interpretation

Looking at the SHAP values in *Figure 10* several things can be seen. In four out of five folds, a higher *after SVI minimum* resulted in a better chance of someone being placed in the no syncope group. Looking at the mean of both groups in *Table 3*, it can be seen that the mean of the no syncope group was also higher. However, in the second fold, where it was the second most important feature, it was the other way around.

For the feature *before HR partial autocorrelation*, three out of three folds showed that the higher the value, the more likely it was to be classified in the no syncope group. This corresponded to the mean in both groups.

The feature *before SVI CWT coefficient 5* was used twice, one time a lower value resulted in a greater chance of being placed in the no syncope group and the other time a lower value resulted in a greater chance of being placed in the syncope group. The mean of the no syncope group was higher than that of the syncope group.

The feature *before SVI CWT coefficient 9* was also used in two folds. Both showed that the higher the value, the greater the chance of no syncope. However, the mean was higher for the syncope group.

For the *before HR FFT abs* feature, a distinction in the effect of SHAP value was not clearly identifiable. However, some higher values seem to be more inclined towards the no syncope group. The mean value of the syncope group appeared to be higher than that of the no syncope group, but the differences were not significant.

### 3.3.3 Statistical Test

A Welch's t-test was performed to compare the means of the syncope and no syncope groups across all features. There was a significant difference between the means of these groups for all features ( $p < 0.005$ ), except for the *before HR FFT abs*, which did not show a statistically significant difference ( $p = 0.16$ ).

*Table 3. The mean and standard deviation (std) of the frequently selected features from both classes and the p-value from the t-test.*

Feature	No Syncope mean (std)	Syncope mean (std)	p-value
<b>After SVI minimum</b>	26.67 (10.02)	21.26 (9.78)	<0.0001
<b>Before HR partial autocorrelation</b>	-0.053 (0.17)	-0.005 (0.14)	0.0031
<b>Before SVI CWT coefficient 5</b>	6.85 (3.91)	5.30 (3.35)	<0.0001
<b>Before SVI CWT coefficient 9</b>	-0.92 (3.72)	0.17 (2.63)	0.001
<b>Before HR FFT abs</b>	25.58 (41.27)	31.50 (44.76)	0.16

### 3.3.4 Feature Distribution

The two most frequently used features are shown in the violin plots of *Figure 11*. We performed a Welch's t-test to compare the means of the syncope and no syncope groups on the *after SVI minimum* and *before HR partial autocorrelation* features. There was a significant difference between the means of these groups on *after SVI minimum* ( $p < 0.0001$ ) and *before HR partial autocorrelation* ( $p = 0.0031$ ).

Figure 12 showed that when the two features were plotted together in a scatterplot, it was possible to distinguish between the two groups to some extent. However, there was still considerable overlap between the two groups.

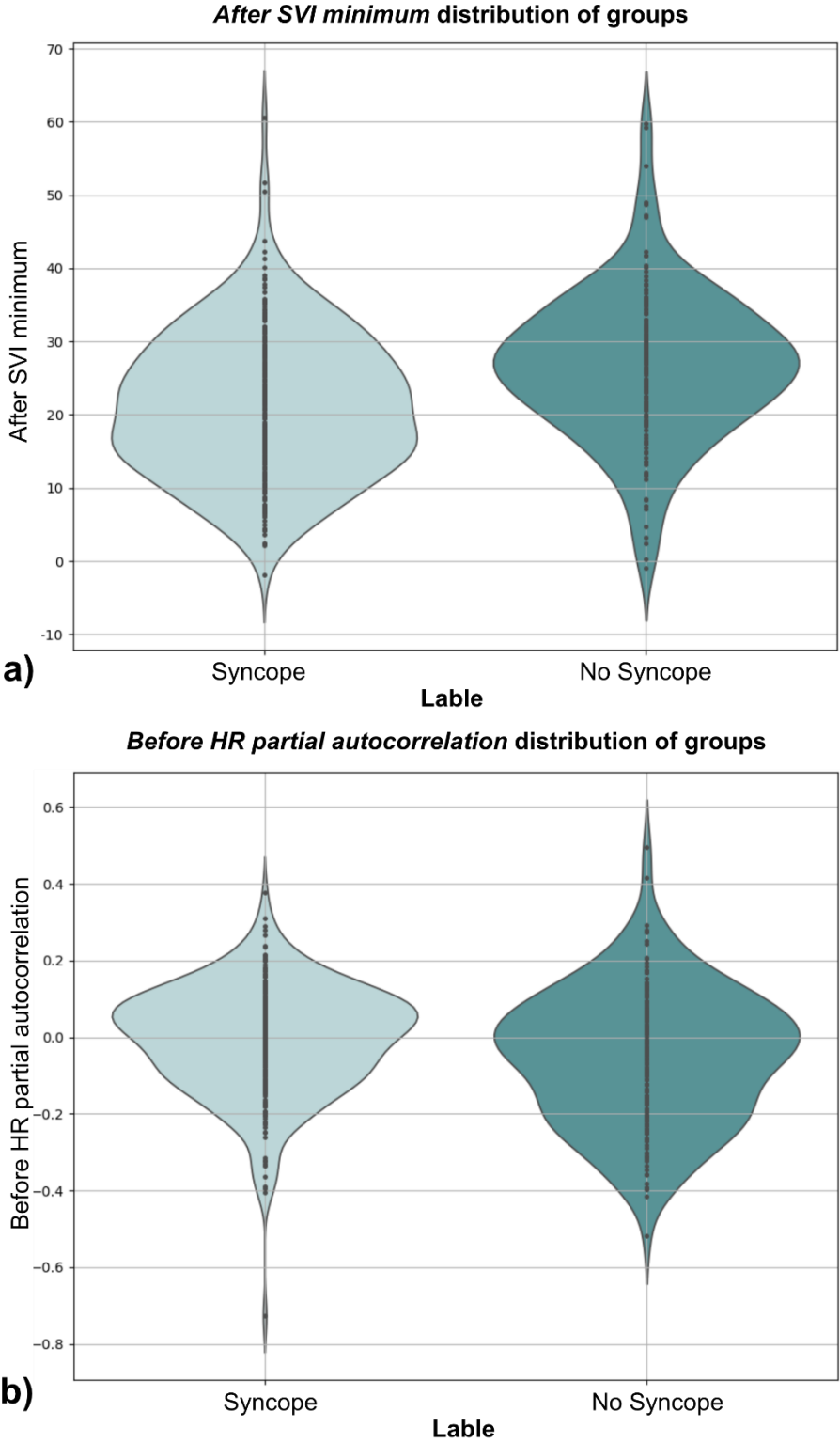


Figure 11. a) A violin plot of the after SVI minimum feature for the syncope and no syncope class. b) A violin plot of the before HR partial autocorrelation feature for the syncope and no syncope class.

### Feature scatter plot

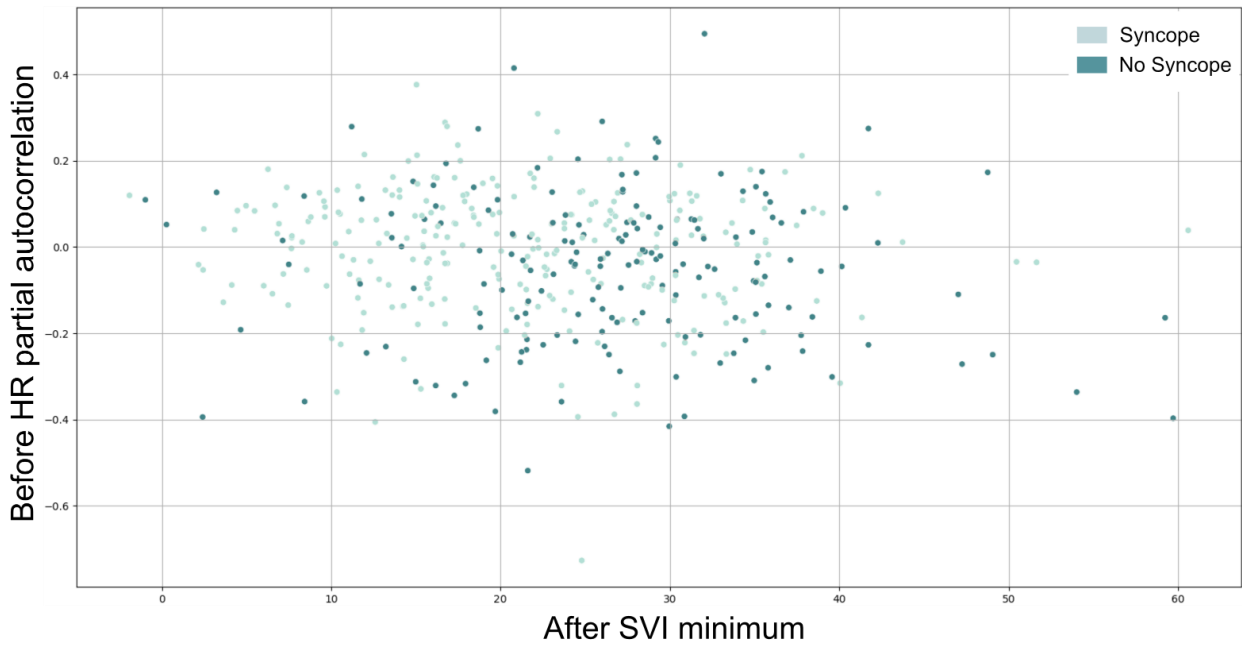


Figure 12. A 2D-scatter plot of the after SVI minimum and before HR partial autocorrelation features for the syncope and no syncope class.

## 4. Discussion

---

This study introduced a novel approach to the early prediction of VVS during HUTT. By developing an automated ML pipeline, we demonstrated that the 3 minutes before and after the tilt can provide predictive insight into the occurrence of syncope 20 to 30 minutes later, although with limited sensitivity and specificity. The best performing model in our analysis was the RF model, which obtained an AUC of 61%, sensitivity of 70%, and specificity of 45%. All selected features contributed to the classification of both the syncope and no syncope groups across all folds. Five features were selected repeatedly during CV, with the *after SVI minimum* and the *before HR partial autocorrelation* features being the most frequently used features. In particular, the *after SVI minimum feature* was identified as a valuable feature in each fold. Both features showed significant differences, indicating that higher values for these features were associated with a greater likelihood of not experiencing syncope during HUTT. However, there are several considerations that should be taken into account when interpreting these results.

