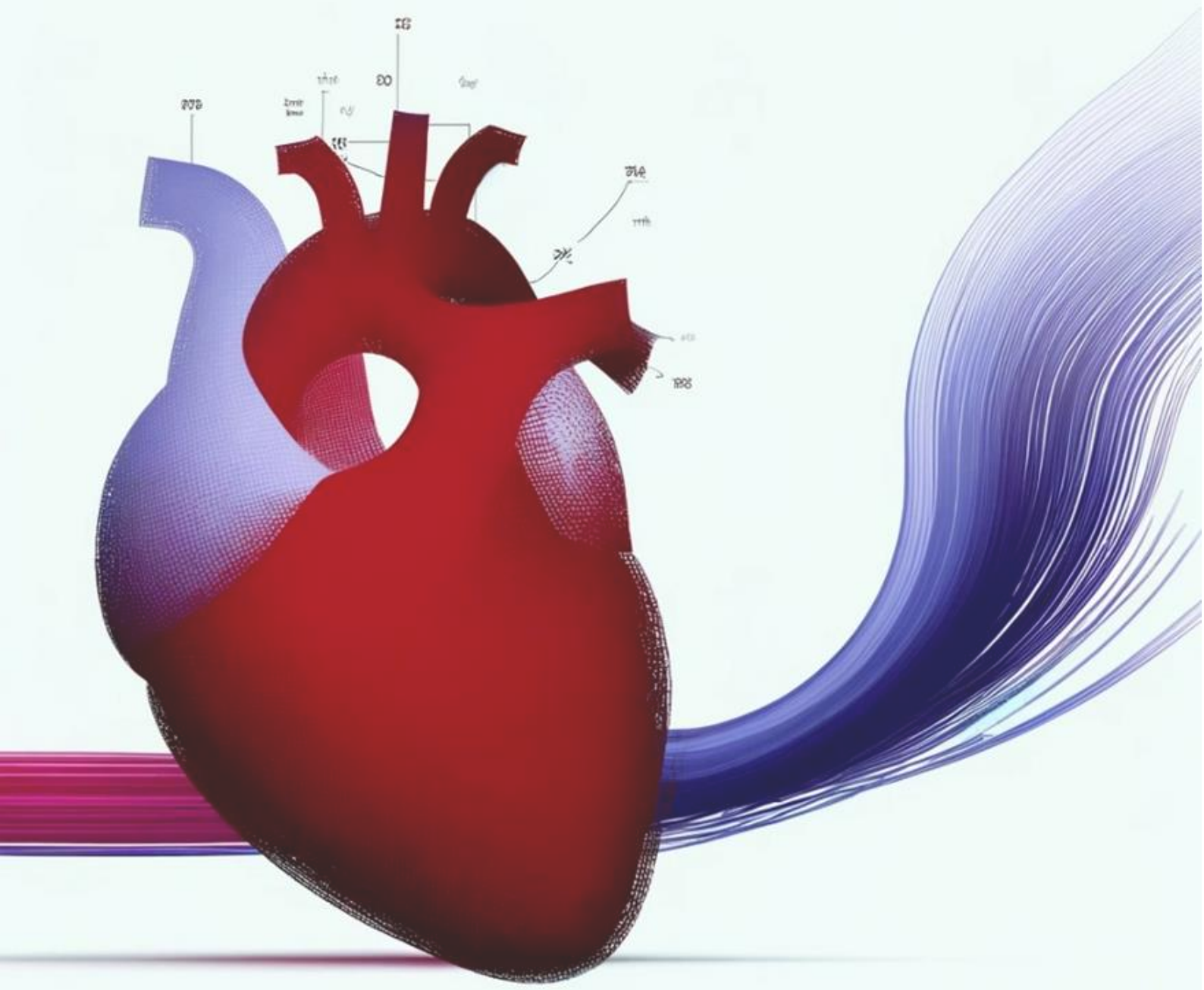
Proof-of-concept of personalized CIED-derived modeling for ambulatory heart failure monitoring

Master thesis

Annefleur Kluft



PROOF-OF-CONCEPT OF PERSONALIZED CIED-DERIVED MODELING FOR AMBULATORY HEART FAILURE MONITORING

Annefleur Kluft
Student number : 4474937
May 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)
Dept. of Biomechanical Engineering, TUDELFT
Cardiology, LUMC

Supervisor(s):
S.L.M.A. Beeres, MD, PhD
J. Dauwels, Ir, PhD
A.D. Egorova, MD, PhD

Thesis committee members:
A.D. Egorova, MD, PhD (chair)
J. Dauwels, Ir, PhD
M.C. Den Haan, MD, PhD

Preface

This report is the end product of my graduation internship at the Cardiology department of the Leiden University Medical Centre. With this report, I finish my time as a student at the Delft University of Technology.