### 4.1. Model Performance

The use of ML for early prediction and classification of syncope has been examined in several studies, with SVM often used as the classification algorithm.<sup>17,30–33</sup> However, this study found that RF outperformed SVM, achieving an AUC of 61% compared to 45% for SVM. This difference could be explained by the nature of the features that were selected, which seem to be less separable for SVM. This is evident in the distribution of the two most common features, which show significant overlap between the syncope and no syncope groups, making it more difficult to maximize the classification margin, even with a kernel. SVM training took longer than RF training, likely due to the complexity of optimizing decision boundaries with overlapping data. XGBoost also outperformed SVM but had an AUC of only 54%, which was lower than that of the RF. This result was unexpected, considering that both models were based on similar principles, with XGBoost using an additional boosting mechanism. This boosting process works by sequentially adjusting the model to minimize errors made on the training data. However, if the data set contains a significant amount of noise, this adjustment can have adverse effects. Overfitting can occur as the model becomes overly tuned to the training data, resulting in poorer generalization to unseen data. However, both models demonstrated comparable AUC on the training set, indicating that the more complex classifier is not necessarily the best option for this classification task. In addition, within the syncope group, no distinction was made between those who had syncope before and those who had syncope after receiving NTG. Although the subgroup who experienced syncope before NTG administration was relatively small, the underlying pathophysiology could be different from those who experienced syncope after NTG administration. This may complicate the classification of individuals into the syncope or no syncope group, especially given the differences in the distribution of patients within the syncope group. Further research could also explore alternative classification approaches, such as multi-class models, which may provide additional insight into the different physiological responses of syncope subgroups. The variation between patients and the influence of noise may have contributed to the limited overall AUC of 61%. This highlights the importance of addressing data quality and feature separability in future studies to improve predictive performance.

When the performance of the ML pipeline in this study is compared to the existing literature, this difference in performance can be seen. However, the better performance in other studies may have been influenced by variations in data selection, feature choices, or patient groups. For instance, the study by He et al. (2021) achieved a sensitivity of 86% and a specificity of 82%.<sup>38</sup> This article used SVM

in combination with both HR and BP features. An important difference here is in the selection of the data, they used 3-minute epochs from 3 minutes before the tilt to 15 minutes after the tilt. Similarly, the studies by Khodor et al. (2014, 2016) reported sensitivities of 88.5% and 87.5%, with specificities of 80.6% and 93.8%, respectively.<sup>17,32</sup> Again, these were based on SVM with HR and BP features and 15 minutes of data after the tilt. All three studies used data closer to the syncope event, which is less suitable for true early prediction. The most likely reason for this choice of data was also shown by Khodor et al. (2016), who compared the performance of the model using 5 minutes after the tilt versus 15 minutes after the tilt. The sensitivity increased from 69.6% to 87.5% and specificity increased from 66.9% to 93.8% when using 15 minutes versus 5 minutes.<sup>17</sup> Another factor contributing to the differences in performance between this study and the studies by Khodor et al. (2014, 2016) is the selection of subjects. They focused on healthy subjects, while this study focused on those already suspected of having VVS.<sup>17,32</sup> However, there may be physiological differences in baroreceptor responses between these two groups of subjects. For example, healthy people may be less likely to faint during HUTT, or they may respond differently from those suspected of having VVS. Given the current understanding of these underlying mechanisms, it is unclear whether such comparisons can be made, highlighting the need for further research into these physiological differences.

Couceiro et al. (2016) obtained an even better performance, reporting a sensitivity of 95.2% and a specificity of 95.4%.<sup>46</sup> However, they used a variable data epoch per patient, which was compared to a reference window at the beginning of the tilt. This means that the data epoch could have been drawn from the entire after tilt period, raising the question of whether true early prediction was consistently achieved. In addition, the study included only 43 subjects, which limits the reliability of the results.

There have also been studies that have used neural networks as a ML method. They have also reported higher performance results, albeit with smaller subject groups or different feature sets.<sup>37,47,48</sup> However, the data that were used came only from the supine position or within 5 minutes of tilting. While the current study focused on well-performing classical ML models, future research could further explore the application of neural networks to assess their potential advantages in this context.

## 4.2. Feature Analysis

Although the maximum AUC observed was only 61%, it is important to note that certain features were consistently selected across all folds. This suggests that the selected features do indeed contain relevant information that allows early distinction between patients who experience syncope during HUTT and those who do not. This indicates that pathophysiological changes between the two groups can be detected within the first few minutes of the test. For example, a lower minimum SVI after tilt appeared to be associated with a higher likelihood of syncope, while a lower or even negative partial autocorrelation of the HR with a lag of 4 may also indicate an increased risk of syncope. However, it is important to consider that for two of the five most common features, including the *after SVI minimum* feature, it was not always clear whether higher or lower values would predict a greater likelihood of syncope. In some cases, such as in one of the folds, a lower *after SVI minimum* value unexpectedly indicated a lower chance of syncope during HUTT. This could be due to the influence of another feature in that particular fold which may have altered the predictive value of the SVI minimum. This inconsistency could also be seen in the before SVI CWT coefficient 5 and before HR FFT abs features. In order to better understand this interaction, follow-up research should be performed to explore the effect of removing specific features, retraining the model and observing any changes in the results. This could provide more clarity on how different feature interactions affect predictions.

The finding that a higher minimum SVI indicates a lower likelihood of syncope is consistent with the existing literature, which shows a decrease in SV relative to BP approximately 9 minutes before the

onset of syncope.<sup>13</sup> While it is known that SV usually decreases within the first few minutes after tilt, the current study demonstrates that the extent of this early decrease in SVI may actually provide an early indication of whether an individual is likely to experience syncope during HUTT.<sup>6,49</sup> Additionally, three of the five most frequently selected features were related to SVI, emphasizing its potential predictive value. This study adds to prior knowledge and highlights the importance of SVI features, both before and after the tilt, in the early prediction of syncope.

The partial autocorrelation between the current HR and the HR from 4 seconds prior in the data before the tilt was also selected as useful feature. This correlation could potentially provide insight into the course and responsiveness of the HR, with a lower or even negative correlation during the stabilization phase indicating that a person is less likely to experience syncope during HUTT. This suggests that HR dynamics before the tilt could also provide valuable information about the likelihood of syncope later in the test.

The recurrence of HR and BP features in the current study was consistent with findings in the literature where HR and BP variables were commonly used for analysis. Specifically, time-frequency and frequency domain features are often used.<sup>17,30–32,38,48,50,51</sup> In this study, features based on similar principles were also selected several times, such as the *before SVI cwt coefficients* and the *before HR FFT abs*, further emphasizing the importance of these domains in syncope prediction during HUTT.

We found that EEG-based features did not add significant value to the classification. However, other types of measurements may still provide useful features. For example, Schang et al. (2006, 2007) looked at features derived from transthoracic impedance measurements, which assess impedance across the thorax.<sup>33,47</sup> In their study, principal component analysis was used to separate the two groups.<sup>47</sup> Nonetheless, considerable overlap between the groups was observed, suggesting that further research is needed to determine whether transthoracic impedance features could provide additional value in improving classification accuracy.

### 4.3. Limitations

Despite the new insights, several limitations in this study should be acknowledged. For instance, in the current study, raw data were used without any further outlier or artifact removal, except for excluding data 20 seconds before and after the tilting point. This decision was based on the preliminary results showing that the tested outlier and artifact removal method, as well as exponential smoothing, did not improve or even worsened model performance compared to using raw data. However, when evaluating the overall performance of the models, it appears that noise in the data may still have affected the results. Future research should focus on investigating the potential impact of outliers and artifacts on the accuracy of the model and explore alternative methods, such as ML based approaches or clustering techniques, for better noise reduction in the data.<sup>52,53</sup>

It became clear that there were several points in the data collection process where artifacts could be introduced. For example, the patient squeezing their hand can reduce blood flow, causing the finger BP measurement to no longer accurately reflect the overall BP. Another potential source of noise was that there would sometimes be a rise in BP when the doctor entered the room to start the tilt. These are just a few examples of external factors that can influence the measurements and potentially affect the accuracy of correctly classifying the patient. However, it is also important to recognize that some of these artifacts may contain valuable information. For example, an increase in BP when the physician arrives could indicate increased sensitivity to changes in HR, possibly indicating a lower likelihood of syncope during HUTT, or vice versa. Therefore, the decision to remove artifacts and outliers must be carefully considered as valuable predictive information may be lost.

Another point to consider is that the data selection and pre-processing methods were optimized based on the entire dataset, which may have introduced bias into the final model. This could result in overfitting to the current dataset and decreased robustness when applied to new data. Although CV was used to minimize this risk, it is important to further validate the model on new, unseen data, ideally from another institution, to more accurately assess its robustness and generalizability.

Finally, there was a slight percentage variation in the average performance of the model over the multiple runs of the 5-fold CV. This variation can be attributed to the randomization of the groups in each run and the separation into different sets, which complicates the comparison between models. Nevertheless, this highlights the need for further improvements in the robustness of the model, as consistent performance across multiple runs is critical for reliability.

#### 4.4. Future Research

This study has provided further insight into the pathophysiology of VVS, highlighting early differences that may exist between individuals during HUTT. It also served as a first step in exploring the potential for early detection of syncope. However, the clinical relevance needs to be considered.

In the future, it may be possible to avoid inducing (pre)syncope in patients in order to make a diagnosis, but we would need to keep in mind that a part of the diagnostic process involves recognizing the syncope symptoms, which would be lost with early detection. For now, the focus should be on improving the accuracy of the ML model to stop the test earlier for those who are unlikely to experience syncope during HUTT. However, it is still important to test this on larger datasets from multiple institutions. There is also a need to investigate how this can be implemented in practice and what adaptations would be required for use in clinical settings.

## 5. Conclusion

---

This study developed an automated ML pipeline using a novel approach for the early prediction of VVS during HUTT, which can provide predictive insight into the occurrence of syncope 20 to 30 minutes later. The RF model was the best performing model with an AUC of 61%, sensitivity of 70%, and specificity of 45%. Important features such as the *after SVI minimum* and the *before HR partial autocorrelation* features were frequently selected across the CV folds, suggesting their relevance in predicting syncope.

Despite the relatively low overall model performance, the results indicate that early pathophysiological changes, particularly in the SVI and HR dynamics, could provide valuable information for predicting syncope during HUTT. However, challenges remain, including the presence of noise in the data and potential feature interactions, which require further investigation. Follow-up studies are needed to validate the model on larger, more diverse data sets. In addition, future research should focus on improving the accuracy and robustness of the model by exploring alternative ML approaches while considering practical implementation in clinical settings. Ultimately, early detection of syncope has the potential to make HUTT more efficient, reduce patient discomfort, and avoid unnecessary testing.