The clinical internship year during my Master's was the first time I worked on the HeartLogic project. I was very excited about the value of the integration of the medical and technical field which is evident in this project. Moreover, the clinical aspect of Cardiology motivated me to extend my time as a student to study Medicine. After this first internship, I was convinced to return to the Cardiology department for my graduation internship. I would like to thank Anastasia Egorova and Saskia Beeres for their positive attitude and flexibility in making this happen and for welcoming me at the department. Furthermore, your enthusiasm for this project is contagious and motivating.

I am very grateful to state that during this project, I learned a lot about machine learning and improved my programming skills. Justin Dauwels, thank you for teaching me throughout this project and for your advice on important decisions.

I would like to express my gratitude to all the colleagues at the Cardiology department who made my time there so enjoyable. Also, thanks to Michelle Feijen, who helped with shaping this project in its early phase. Lastly, I would like to thank Melina den Haan for participating in my thesis committee.

I am proud to present to you the results of my graduation project.

Table of Contents

List of abbreviations

ACE-I	Angiotensin-converting enzyme inhibitor
AF	Atrial fibrillation
ARB	Angiotensin 2 receptor blocker
ARNI	Angiotensin receptor neprilysin inhibitor
AT	Atrial tachycardia
AUPRC	Area under precision-recall curve
AUROC	Area under receiver operating characteristic
BMI	Body Mass Index
CABG	Coronary artery bypass graft
CIED	Cardiac implantable electronic devices
COPD	Chronic obstructive pulmonary disease
CRT(-D)	Cardiac resynchronization therapy (with defibrillator)
CV	Cross-validation
CVA	Cerebral vascular accident
ICD	Implantable cardioverter defibrillator
IQR	Interquartile range
eGFR	estimated Glomerular filtration rate
HF	Heart failure
HRV	Heart rate variability
IV	Intravenous
ML	Machine learning
MRA	Mineral corticoid inhibitor
NT-proBNP	N-terminal pro B-type natriuretic peptide
LOOCV	Leave-one-out cross-validation
LV	Left ventricular
LVEF	Left ventricular ejection fraction
NYHA	New York Heart Association
NPV	Negative predictive value
PPV	Positive predictive value
PPY	Per patient-year
RBF	Radial basis function
SMOTE	Synthetic minority oversampling technique
S1	First heart sound
S3	Third heart sound
SD	Standard deviation
SHFM	Seattle Heart Failure Model
SVC	Support vector classifier
TIA	Transient ischemic attack
XGBoost	Extreme gradient boosting

Abstract

Introduction: Heart failure (HF) poses a significant burden on public health. This can be largely attributed to recurrent hospitalizations in consequence of HF decompensation. Detection of early signs of impending fluid retention may facilitate timely medical intervention and thereby prevent hospitalizations. Monitoring of Cardiac Implantable Electronic Devices (CIEDs)-derived parameters has been proposed as promising solution, as the sensor inherent in CIEDs provide the ability to continuously monitor physiological signals. The aim of this study was to develop personalized machine learning (ML) models that can identify upcoming HF decompensation based on CIED-derived parameters.

Methods: Two ML models, a support vector classifier (SVC) and an extreme gradient boosting (XGBoost) model, were developed for all patients. Features known to be associated to HF decompensation were extracted from daily CIED data. The output of the models is the daily classification of the patient's HF status, either 'stable' or 'unstable'. Model performance was evaluated through area under the precision-recall curve (AUPRC). First, the models were tested on a development dataset with leave-one-out cross-validation, and subsequently on an independent test set.

Results: In total, for 62 patients two models were developed. The average AUPRC on the independent test set of the XGBoost models was 0.63 ± 0.28 and of the SVC models was 0.57 ± 0.26. Finally, for each patient, the model that resulted in the highest AUPRC was selected. The final models achieved an AUPRC on the independent test set of 0.61 ± 0.28.

Conclusion: The findings of this study show promising results for the use of personalized CIED-derived models. However, significant variability in model performance across patients highlight the need for further research.

1

Introduction

Heart failure (HF) is a clinical syndrome that exerts a significant impact on public health. It is the result of a structural and/or functional abnormality of the heart (1). Today, HF affects more than 60 million people in the world, corresponding to a prevalence of 1-2% of adults (2). As society ages, this number is expected to increase. In the Netherlands, costs related to HF constitute about 0.5% of the total health care budget (3). This can be largely attributed to the (recurrent) hospitalizations in consequence of worsening HF, or 'decompensation' (4). Yearly, an estimated 13% of patients are hospitalized at least once (5). Moreover, (recurrent) hospitalizations are an indication of disease progression and are significantly associated with increased mortality (6).

For these reasons, prevention of recurrent hospitalizations is one of the major goals in the treatment of HF (1). To prevent hospitalizations, detection of impending fluid retention is key. Early detection could facilitate timely therapeutic adjustments and avoid hospitalizations. In the last decades, telemonitoring has gained interest as a promising solution to detect worsening HF at an early stage.

The first telemonitoring strategies consisted of assessments of body weight, heart rate and blood pressure at home and monitoring of symptoms through telephone contact. Meta-analyses and Randomized Controlled Trials (RCTs) have demonstrated small and heterogeneous, but significant reductions in all-cause mortality and HF hospitalizations with these strategies (7-9).

Simultaneously, Cardiac Implantable Electronic Devices (CIEDs) have been proven to be significant in the treatment of HF, to prevent sudden cardiac death and/or improve cardiac function (10, 11). The sensors inherent in these devices have facilitated the next step in telemonitoring. Sensing of electrical impedance, the RR interval and acceleration allows for monitoring of parameters such as thoracic impedance, respiratory rate, heart rate variability, night heart rate and physical activity. These parameters have all been proven to correlate with the pathophysiologic process of worsening HF (12-14). However, their individual predictive ability to predict an upcoming episode of worsening HF is limited (15-19).

Nevertheless, the limited success of single sensor monitoring strategies may be attributed to the inherent multifactorial and complex nature of the pathophysiology of worsening HF. This hypothesis together with the aforementioned results of single sensor studies have triggered the development of algorithms that make use of multi-sensor derived parameters as inputs. HeartLogic is one example of a multisensor algorithm, developed by Boston Scientific (Marlborough, MA, USA). The algorithm was developed and validated in the MultiSENSE trial in 2017 (20). The study reported a sensitivity of 70% and specificity of 87.5% for detection of worsening HF. Since then, HeartLogic's diagnostic performance to predict worsening HF has been successfully validated in multiple patient cohorts (21-25). Two studies have reported a positive effect of a HeartLogic guided care path on HF hospitalizations (22, 26). Yet, the question remains whether a standalone alert justifies therapeutic actions. The unexplained alert rate is reported between 0.16 per patient-year (PPY) and 1.47 PPY (20-24). Moreover, the mathematical methods used for development of HeartLogic are unknown to clinicians, essentially making HeartLogic a "black box" (27). The black box problem has been described as one of the phenomena setting back

the integration of machine learning (ML) models in the clinical practice (28). Moreover, HF is a complex and heterogeneous syndrome and its progression is of patient-specific nature (29).

Therefore, the main objective of this thesis was to explore the feasibility of a personalized CIED-derived modeling approach. Conceptually, such an approach would enhance interpretability of the algorithm's outcome and provide a patient-specific risk classification. To this aid, we developed and tested two different ML classifiers that use CIED-extracted features to discriminate between 'stable' HF and HF decompensation, which is referred to as 'unstable' HF. For this purpose, we utilized Extreme Gradient Boosting (XGBoost) and a Support Vector Classifier (SVC).

2

Background

2.1 Pathophysiology of worsening heart failure and its relation to CIED-derived parameters

Heart failure is a complex, chronic clinical syndrome punctuated by episodes of acute clinical deterioration, which further advance disease progression (30). These episodes of acute clinical deterioration, or decompensation, are associated with signs and symptoms of fluid retention, such as: (increased) dyspnea, orthopnea, weight gain, peripheral oedema, and fatigue (1).