## 6. References

---

1. Ganzeboom KS, Mairuhu G, Reitsma JB, et al. Lifetime Cumulative Incidence of Syncope in the General Population: A Study of 549 Dutch Subjects Aged 35–60 Years. *J Cardiovasc Electrophysiol* 2006;17(11):1172–1176; doi: 10.1111/J.1540-8167.2006.00595.X.
2. Brignole M, Moya A, De Lange FJ, et al. 2018 ESC Guidelines for the diagnosis and management of syncope. *Eur Heart J* 2018;39(21):1883–1948; doi: 10.1093/EURHEARTJ/EHY037.
3. Soteriades ES, Evans JC, Larson MG, et al. Incidence and prognosis of syncope. *N Engl J Med* 2002;347(12):878–885; doi: 10.1056/NEJM0A012407.
4. Aydin MA, Salukhe T V, Wilke I, et al. Management and therapy of vasovagal syncope: A review. *World J Cardiol* 2010;2(10):308; doi: 10.4330/WJC.V2.I10.308.
5. Brignole M, Moya A, De Lange FJ, et al. Practical Instructions for the 2018 ESC Guidelines for the diagnosis and management of syncope. *Eur Heart J* 2018;39(21):e43–e80; doi: 10.1093/EURHEARTJ/EHY071.
6. Jardine DL, Wieling W, Brignole M, et al. The pathophysiology of the vasovagal response. *Heart Rhythm* 2018;15(6):921–929; doi: 10.1016/J.HRTHM.2017.12.013.
7. Ng J, Sheldon RS, Ritchie D, et al. Reduced quality of life and greater psychological distress in vasovagal syncope patients compared to healthy individuals. *Pacing and Clinical Electrophysiology* 2019;42(2):180–188; doi: 10.1111/PACE.13559.
8. Jorge JG, Raj SR, Teixeira PS, et al. Likelihood of injury due to vasovagal syncope: a systematic review and meta-analysis. *EP Europace* 2021;23(7):1092–1099; doi: 10.1093/EUROPACE/EUAB041.
9. Benditt DG, Ferguson DW, Grubb BP, et al. Tilt table testing for assessing syncope. *J Am Coll Cardiol* 1996;28(1):263–275; doi: 10.1016/0735-1097(96)00236-7.
10. Bartoletti A, Alboni P, Ammirati F, et al. “The Italian Protocol”: a simplified head-up tilt testing potentiated with oral nitroglycerin to assess patients with unexplained syncope. *Europace* 2000;2(4):339–342; doi: 10.1053/EUPC.2000.0125.
11. Russo V, Parente E, Tomaino M, et al. Short-duration head-up tilt test potentiated with sublingual nitroglycerin in suspected vasovagal syncope: the fast Italian protocol. *Eur Heart J* 2023;44(27):2473–2479; doi: 10.1093/EURHEARTJ/EHAD322.
12. Forleo C, Guida P, Iacoviello M, et al. Head-up tilt testing for diagnosing vasovagal syncope: a meta-analysis. *Int J Cardiol* 2013;168(1):27–35; doi: 10.1016/J.IJCARD.2012.09.023.
13. van Dijk JG, van Rossum IA, Thijs RD. The pathophysiology of vasovagal syncope: Novel insights. *Auton Neurosci* 2021;236; doi: 10.1016/J.AUTNEU.2021.102899.
14. Van Dijk J, Ghariq M, Kerkhof FI, et al. Novel Methods for Quantification of Vasodepression and Cardioinhibition During Tilt-Induced Vasovagal Syncope. *Circ Res* 2020;127(5):E126–E138; doi: 10.1161/CIRCRESAHA.120.316662.

15. van Rossum IA, de Lange FJ, Benditt DG, et al. Variability of cardioinhibition in vasovagal syncope: differences between subgroups during cardioinhibition and beyond. *Clinical autonomic research* 2023;33(6):749–755; doi: 10.1007/S10286-023-00991-5.
16. Gilmore S, Hart J, Geddes J, et al. Classification of orthostatic intolerance through data analytics. *Med Biol Eng Comput* 2021;59(3):621–632; doi: 10.1007/S11517-021-02314-0.
17. Khodor N, Carrault G, Matelot D, et al. Early syncope detection during head up tilt test by analyzing interactions between cardio-vascular signals. *Digit Signal Process* 2016;49:86–94; doi: 10.1016/J.DSP.2015.11.005.
18. Wiens J, Shenoy ES, Investigator F, et al. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases* 2018;66(1):149; doi: 10.1093/CID/CIX731.
19. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med* 2022;28(1):31–38; doi: 10.1038/S41591-021-01614-0.
20. Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. *NPJ Digit Med* 2020;3(1); doi: 10.1038/S41746-019-0216-8.
21. Kather J, Pearson A, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019;25(7):1054–1056; doi: 10.1038/S41591-019-0462-Y.
22. Smiti A. When machine learning meets medical world: Current status and future challenges. *Comput Sci Rev* 2020;37:100280; doi: 10.1016/J.COSREV.2020.100280.
23. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 2019;19(1):1–18; doi: 10.1186/S12874-019-0681-4/TABLES/5.
24. Rahmani AM, Yousefpoor E, Yousefpoor MS, et al. Machine Learning (ML) in Medicine: Review, Applications, and Challenges. *Mathematics* 2021, Vol 9, Page 2970 2021;9(22):2970; doi: 10.3390/MATH9222970.
25. Anonymous. Feature Extraction Explained - MATLAB & Simulink. n.d. Available from: <https://nl.mathworks.com/discovery/feature-extraction.html> [Last accessed: 5/28/2024].
26. Vieira S, Garcia-Dias R, Lopez Pinaya WH, et al. A step-by-step tutorial on how to build a machine learning model. *Machine Learning: Methods and Applications to Brain Disorders* 2020;343–370; doi: 10.1016/B978-0-12-815739-8.00019-5.
27. Christ M, Braun N, Neuffer J, et al. Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* 2018;307:72–77; doi: 10.1016/J.NEUCOM.2018.03.067.
28. Reitermanová Z. Data Splitting. 2010.
29. Kursu MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw* 2010;36(11):1–13; doi: 10.18637/JSS.V036.I11.
30. Hussain S, Raza Z, Giacomini G, et al. Support Vector Machine-Based Classification of Vasovagal Syncope Using Head-Up Tilt Test. *Biology (Basel)* 2021;10(10); doi: 10.3390/BIOLOGY10101029.

31. Hussain S, Raza Z, Kumar T, et al. Diagnosing Neurally Mediated Syncope Using Classification Techniques. *J Clin Med* 2021;10(21); doi: 10.3390/JCM10215016.
32. Khodor N, Matelot D, Carrault G, et al. Kernel based support vector machine for the early detection of syncope during head-up tilt test. *Physiol Meas* 2014;35(10):2119–2134; doi: 10.1088/0967-3334/35/10/2119.
33. Schang D, Feuilloy M, Plantier G, et al. Early prediction of unexplained syncope by support vector machines. *Physiol Meas* 2007;28(2):185–197; doi: 10.1088/0967-3334/28/2/007.
34. Suthaharan S. *Machine Learning Models and Algorithms for Big Data Classification*. Integrated Series in Information Systems. Springer, Boston, MA; 2016.; doi: 10.1007/978-1-4899-7641-3\_10.
35. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016*;13-17-August-2016:785–794; doi: 10.1145/2939672.2939785.
36. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* 2020;415:295–316; doi: 10.1016/J.NEUCOM.2020.07.061.
37. Fortrat JO, Schang D, Bellard E, et al. Cardiovascular variables do not predict head-up tilt test outcome better than body composition. *Clinical autonomic research* 2007;17(4):206–210; doi: 10.1007/S10286-007-0423-2.
38. He Z, Du L, Du S, et al. Machine learning for the early prediction of head-up tilt testing outcome. *Biomed Signal Process Control* 2021;69:102904; doi: 10.1016/J.BSPC.2021.102904.
39. Kefalas M, Koch M, Geraedts V, et al. Automated Machine Learning for the Classification of Normal and Abnormal Electromyography Data. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020* 2020;1176–1185; doi: 10.1109/BIGDATA50022.2020.9377780.
40. Christ M, Braun N, Neuffer J, et al. Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* 2018;307:72–77; doi: 10.1016/J.NEUCOM.2018.03.067.
41. Pedregosa F, Michel V, Grisel OLIVIERGRISEL O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011;12(85):2825–2830.
42. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature* 2020;585(7825):357–362; doi: 10.1038/S41586-020-2649-2.
43. Stein B Van, Wang H, Back T. Automatic Configuration of Deep Neural Networks with Parallel Efficient Global Optimization. *Proceedings of the International Joint Conference on Neural Networks 2019*;2019-July; doi: 10.1109/IJCNN.2019.8851720.
44. Lundberg SM, Allen PG, Lee S-I. *A Unified Approach to Interpreting Model Predictions*. 2017.
45. Anonymous. Tsfresh.Feature\_extraction Package — Tsfresh 0.20.2.Documentation. n.d. Available from: [https://tsfresh.readthedocs.io/en/main/api/tsfresh.feature\\_extraction.html#tsfresh.feature\\_extraction.feature\\_calculators.spkt\\_welch\\_density](https://tsfresh.readthedocs.io/en/main/api/tsfresh.feature_extraction.html#tsfresh.feature_extraction.feature_calculators.spkt_welch_density) [Last accessed: 9/21/2024].

46. Couceiro R, Carvalho P, Paiva RP, et al. Real-Time Prediction of Neurally Mediated Syncope. *IEEE J Biomed Health Inform* 2016;20(2):508–520; doi: 10.1109/JBHI.2015.2408994.
47. Schang D, Bellard E, Plantier G, et al. Comparison of computational algorithms applied on transthoracic impedance waveforms to predict head-up tilt table testing outcome. *Comput Biol Med* 2006;36(3):225–240; doi: 10.1016/J.COMPBIOMED.2004.09.004.
48. Myrovali E, Fragakis N, Vassilikos V, et al. Efficient syncope prediction from resting state clinical data using wavelet bispectrum and multilayer perceptron neural network. *Med Biol Eng Comput* 2021;59(6):1311–1324; doi: 10.1007/S11517-021-02353-7.
49. Jardine DL. Vasovagal Syncope: New Physiologic Insights. *Cardiol Clin* 2013;31(1):75–87; doi: 10.1016/J.CCL.2012.10.010.
50. Klemenc M, Štrumbelj E. Predicting the outcome of head-up tilt test using heart rate variability and baroreflex sensitivity parameters in patients with vasovagal syncope. *Clin Auton Res* 2015;25(6):391–398; doi: 10.1007/S10286-015-0318-6.
51. Ciliberti MAP, Santoro F, Di Martino LFM, et al. Predictive value of very low frequency at spectral analysis among patients with unexplained syncope assessed by head-up tilt testing. *Arch Cardiovasc Dis* 2018;111(2):95–100; doi: 10.1016/J.ACVD.2017.04.006.
52. Nazari Z, Yu SM, Kang D, et al. Comparative Study of Outlier Detection Algorithms for Machine Learning. *ACM International Conference Proceeding Series* 2018;47–51; doi: 10.1145/3234804.3234817.
53. Liu W, Hua G, Smith JR. Unsupervised one-class learning for automatic outlier removal. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2014;3826–3833; doi: 10.1109/CVPR.2014.483.

# 7. Appendix

---

## A. Preliminary Study

### 1. Objective

This study was intended as a preliminary study to test which dataset is most suitable to extract features and to test different ways of selecting features. This includes evaluating different techniques for handling missing data, determining which segments of the data are optimal for analysis, assessing whether there are any artifacts and outliers that need to be removed, and the feature selection. The aim is therefore to:

- 1) Determine which portion of the data, along with any necessary pre-processing, can best be used for HR and BP data in patients with suspected VVS during HUTT.

### 2. Methods

The same dataset and ML pipeline was used as described in the primary study unless otherwise specified in the methods.

#### 2.1. Data Selection

The age, sex, heart rate (HR) and blood pressure (BP) data of all included patients were used. The three patients excluded on the basis of body mass index (BMI) were still included in this part of the study.