Sensing of physiological signals by means of CIEDs has allowed for a unique insight into the pathophysiology of these episodes (13). A graphical presentation of the progression of stable HF in an euvolemic 'stable' state to decompensated HF is presented in figure 1. The main proposed pathophysiologic mechanism is that reduced contractility of the heart results in increased intracardiac filling pressures (31). Studies have identified that about three to four weeks before hospitalization, cardiac filling pressures increase (32). Changes in contractility and cardiac filling pressures are reflected by the first and third heart sound (S1 and S3), respectively. CIED-derived accelerometer data embedded in the pulse generator of the device allow for quantification of S1 and S3 (33). The accelerometer measures accelerations resulting from vibrations in the right ventricular wall through the right ventricular lead. These measurements correspond to the auscultated heart sounds. The accelerometer based S1 significantly correlates with the left ventricular pressure derivative (dP/dt), a measure of contractility (34). Moreover, increased S3 is indicative of elevated filling pressures (34). Specifically, the vibrations that are responsible for S3 are produced by rapid deceleration of the early diastolic transmitral flow caused by a stiff left ventricle (35, 36).

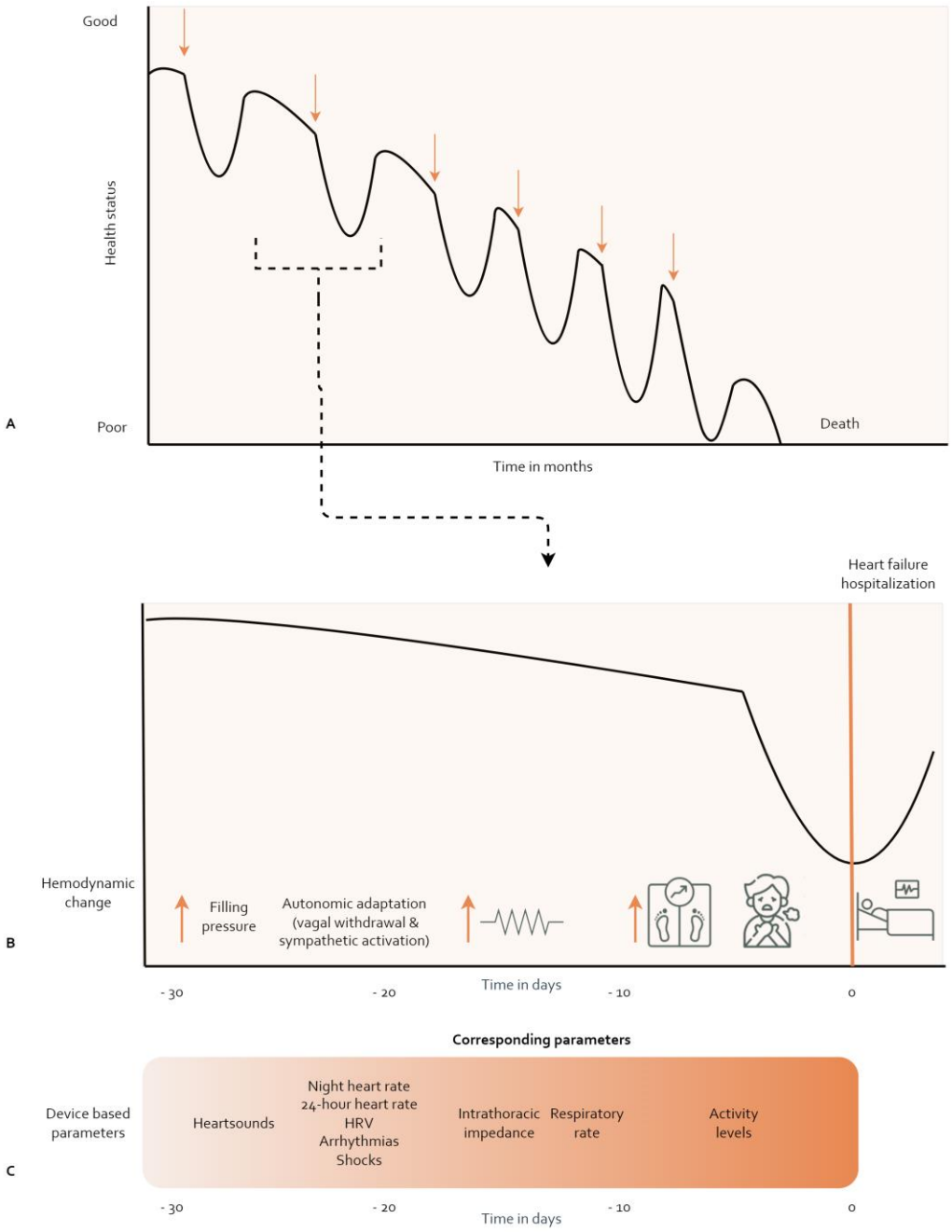
Persistently increased pressures trigger interstitial fluid accumulation. Moreover, reduced cardiac output leads to activation of neurohormonal and sympathetic responses, which exert stress on peripheral vasculature and further exacerbate fluid retention (31). Increased heart rate and decreased heart rate variability, as monitored by the CIED, are indicative of increased sympathetic control (12). In addition, fluid accumulation is reflected by measurements of thoracic impedance. Thoracic impedance is measured between the device case and the right ventricular lead (37). Since electrical current conducts more rapidly through fluid than air, thoracic impedance decreases in case of fluid accumulation. A change in thoracic impedance may be identified about two weeks before hospitalization (38).

Notably, symptoms of worsening HF can also be linked to parameters derived from CIEDs. For instance, patients who experience fatigue may decrease their activity levels. Moreover, an increased respiratory rate is associated with a rapid shallow breathing pattern, causing patients to experience dyspnea.

Lastly, atrial fibrillation (AF) is a frequent and clinically significant comorbidity of HF; about 25% of HF patients experience sustained AF (39). The proportion of HF patients with AF increases with HF progression and age. One mechanism of AF that contributes to worsening HF is the result of the lack of effective atrial contractions and irregularity in the timing of the diastole. This causes the left atrial pressure to increase, whereas blood pressure,

stroke volume and cardiac output decrease. Moreover, AF can directly contribute to left ventricular dysfunction through a mechanism known as tachycardia-induced cardiomyopathy.

Figure 1: A) *Schematic representation of the clinical course of chronic HF*, B) *Hemodynamic changes leading to an episode of decompensation*, C) *Cardiac Implantable Electronic Device-derived parameters corresponding to hemodynamic changes*.



HRV, heart rate variability

2.2 Technical background

Machine learning (ML) is the science within the field of Artificial Intelligence that provides systems with the ability to learn from data (40). It uses experiences, training data, to perform a task, aiming to optimize its performance in executing that task. Supervised learning is a type of ML in which the training data and the desired outcome are both fed to the system (41). The supervised learning algorithms compute relationships between the data and the desired outcome. Typically, the algorithms are trained for tasks of classification or regression.
After model training, ML models are evaluated on test data that the model has not seen yet. A robust ML model generalizes well to new observations. Conversely, ML models that perform well on the training data, but not on the test data, are overfitted. Such models follow the noise in the training data rather than detect the relevant patterns (41).

Extreme Gradient Boosting

Within supervised learning, ensembling is a method of combining predictors (40). Specifically, Extreme Gradient Boosting (XGBoost) is a decision tree ensemble (42). A decision tree is a model that classifies or predicts by learning decision rules based on data features (43). The XGBoost ensemble is created by sequentially constructing decision trees (44). To do so, it leverages an approach known as gradient boosting. Each new tree is fit to the residual errors made by the previous tree. Specifically, each subsequent tree is fit to minimize the objective function. The objective function is a composite of the loss function L and regularization terms Ω, which is given as in Equation 1 (45).