#### 2.2. ML Pipeline

In this preliminary study, only a RF classifier was utilized. In order to conserve time, only one classifier was used. RF was chosen due to its simplicity, robustness against overfitting, and capability to capture complex interactions between variables.<sup>1</sup>

In addition, the pipeline parameters are the same as described in the main study, except that the maximum iterations are set to 2 rather than 100.

#### 2.3. Filling of Missing Data

Three different interpolation methods were compared for their effectiveness in filling the non-numeric cells in the data: linear, quadratic, and cubic. To evaluate their performance, five consecutive data points were randomly removed from a copy of the original data and then filled using one of the interpolation methods. The interpolated values were then compared to the original data, and the mean error percentage and standard deviation (std) were calculated. This process was repeated for 30 random iterations and the results were averaged to determine the most effective method. (*Figure 1*)

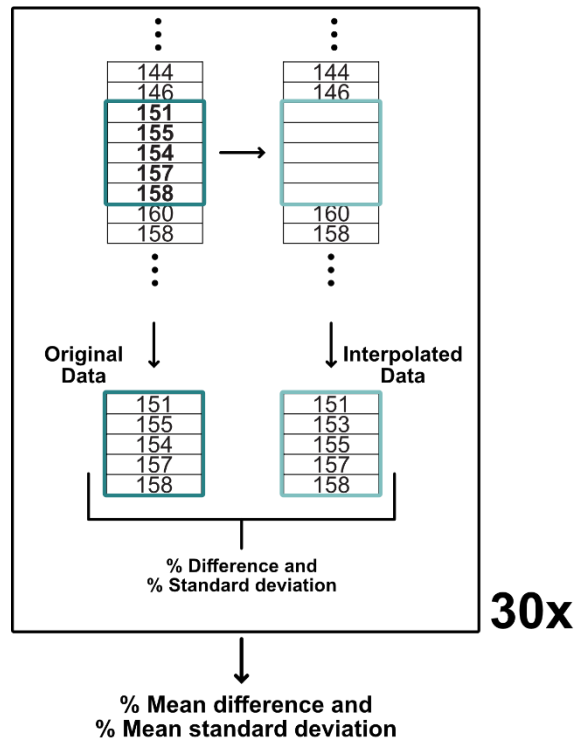


Figure 1. Schematic illustration of the interpolation testing, which involves taking a random sample of data, interpolating it, and comparing it to the original data.

## 2.4. Selecting Data

Two different options were considered to select the appropriate variables; the basic BP and HR variables, and the addition of extra variables provided by the device software. A summary of the basic and extra HR and BP variables used can be found in *Appendix B*.

To assess the impact of data length on model performance, a comparison was made using different durations and data epochs. The dataset was tested with data intervals of 1, 3, and 5 minutes, and different approaches were evaluated. These approaches included analyzing data from the entire period around the tilt, separating data into before and after the tilt, subtracting features extracted from before and after tilt data, and considering only the data collected after the tilt. *Figure 2* illustrates the different options.

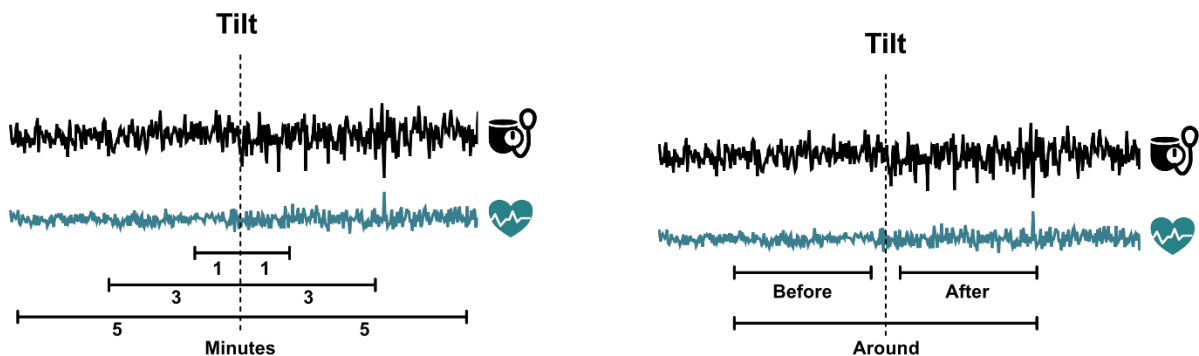


Figure 2. The different ways in which data can be selected. This can be done by taking different lengths of data or by specifying the part of the data to be selected.

## 2.5. Artifacts and Outliers

To reduce the artifacts produced by tilting, we also looked at removing the data points around the tilting point. This involved two approaches: one where the data around the tilt point was retained when using a single epoch, and another where 20 seconds of data on either side of the marked tilt point were excluded. This second approach was applied when the data were divided into multiple epochs.

The influence of artifacts and outliers was tested by processing the data in three different ways. We examined the use of raw data compared to two different artifact removal methods. One of these methods was the simple exponential smoothing algorithm with an alpha of 0.05.<sup>2</sup> This method was used to filter out artifacts and outliers, using the shape of the signal as the main information component. The other method was an outlier detection function that uses a threshold, mean and std of the previous 10 point window to detect outliers. Once an outlier was detected, that point and the three surrounding points were removed from that variable and the other variables, since a single variable was often used to calculate other variables in which the outliers are also present. An overview of the exact settings and an example can be found in the *Appendix K*. The removed data points were filled using the best tested interpolation method.

## 2.6. Feature Selection Using Boruta

In order to examine the effect of feature selection, different methods of Boruta usage were tested. A comparison was made between the original method, where Boruta was run through each fold to select the best features, different Boruta thresholds ('perc'), including 85, 95 and 98 percent, and Boruta in a loop (5 and 10 iterations), where all features of the iterations were combined. Finally, we also investigated whether manually selecting the best performing features of a fold works well, where a 10-fold CV was performed, selecting the features of the best and worst performing fold (based on 60, 70, or 80 percent of the data). These features were then tested on the remaining part of the data (40, 30, or 20 percent of the data) in the rest of the pipeline. In this way, the feature selection was tested separately from the automated model using Boruta.

## 2.7. Performance Evaluation

The mean AUC of the 5-fold CV was used as performance value for the comparison of the effectiveness of the different models.

# 3. Results

## 3.1. Filling of Missing Data

*Table 1. The average mean and standard deviation (std) percentage of 30 interpolation comparisons. It was tested per variable for linear, quadratic and cubic interpolation. The meaning of the variables is further explained in Appendix B.*

SYS			DIA			MAP		
Method	Mean	Std	Method	Mean	Std	Method	Mean	Std
Linear	-0,14%	4,51%	Linear	-0,27%	3,98%	Linear	-0,07%	3,47%
Quadratic	-0,3%	5,96%	Quadratic	0,25%	5,94%	Quadratic	-0,1%	5,03%
Cubic	-0,38%	6,5%	Cubic	-0,29%	5,49%	Cubic	0,03%	5,2%

IBI			LVET			SVI		
Method	Mean	Std	Method	Mean	Std	Method	Mean	Std
Linear	-0,29%	8,87%	Linear	0,003%	2,45%	Linear	-0,29%	10,85%
Quadratic	-0,45%	14,46%	Quadratic	-0,07%	4,24%	Quadratic	-1,32%	17,82%
Cubic	-0,88%	13,64%	Cubic	-0,07%	4,26%	Cubic	-0,68%	19,49%

CI			HR			SVRI		
Method	Mean	Std	Method	Mean	Std	Method	Mean	Std
Linear	-0,33%	13,56%	Linear	0,15%	7,56%	Linear	-0,09%	5,51%
Quadratic	-0,87%	28,29%	Quadratic	0,59%	11,29%	Quadratic	-0,21%	6,38%
Cubic	-0,91%	22,32%	Cubic	-1,05%	13,72%	Cubic	-0,11%	6,34%

It can be seen in *Table 1* that for the all variables, the linearly interpolated data deviates, on average, less than 0.5% from the original data and has a smaller error than the quadratic and cubic interpolation methods.

### 3.2. Selecting Data

In the two datasets where the only distinction was between using basic HR and BP variables versus using both basic and extra HR and BP variables, the results showed the same mean AUC of 61%. A similar approach was applied to a 3-minute dataset after the tilt, which yielded an AUC of 57% when only the basic variables were used and an AUC of 59% when the extra variables were added.

In *Table 2* it can be seen that the overall performance of the model works just as good and sometimes even better with using 3 minutes of data compared to 5 minutes.

*Table 2. AUC percentage comparing three different feature sets when using data from one, three and five minutes of data around the tilt.*

Features used	% AUC 1 minute	% AUC 3 minutes	% AUC 5 minutes
After the tilt with raw data	51	59	56
After the tilt with smoothed data	46	53	57
Before and after the tilt separated with smoothed data	51	57	56

Different epochs and corresponding results can be found in *Table 3* with the best results coming from the features of data split from 3 minutes before the tilt and 3 minutes after the tilt. An additional epoch of 3-5 minutes did not improve the results.

*Table 3. AUC percentage of five different feature sets each of which uses a different epoch of data. It is tested on two different variable sets.*

Features used	% AUC when using Basic variables	% AUC when using Extra variables
0-3 and 3-5 minute epochs before and after the tilt		59
3 minutes after the tilt	57	59
3 minutes before and after the tilt separated	61	61
3 minutes around the tilt in one epoch	57	
3 minutes before and after the tilt subtracted after feature extraction	49	



### 3.3. Artifacts and Outliers

When looking at the different methods for detecting outliers and artifacts, the raw data performs better for all three different feature sets compared to the outlier detection algorithm and exponential smoothing (Table 4.)

Table 4. AUC percentage from three different feature sets testing 3 different outlier and artifact detection methods including the use of raw data, an outlier detection algorithm and an exponential smoothing algorithm

Features used	% AUC when using raw data	% AUC when using outlier detection	% AUC when using exponential smoothing
3 minutes around the tilt in one epoch	57	52	52
3 minutes after the tilt	59	57	53
3 minutes before and after the tilt separated	61	53	57

### 3.4. Feature Selection Using Boruta

When applying different Boruta thresholds, the performance variation was minimal. With an 85% threshold, an AUC of 62% was achieved. However, the number of features used in the model was significantly higher, with an average of 479 features, compared to only 6 features selected with a 100% threshold. (Table 5)

Table 5. AUC percentage of the feature set based on the separated data before and after the tilt with a varying Boruta selection threshold. Together with the average number of features selected by Boruta.