$$Objective(\theta) = L(\theta) + \Omega(\theta). \quad (1)$$

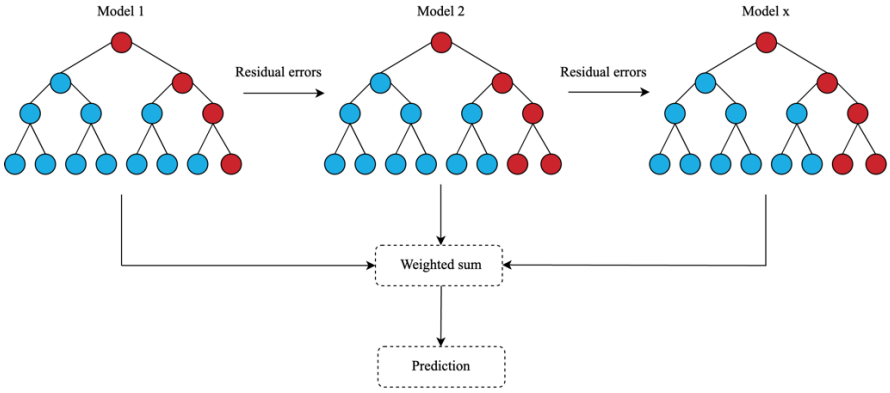
The regularization terms penalize model complexity and restrict the model's training process to avoid overfitting (46). The loss function indicates the difference between the prediction $\hat{y}_i$ and the true y_i. In binary classification problems, the loss function is often represented by the logistic loss (Equation 2) (45). For a single data point, a simplified version of the logistic loss can be described by Equation 3, in which p denotes the probability.

$$L(\theta) = \sum [y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})]. \quad (2)$$

$$L(y, p) = y\ log(p) + (1 - y) \log(1 - p). \quad (3)$$

In conclusion, the XGBoost ensemble gradually performs better as more trees are added, constrained by the regularization terms to create a robust model. A simplified version of the structure of the XGBoost model is presented in figure 2.

Figure 2: *Schematic presentation of Extreme Gradient Boosting.*



Hyperparameters are parameters of the learning algorithm that are set prior to model training. A subset of key hyperparameters of the XGBoost algorithm are depicted in table 1.

Table 1: *An overview of important hyperparameters for the XGBoost model (45, 46).*

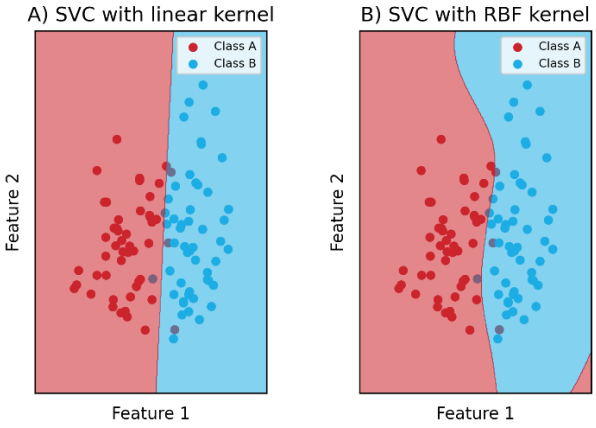
Parameter	**Explanation**	**The risk of overfitting as the parameter increases**
Learning rate	The value of the learning rate scales the effect of each newly added tree.	↑
Maximum depth	Specifies the maximum allowed depth for each tree.	↑
Minimum child weight	Minimum sum of Hessian values (second order derivative of the loss function) needed for a tree to split.	↓
Subsample	Fraction of the training set that is sampled to construct each tree.	↑
Colsample by tree	Proportion of features used to construct each tree.	↑
Number of estimators	Total number of trees in the model.	↑
Scale position weight	Controls the balance of weights in both classes. Primarily used in datasets with significant class imbalance.	Not applicable

XGBoost, Extreme Gradient Boosting

Support Vector Classifier

The Support Vector Classifier (SVC) is a popular machine learning method that classifies data by fitting a hyperplane to a multidimensional feature space to separate classes (40). The kernel function describes the transformation of the data to this multidimensional feature space (41). The variety in kernel functions enable the SVC to identify both linear and more complex classification patterns. The goal is to find the decision boundary with the largest possible margin between the different classes. Figure 3 illustrates the decision boundary with a linear and a radial basis function kernel.