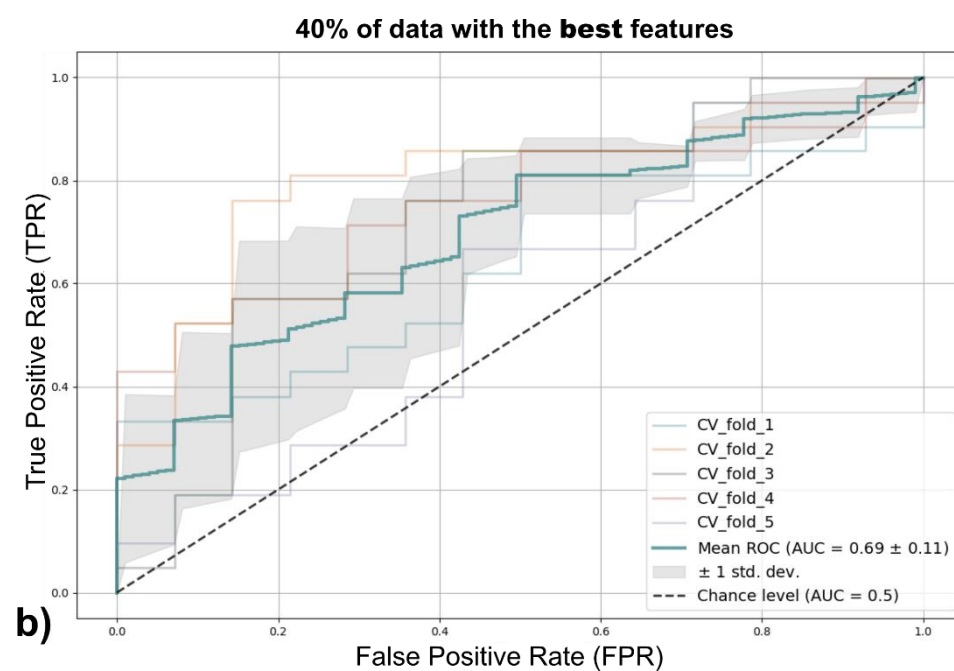
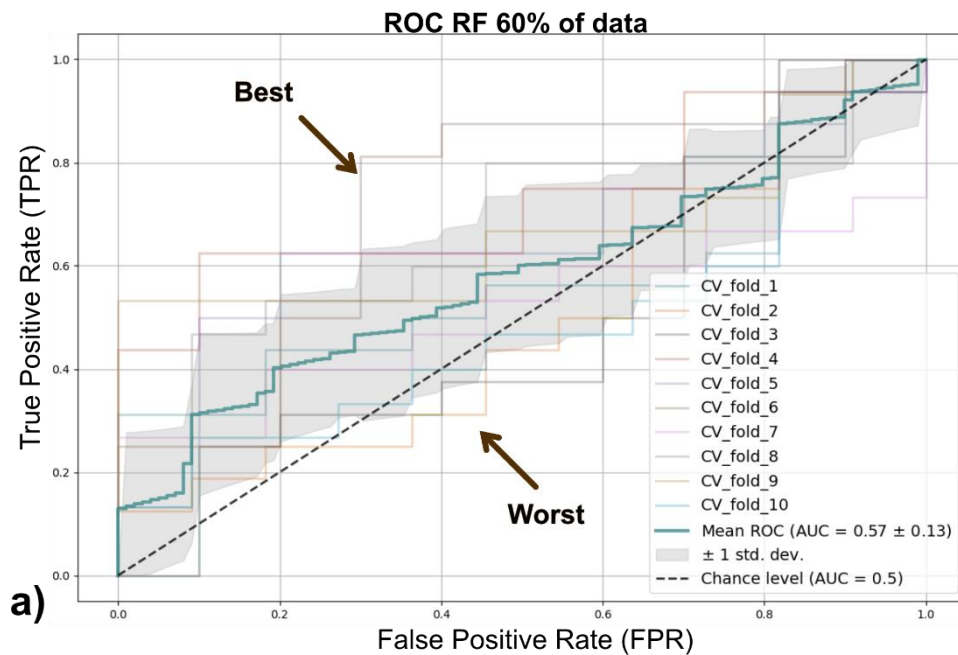
Features	% AUC when using perc=85 in Boruta	% AUC when using perc=95 in Boruta	% AUC when using perc=98 in Boruta	% AUC when using perc=100 in Boruta
3 minutes before and after the tilt separated	62	60	60	61
Mean amount of selected features	479	152	54	6

Using Boruta in a loop also resulted in more feature selections, but on average, no more than 17 features were selected. Despite this, no improvement in performance was observed. (Table 6)

Each of the methods used above had a std ranging from 4% and 8%.

Table 6. AUC percentage of the feature set based on the separated data before and after the tilt with a loop around the Boruta feature selector. Together with the average number of total features selected by Boruta.

Features	% AUC when using 1 iteration	% AUC when using 5 iterations	% AUC when using 10 iterations
3 minutes before and after the tilt separated	61	60	61
Mean amount of selected features	6	12	17



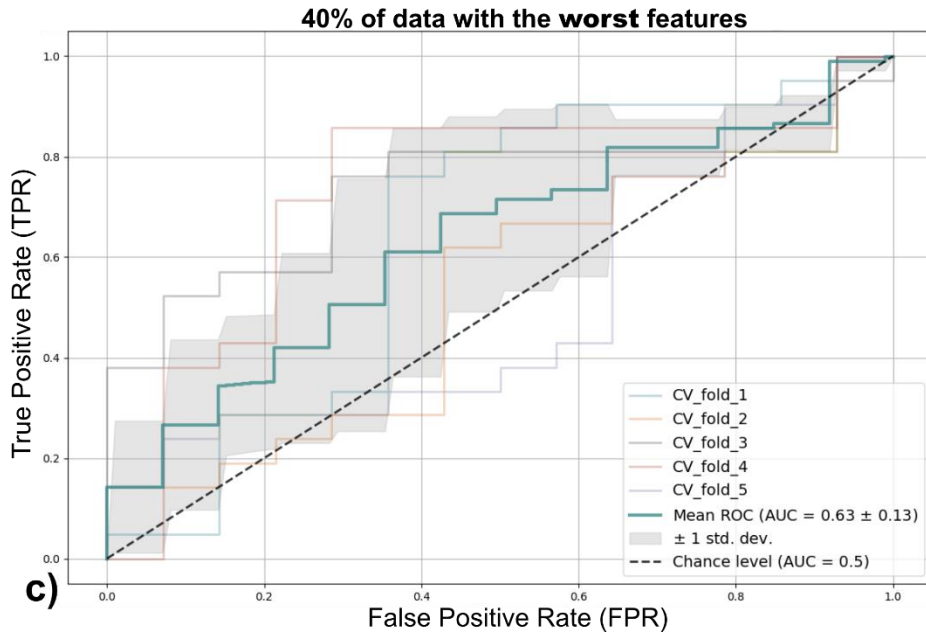


Figure 3. An example of feature selection based on the best and worst performing folds. a) 10-fold cross-validation on 60% of the data, where the features of the best and worst performing fold were selected. b) 5-fold cross-validation of 40% of the data with the features from the best performing fold. c) 5-fold cross-validation of 40% of the data with the features from the worst performing fold.

Figure 3 shows an example of a 10-fold CV on the split data, with feature selection based on the best and worst performing folds. This method was chosen to test whether predefined features are better for classification. Previous results on the whole dataset using pre-defined features based on the best performing fold gave better classification results. In order to test this in a reliable way, it was decided to validate part of the data separately during this test. Table 7 shows that on average the validation group outperforms the training data in all cases. However, the std was between 10% and 15%, which was higher than the std observed with the previously mentioned methods.

Table 7. AUC percentage of feature selection based on best and worst performing folds. The results are based on the features extracted from the separated data of 3 minutes before and after the tilt. The method was tested on different data splits where 80, 70 or 60% of the data were used to select the best and worst features. The remaining 20, 30 or 40% of the data were used to test the validity of this method with the selected best and worst features.

Features: 3 minutes before and after the tilt separated	% AUC on training of data for features	% AUC of validation data with the best features	% AUC of validation data with the worst features
80/20 split	57	79	-
70/30 split	53	68	69
60/40 split	57	69	63

## 4. Discussion

In this preliminary study, various datasets were evaluated for feature extraction and selection, focusing on the filling of missing data, data selection, artifacts and outliers, and feature selection. Linear interpolation was found to be the most effective method for the filling of missing data. Although the performance differences between the different datasets were small, feature sets that included the extra variables, separated 3 minutes of data before and after tilt, and used raw data generally produced the best results, with an average AUC of 61%. For feature selection the original Boruta method proved to be the most suitable in this context, offering a balance between automation, feature reduction, and model stability.

### 4.1. Filling of Missing Data

The study by Hussain et al. (2021) investigated the prediction of syncope using ML on HUTT data and addressed missing data through imputation by averaging indicators across patients.<sup>3</sup> However, this approach may not account for variability between patients. Therefore, interpolation was selected for the current study due to its effectiveness with relatively continuous data and predictable trends.<sup>4</sup> This method is particularly effective when only a few data points are missing, which was generally the case with the missing data points in this dataset. It was validated by testing on sequences with up to five consecutive missing data points. However, instances of more than 50 consecutive missing data points, although rare, could have affected the results, as the accuracy of interpolation tends to decrease with an increase of the number of missing data points.

### 4.2. Selecting Data

The integration of ML with signal analysis of HUTT data opens new possibilities for early prediction and accurate classification of syncope by combining traditional medical assessment with ML classification techniques. However, to date, research using classical ML models for early prediction shows that it would be better to use 15 minutes of data after the tilt compared to 5 minutes. This comparison was made by Khodor et al. (2016) between classification at 5 minutes after the tilt and classification at 15 minutes after the tilt. Sensitivity increased from 69.6% to 87.5% and specificity increased from 66.9% to 93.8% when using 15 minutes versus 5 minutes after the tilt.<sup>5</sup> While several studies have favored a 15-minute after the tilt period for early syncope prediction, this approach may counteract the goal of reducing the duration of HUTT, particularly when employing protocols like the 'Fast Italian Protocol.'<sup>5,6</sup>

Although several studies have used data from 15 minutes after tilt to predict syncope, it remains unclear whether subjects who had already experienced syncope during this period were excluded. If not, these models are diagnostic rather than predictive. While several studies refer to 15 minutes as early prediction, true early prediction would occur in the first few minutes before or just after the tilt, well before the onset of syncope. In contrast, our study focused on earlier time points. We compared 3-minute and 5-minute epochs to investigate whether syncope could be predicted more accurately in a shorter time frame. Our results indicate that the difference in performance between these epochs was minimal. In addition, we found that using data segmented into before and after tilt phases produced better classification results than using after tilt data alone.

### 4.3. Artifacts and Outliers

In the context of pre-processing HUTT syncope data for ML applications, there is limited information available in the literature. Pre-processing typically involves normalizing and resampling the data, occasionally applying filters, or transforming the data into the frequency domain for feature extraction.<sup>6-11</sup> Most studies do not address the extraction of artifacts or outliers from the data. This indicates that using raw data for HR and BP variables generally yields the best results. These findings

are consistent with the results of the current study. However, this does not eliminate the potential impact of outliers or artifacts on classification performance.

#### 4.4. Feature Selection Using Boruta

The original Boruta method was selected as the optimal feature selection technique within this pipeline. However, splitting the data gave better results, but only a smaller set could be used for training and validating the model. The use of a smaller set and the variation within that set may have been a reason for the better performance. This could have led to less stability and reliability, as indicated by a larger std in the results. A future study with a larger cohort of patients could further investigate these findings and test whether the increase in performance is consistent and whether manual feature selection could further improve the results.

Although Boruta has not yet been applied to syncope data from HUTT, it has produced good results in research with electroencephalogram (EEG) and electromyography (EMG) data.<sup>12,13</sup> However, there are other feature selection methods that could be explored. For example, a study by Khodor et al. (2016) compared three methods and found that sequential forward selection (SFS) performed best. SFS works by adding features to a candidate set one at a time until further additions no longer improve performance based on a given criterion.<sup>5</sup> Future studies could test whether alternative methods such as SFS provide better results with HUTT data.

#### 4.5. Limitations

There are several limitations to this study. First, the results are based exclusively on a RF model. While RF performed well, there are other models that may be better suited for this type of analysis. This will be more of a focus in the primary study. The choice of dataset and pre-processing was based on the best performing RF model, which may have influenced the overall pipeline, potentially biasing the model development towards RF specific features, when other classifiers might provide better results.

In addition, only the tsfresh package was used for feature extraction. This method was chosen because of its wide range of calculated feature types. However, it lacks features derived from medical expertise that have been used in previous studies.<sup>5,9</sup> Although most relevant or similar features are included, important features that could better distinguish between the two groups may still be missing.

Another limitation is that only two different iterations were used for HO, due to the broad purpose of the preliminary study and to save time. Although each iteration considered ten different outcomes along with a 10-fold CV, this remains a limitation and should be addressed in further research.

In testing various options to improve the model, the primary focus was on optimizing the data for feature extraction and selection. Despite the robust performance of the pipeline in previous research, it is possible that other settings could have been optimized. Future research should explore these possibilities more thoroughly.

Finally, it should be taken into account that not every possible combination of features was tested. Some combinations were chosen for comparison purposes, but features based on different datasets may yield different results that were not explored. Additionally, there was variation in the outcomes of the same models because, with the 5-fold split, different groups were created each time the model was run.

## 5. Conclusion

In this preliminary study, we compared several datasets for feature extraction and selection. The results suggest that the most effective approach for further research is to use a feature set derived from raw data, focusing on the 3 minutes before and after the tilt in separated epochs, and to use linear interpolation to fill in the missing data. In addition, using the original Boruta algorithm proves to be an effective method for feature selection.

## 6. References

1. Breiman L. Random forests. *Mach Learn* 2001;45(1):5–32; doi: 10.1023/A:1010933404324/METRICS.
2. Seabold S, SciPy JP-, 2010 undefined. *Statsmodels: econometric and statistical modeling with python*. Proceedings of the 9th Python in Science Conference 2010.
3. Hussain S, Raza Z, Kumar T, et al. Diagnosing Neurally Mediated Syncope Using Classification Techniques. *J Clin Med* 2021;10(21); doi: 10.3390/JCM10215016.
4. Couceiro R, Carvalho P, Paiva RP, et al. Real-Time Prediction of Neurally Mediated Syncope. *IEEE J Biomed Health Inform* 2016;20(2):508–520; doi: 10.1109/JBHI.2015.2408994.
5. Khodor N, Carrault G, Matelot D, et al. Early syncope detection during head up tilt test by analyzing interactions between cardio-vascular signals. *Digit Signal Process* 2016;49:86–94; doi: 10.1016/J.DSP.2015.11.005.
6. He Z, Du L, Du S, et al. Machine learning for the early prediction of head-up tilt testing outcome. *Biomed Signal Process Control* 2021;69:102904; doi: 10.1016/J.BSPC.2021.102904.
7. Klemenc M, Štrumbelj E. Predicting the outcome of head-up tilt test using heart rate variability and baroreflex sensitivity parameters in patients with vasovagal syncope. *Clin Auton Res* 2015;25(6):391–398; doi: 10.1007/S10286-015-0318-6.
8. Ciliberti MAP, Santoro F, Di Martino LFM, et al. Predictive value of very low frequency at spectral analysis among patients with unexplained syncope assessed by head-up tilt testing. *Arch Cardiovasc Dis* 2018;111(2):95–100; doi: 10.1016/J.ACVD.2017.04.006.
9. Schang D, Feuilloy M, Plantier G, et al. Early prediction of unexplained syncope by support vector machines. *Physiol Meas* 2007;28(2):185–197; doi: 10.1088/0967-3334/28/2/007.
10. Khodor N, Matelot D, Carrault G, et al. Kernel based support vector machine for the early detection of syncope during head-up tilt test. *Physiol Meas* 2014;35(10):2119–2134; doi: 10.1088/0967-3334/35/10/2119.
11. Hussain S, Raza Z, Giacomini G, et al. Support Vector Machine-Based Classification of Vasovagal Syncope Using Head-Up Tilt Test. *Biology (Basel)* 2021;10(10); doi: 10.3390/BIOLOGY10101029.
12. Kefalas M, Koch M, Geraedts V, et al. Automated Machine Learning for the Classification of Normal and Abnormal Electromyography Data. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020* 2020;1176–1185; doi: 10.1109/BIGDATA50022.2020.9377780.

13. Geraedts VJ, Koch M, Contarino MF, et al. Machine learning for automated EEG-based biomarkers of cognitive impairment during Deep Brain Stimulation screening in patients with Parkinson's Disease. *Clinical Neurophysiology* 2021;132(5):1041–1048; doi: 10.1016/J.CLINPH.2021.01.021.

## B. Variables

Table 1. Basic BP and HR Variables.

Variable	Meaning
reSYS	systolic brachial pressure
reDIA	diastolic brachial pressure
reMAP	mean arterial pressure
HR	Pulse rate derived from atrial pressure signal
IBI	Inter Beat interval (heart rate variability)
LVET	Left Ventricular Ejection Time, the time of ejection of blood from the left ventricle beginning with aortic valve opening and ending with aortic valve closure.
SVI	Stroke Volume Indexed (Stroke volume/body mass area)
CI	Cardiac output Indexed (Cardiac output /body mass area)
SVRI	Total Peripheral Resistance (TPR) Indexed, Parameter of Windkessel model used to reconstruct flow from pressure. (TPR /body mass area)

Table 2. Extra BP and HR Variables.

Variable	Meaning
mFlow	Model computation of the aortic flow waveform
dP_dt	Maximal steepness of the current upstroke
SPTI	Systolic Pressure Time Index computed as the area under the systolic portion of the arterial pulse
RPP	Rate Pressure Product computed as the product of systolic pressure and pulse rate and is indexed for cardiac oxygen demand per min (=SYS*HR)
DPTI	Diastolic Pressure Time Index computed as the area under the diastolic portion of the arterial pulse
DPTI_SPTI	Ratio is an index of cardiac oxygen supply / demand
ZAo	Parameter of Windkessel model used to reconstruct flow from pressure: ascending aorta characteristic impedance (Z) at diastolic pressure, impedance of arterial system
Cwk	Parameter of Windkessel model used to reconstruct flow from pressure: Windkessel compliance (C), total arterial compliance at diastolic pressure



## C. Python packages and versions

Table 1. Python packages for the Machine learning Pipeline.

Package	Version
<b>boruta</b>	0.3
<b>matplotlib</b>	3.5.1
<b>mipego</b>	2.0.0
<b>numpy</b>	1.22.4
<b>pandas</b>	1.3.5
<b>scikit-learn</b>	1.0.2
<b>scipy</b>	1.7.3
<b>shap</b>	0.38.1
<b>tabulate</b>	0.8.9
<b>xgboost</b>	2.0

Table 2. Python packages for the Pre-processing.

Package	Version
<b>matplotlib</b>	3.9.0
<b>numpy</b>	1.26.4
<b>pandas</b>	2.2.2
<b>scikit-learn</b>	1.4.2
<b>scipy</b>	1.7.3
<b>statsmodel</b>	0.14.1
<b>tsfresh</b>	0.20.2

## D. Machine learning Pipeline overview

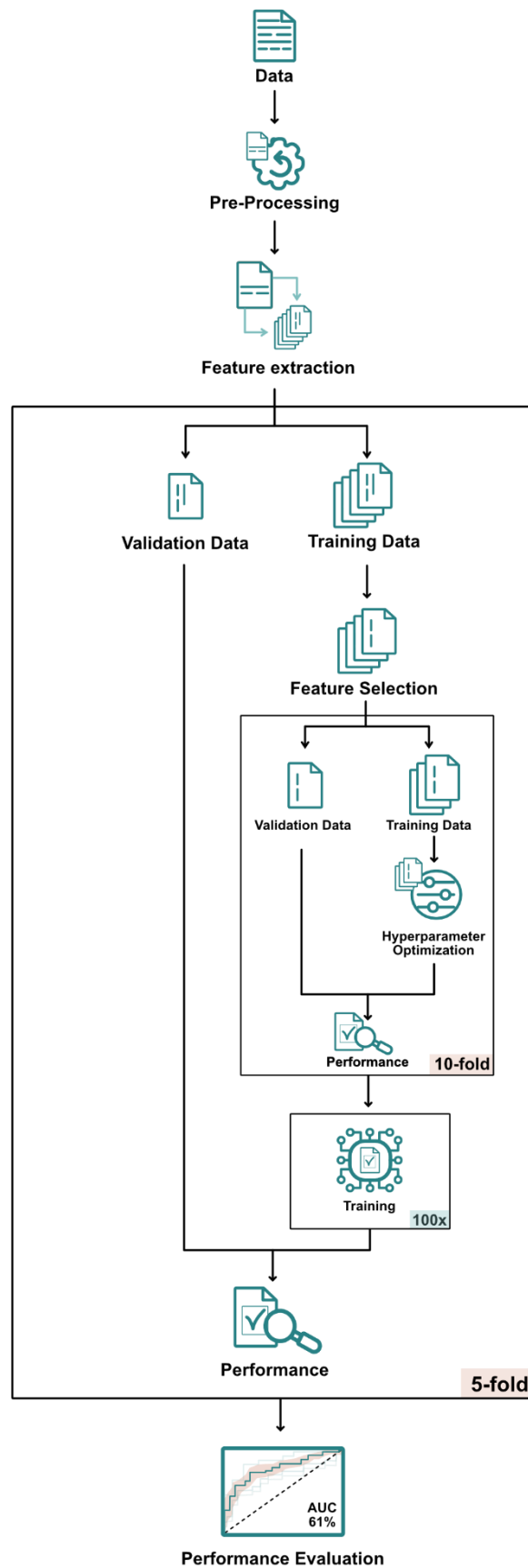


Figure 1. An overview of the Machine learning pipeline that was used.