Figure 3: *Examples of SVC hyperplanes with a linear and a Radial Basis Function (RBF) kernel. Both examples represent a two-class classification problem.* A) *The linear kernel fits a linear decision boundary to separate the classes*, B) *The RBF kernel uses a Gaussian (exponential) function that allows for a non-linear decision boundary.*



RBF, radial basis function; SVC, support vector classifier

The objective function determines the optimal parameters in the hyperplane to maximize the margin and is represented by Equation 4 (47). In this Equation, the first term describes minimizing the weight vector ω to maximize the margin. The second part addresses handling of misclassifications. C is the regularization parameter that determines the penalty of misclassifications and ξ_i are the slack variables represent the degree of misclassifications (48, 49). In general, a higher value of C penalizes misclassifications on the training data, allowing for a more complex decision boundary.

$$min_{\omega,b,\xi} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i . \quad (4)$$

The kernel is one of the key hyperparameters of the SVC. Table 2 provides an explanation of important hyperparameters for the SVC.

Table 2: *An overview of important hyperparameters for the SVC (43, 50).*

Parameter	**Explanation**	**The risk of overfitting as the parameter increases**
Kernel	Describes the transformation of the data to the multidimensional feature space.	Not applicable
C	Controls the trade-off between correct classification of training examples against maximizing the margin of the decision boundary.	↑
Gamma*	Controls the power of individual training examples to the decision boundary.	↑
Class weight	Controls the balance of weights in both classes. Primarily used in datasets with significant class imbalance.	Not applicable

SVC, Support Vector Classifier - *only applicable to non-linear kernels

Class imbalance

A known challenge of ML models is an imbalanced dataset. The class imbalance problem occurs when the number of examples in one class is significantly higher than the number of examples in the other class (51). In the general context of clinical diagnostics, 'healthy' instances often outnumber 'disease' instances. As a result, class imbalance is a common problem in ML for clinical applications (52). Standard ML models often assume a balanced distribution between classes. This means that in case of a two-class classification problem, a sample is classified as the positive class if the predicted probability exceeds 0.5. With an imbalanced dataset, this may introduce bias towards the majority class. Conversely, the model may underfit to the minority data. Additionally, performance metrics such as predictive accuracy may overestimate the model's performance. Multiple approaches have been proposed to deal with class imbalance. For example, oversampling the minority class, downsampling the majority class, adjusting misclassification costs and changing the decision threshold are amongst the strategies to tackle this problem (51).

Synthetic minority oversampling technique (SMOTE) is a tool that addresses class imbalance by oversampling the minority class (53). A simplified version of this technique is represented by Equation 5. The technique randomly selects samples x_i that serve as initial minority class samples. For each selected sample x_i in the minority class, it finds the *k* closest neighbors within the same class. Synthetic samples are created along the vector between the original sample x_i and one of the *k* closest neighbors $x_{neighbor}$. The random scalar c (value between 0 and 1) effectively determines the position of the new sample along the vector.

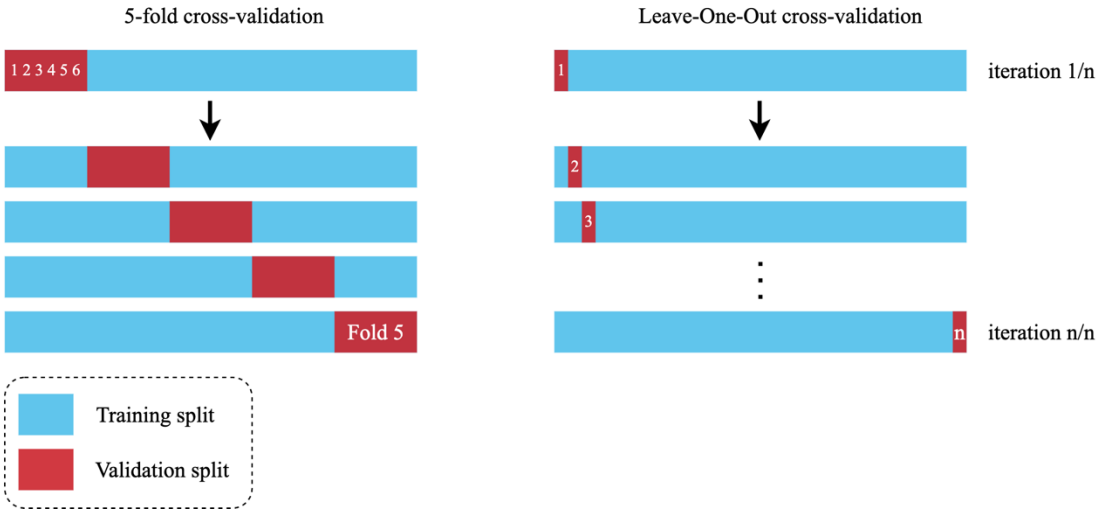
$$(x_{new}) = x_i + c * \left(x_{neighbor} - x_i\right). \quad (5)$$

Model development

In general, ML model development consists of a number of steps: data preparation, selection and finetuning of the model, training and performance evaluation.