## E. Hyperparameter Search Space

Table 1. Random Forest Hyperparameter search space

<b>max_depth</b>	NominalSpace([None] + np.arange(2, 102, 2).tolist(), var_name="max_depth")
<b>n_estimators</b>	OrdinalSpace([1, 1000], var_name="n_estimators")
<b>min_samples_leaf</b>	OrdinalSpace([1, 10], var_name="min_samples_leaf")
<b>min_samples_split</b>	OrdinalSpace([2, 20], var_name="min_samples_split")
<b>bootstrap</b>	NominalSpace([True, False], var_name="bootstrap")
<b>max_features</b>	NominalSpace(["auto", "sqrt", "log2"], var_name="max_features")

Table 2. Support Vector Machine Hyperparameter search space

<b>C</b>	OrdinalSpace([1, 50], var_name="C")
<b>kernel</b>	NominalSpace(["poly", "rbf", "sigmoid"], var_name="kernel")
<b>gamma</b>	NominalSpace(["scale"] + ["auto"], var_name="gamma")
<b>coef0</b>	ContinuousSpace([0, 5], var_name="coef0")
<b>probability</b>	True
<b>class_weight</b>	"balanced"

Table 3. XGBoost Hyperparameter search space

<b>max_depth</b>	OrdinalSpace([1, 10], var_name="max_depth")
<b>gamma</b>	ContinuousSpace([0, 10], var_name="gamma")
<b>min_child_weight</b>	OrdinalSpace([1, 10], var_name="min_child_weight")
<b>learning_rate</b>	ContinuousSpace([0, 1], var_name="learning_rate")
<b>colsample_bytree</b>	ContinuousSpace([0, 1], var_name="colsample_bytree")
<b>reg_alpha</b>	ContinuousSpace([0, 1], var_name="reg_alpha")
<b>reg_lambda</b>	ContinuousSpace([0, 1], var_name="reg_lambda")
<b>subsample</b>	0.7
<b>eval_metric</b>	"auc"

## F. Performance formula's

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$F1 = \frac{2 * \text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

**Abbreviations:** *TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.*

## G. ROC-Curves

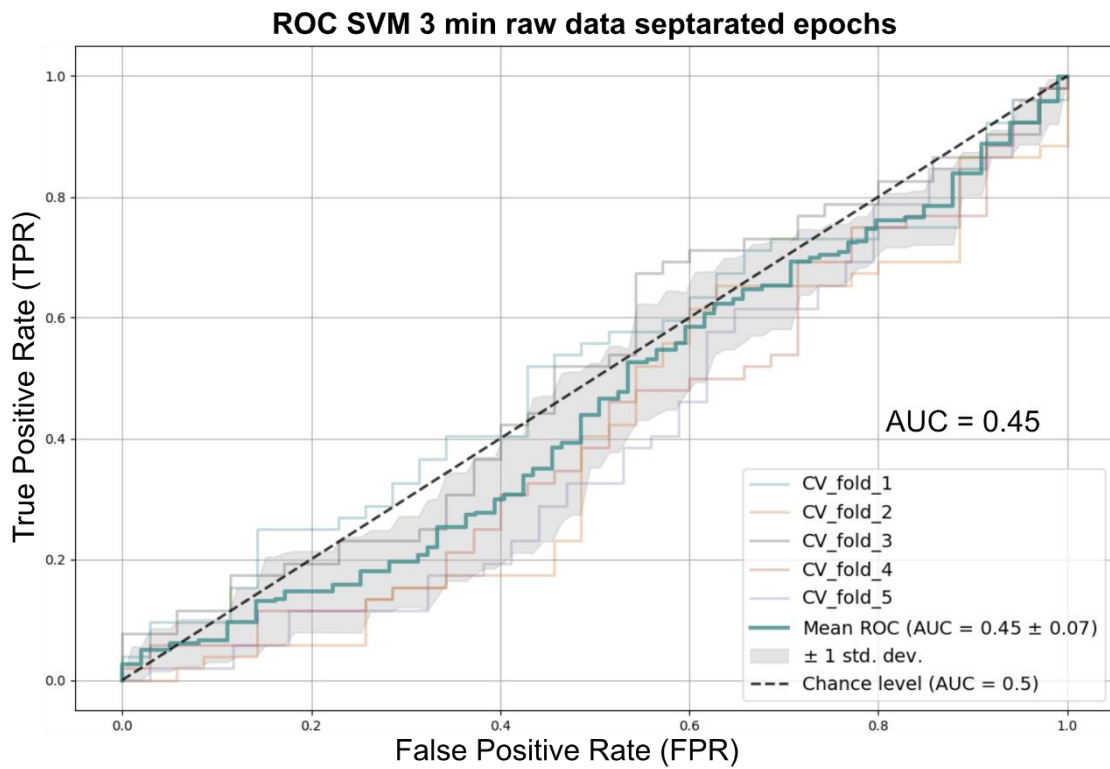


Figure 1. The ROC-curve of the final SVM model with an AUC of 45% with a standard deviation of 7%.

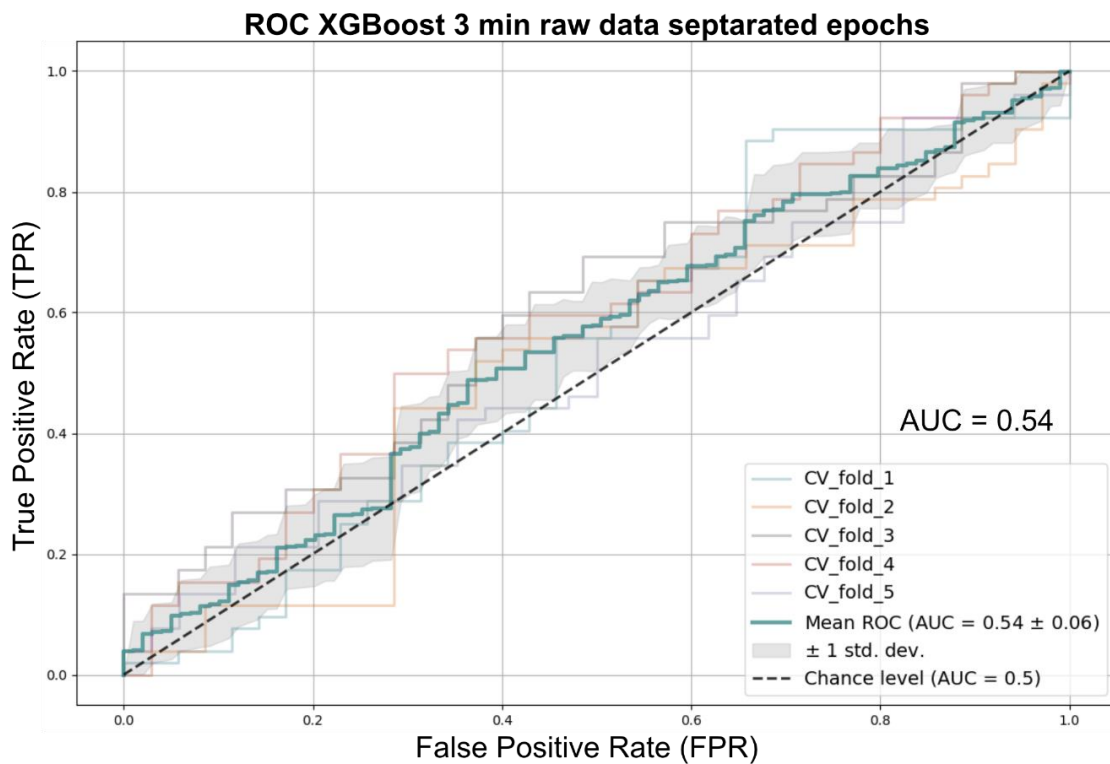


Figure 2. The ROC-curve of the final XGBoost model with an AUC of 54% with a standard deviation of 6%.

## H. Selected Hyperparameters

Table 1. Random Forest Hyperparameters.

Hyperparameter	1-fold	2-fold	3-fold	4-fold	5-fold
max_depth	70	42	100	82	72
n_estimators	292	185	253	22	893
bootstrap	False	False	False	False	False
max_features	'auto'	'sqrt'	'auto'	'auto'	'auto'
min_samples_leaf	4	4	9	1	2
min_samples_split	3	8	10	7	11

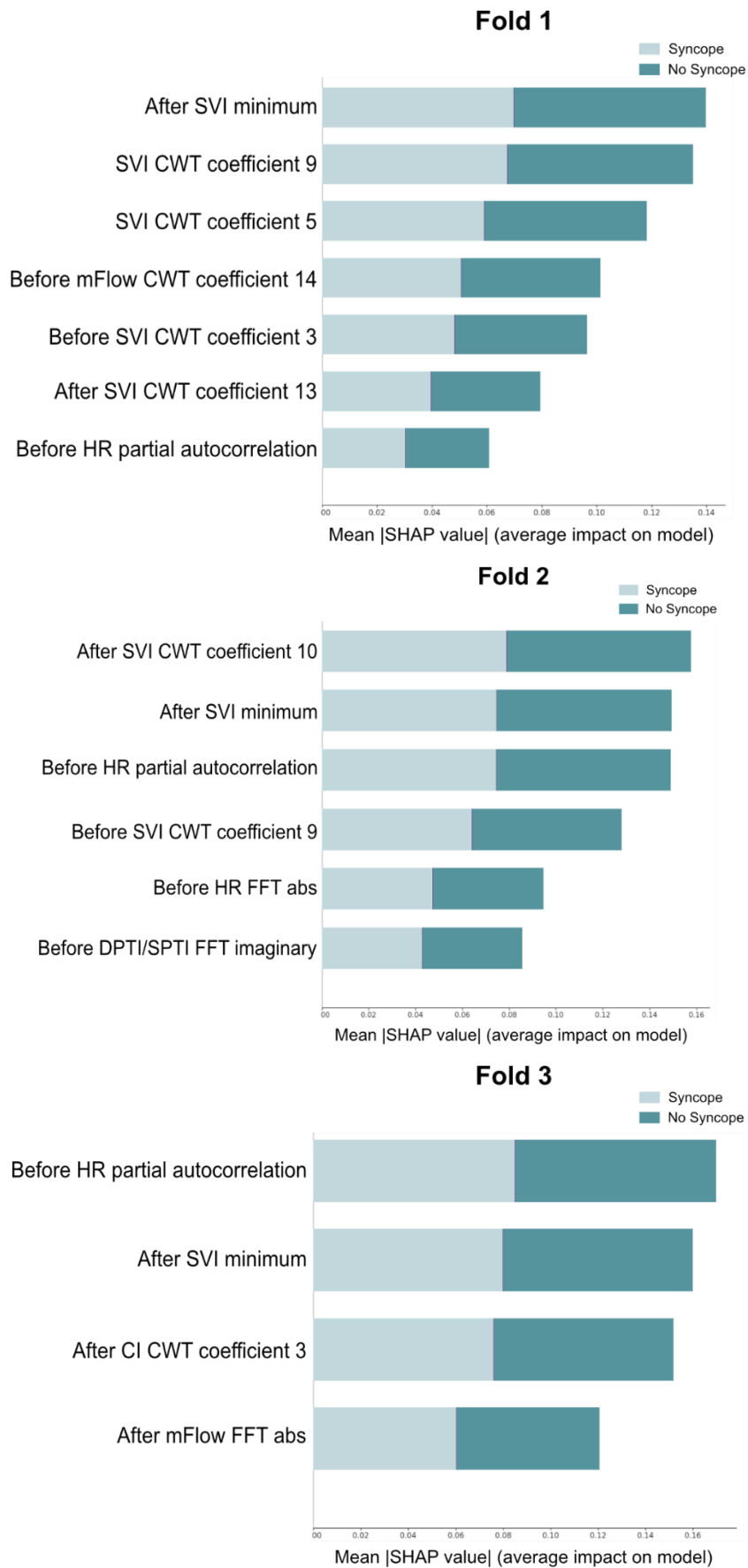
Table 2. XGBoost Hyperparameters.

Hyperparameter	1-fold	2-fold	3-fold	4-fold	5-fold
max_depth	6	5	9	6	9
gamma	1.7	8.5	7.9	3.7	5.1
min_child_weight	1	8	1	2	4
learning_rate	0.12	0.55	0.56	0.99	0.83
colsample_bytree	0.91	0.55	0.91	0.64	0.44
reg_alpha	0.10	0.85	0.83	0.25	0.65
reg_lambda	0.73	0.85	0.06	0.14	0.25

Table 3. Support Vector Machine Hyperparameters.

Hyperparameter	1-fold	2-fold	3-fold	4-fold	5-fold
C	50	36	33	44	48
kernel	'sigmoid'	'sigmoid'	'sigmoid'	'sigmoid'	'sigmoid'
gamma	'scale'	'scale'	'scale'	'scale'	'scale'
coef0	1.11	0.03	1.06	0.22	0.89

# I. SHAP Summary Bar plots



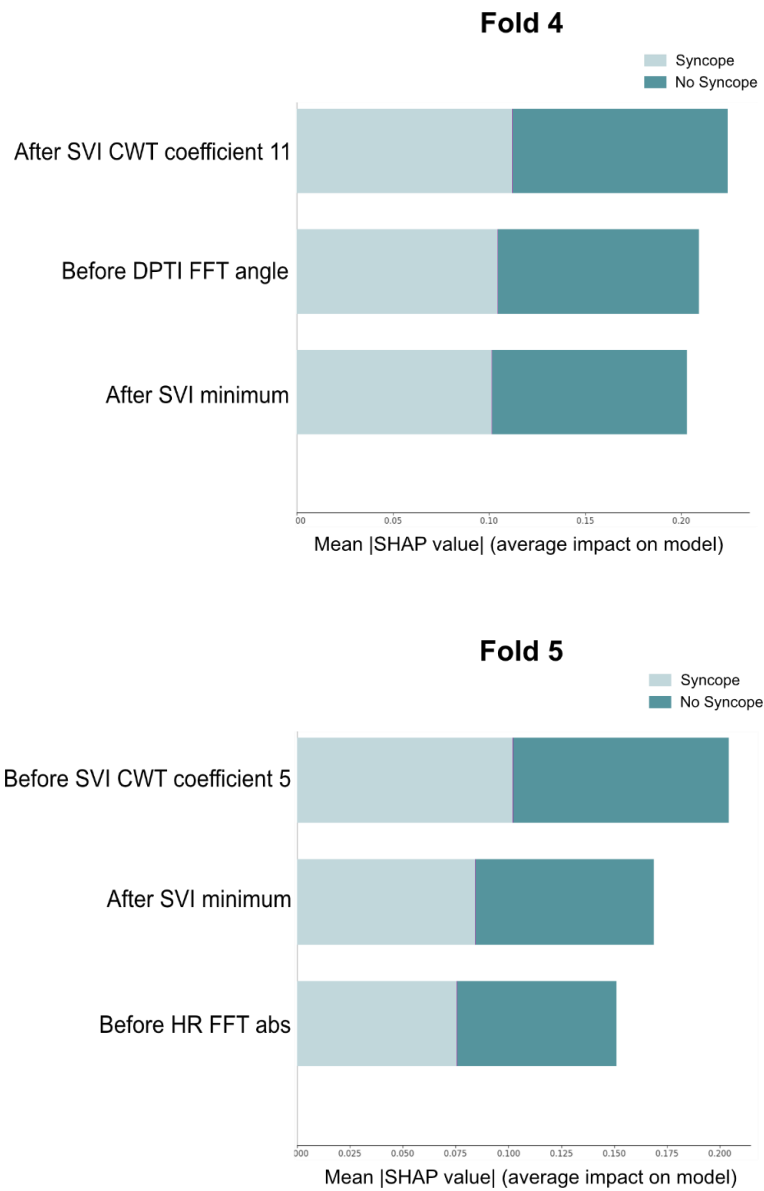


Figure 1. The SHAP summary bar plots for each of the 5 folds. For each fold, the selected features are ordered, with the features that have the greatest influence on the random forest model being at the top. In addition, the two colors represent the influence of that feature on the classification in a certain class.



## J. Selected Features in Random Forest

Table 1. A list of the full feature names based on the variables as mentioned in Appendix K and the feature names according to the tsfresh python package.

Full Feature name	Name in Text and Figures
after_SVI__minimum	After SVI minimum
before_HR__partial_autocorrelation__lag_4	Before HR partial autocorrelation
before_SVI__cwt_coefficients__coeff_5__w_2__widths_(2, 5, 10, 20)	Before SVI CWT coefficient 5
before_SVI__cwt_coefficients__coeff_9__w_2__widths_(2, 5, 10, 20)	Before SVI CWT coefficient 9
before_HR__fft_coefficient__attr_"abs"__coeff_79	Before HR FFT abs
after_SVI__cwt_coefficients__coeff_13__w_5__widths_(2, 5, 10, 20)	After SVI CWT coefficient 13
before_SVI__cwt_coefficients__coeff_3__w_20__widths_(2, 5, 10, 20)	Before SVI CWT coefficient 3
before_mFlow__cwt_coefficients__coeff_14__w_5__widths_(2, 5, 10, 20)	Before mFlow CWT coefficient 14
after_SVI__cwt_coefficients__coeff_10__w_5__widths_(2, 5, 10, 20)	After SVI CWT coefficient 10
before_DPTI_SPTI__fft_coefficient__attr_"imag"__coeff_8	Before DPTI/SPTI FFT imaginary
after_CI__cwt_coefficients__coeff_3__w_10__widths_(2, 5, 10, 20)	After CI CWT coefficient 3
after_mFlow__fft_coefficient__attr_"abs"__coeff_18	After mFlow FFT abs
after_SVI__cwt_coefficients__coeff_11__w_5__widths_(2, 5, 10, 20)	After SVI CWT coefficient 11
before_DPTI__fft_coefficient__attr_"angle"__coeff_1	Before DPTI FFT angle

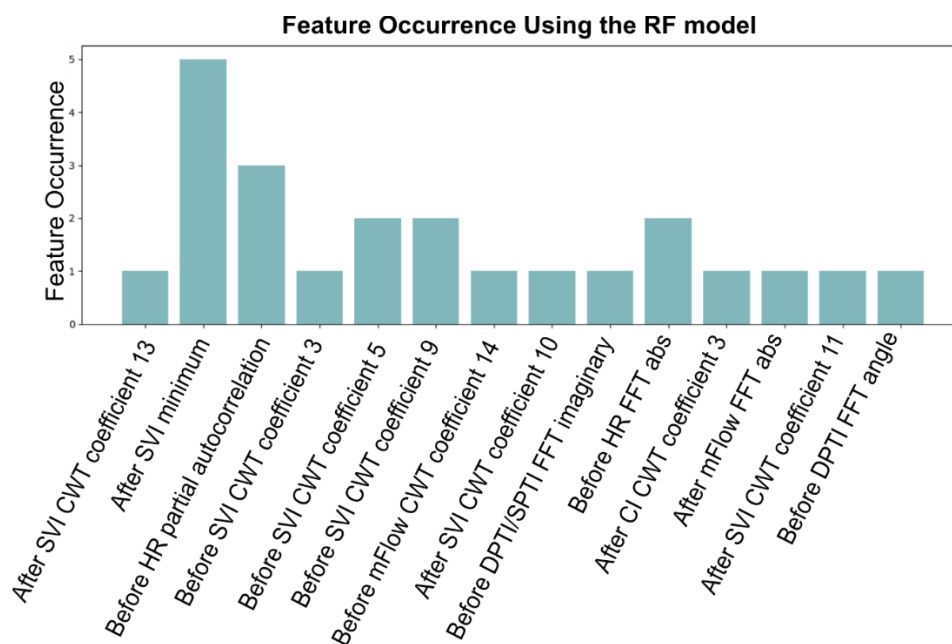


Figure 1. An full overview of the feature occurrence in all 5 folds of the random forest (RF) model.

## K. Outlier and Artifact detection

Table 1. Outlier detection algorithm settings

Variable	Threshold	Average	Standard deviation (std)
reSYS	300		
reDIA	160		
reMAP	250		
HR	40	1	0.5
IBI	300	1	0.1
LVET	70	1	
SVI	10	1	1.5
CI	1	1	1
SVRI	0.5	1	2

**High threshold = threshold + number \* average + number \* std**

**Low threshold = -threshold - number \* average - number \* std**

If the number is not mentioned in the table it is not used in order to calculate the threshold.

Example:

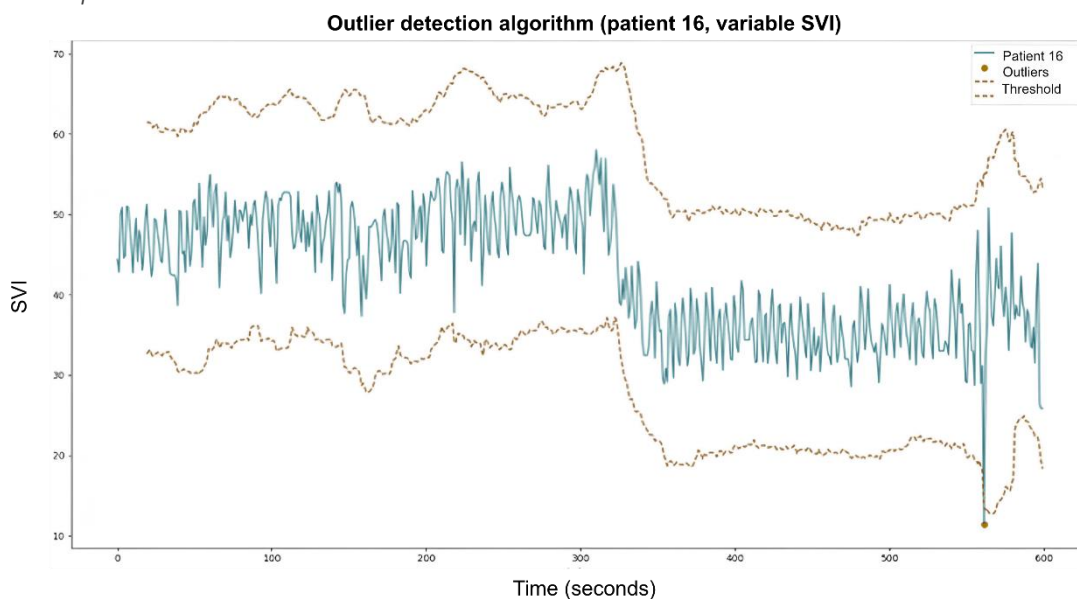


Figure 1. An example patient 16 for the Stroke volume index, where a threshold, the average and std is used to create the high and low threshold for outliers.