Models are typically optimized, trained and evaluated using cross-validation (CV) (41). In *K*-fold CV, the dataset is randomly split in *K* number of subsets, or folds. The model is trained and evaluated *K* times, with a different fold for performance evaluation in each iteration and training on the other *K* - 1 folds. Leave-One-Out cross-validation is another cross-validation strategy that is often preferred for smaller or highly imbalanced datasets. Leave-one-out cross-validation uses each individual observation for evaluation, and the remaining observations for training. The performance of the model is obtained by averaging the performance across all validation observations. In general, the performance estimate obtained with leave-one-out cross-validation is subject to a degree of variance since each iteration uses almost all observations for training. In contrast, the performance estimate obtained with *K*-fold CV may exhibit some bias. A schematic presentation of both methods is provided in Figure 4.

Figure 4: *Schematic presentation of 5-fold cross-validation (CV) and Leave-One-Out cross-validation (LOOCV). The data is split into a training set (blue) and a validation set (red). In 5-fold CV, the data is split in 5 subsets. In each fold, one subset is used for validation and the remaining subsets for training. In LOOCV, each single observation is used for validation once, and all remaining observations are used for training.*



Bayesian optimization

An important step within the process of model development is hyperparameter tuning. During hyperparameter tuning, the performance of the model is optimized by evaluating the performance with different combinations of hyperparameters. Bayesian optimization is an efficient hyperparameter tuning method to iteratively propose new hyperparameter settings based on the performance of prior settings.

The key concept of Bayesian optimization is a sequential model-based approach that is updated throughout the process to drive optimization decisions (54, 55). The aim is to search for the combination of hyperparameters that minimizes the performance loss on the validation set. With each iteration, the algorithm constructs a probabilistic model to capture how hyperparameters affect performance loss. To this aid, often a Gaussian Process is exploited. The output is an estimate of the expected performance loss and the uncertainty of the prediction. Equipped with these outputs, the acquisition function is leveraged to guide the selection of a new combination of hyperparameters (56). These functions represent a trade-off between exploration (investigating unknown settings) and exploitation (focusing on settings with high predicted performance). The search space defines the range and scale of the hyperparameters that are optimized (57).

3

Methods

3.1 Study population and data collection

For this master thesis, we considered all HF patients under active follow-up in the HeartLogic cohort at the Leiden University Medical Center (LUMC) in October 2023. Patients who experienced at least one episode of decompensated HF since activation of HeartLogic were included.

Patients in the HeartLogic guided care path at the LUMC are followed-up according to a standardized protocol. In case of an alert, the HF care team is notified. The patient is contacted by phone within 72 hours for structural evaluation of early signs and symptoms of HF decompensation. For this purpose, a dedicated heart failure questionnaire is used that includes assessment of symptoms, signs, weight, blood pressure and heart rhythm (23). In case 2 or more criteria of HF decompensation are met, the alert is considered true positive. All patients with a true positive alert receive lifestyle advice to limit fluid and salt intake. Additional therapeutic course of action depends on the severity of symptoms and signs. In most cases, oral diuretic therapy is up-titrated for several days (23, 26). Patients without symptoms or signs of HF decompensation or any other suspected diagnoses are followed-up again after two, six and ten weeks. In case no symptoms or signs of HF decompensation arise during this period of follow-up, an alert is deemed false positive.

The follow-up protocol within the HeartLogic research consists of classification of each HeartLogic alert as true or false positive. Moreover, patients are offered the possibility to contact the HF nurses when they experience symptoms or signs of HF decompensation. An episode of HF decompensation without a preceding HeartLogic alert is considered false negative. For the purpose of this thesis, all alert follow-up data was made available.

3.2 Preprocessing and feature extraction

Boston Scientific provided the daily recorded CIED extracted sensor data and the corresponding HeartLogic index from each patient from the moment of inclusion in the HeartLogic cohort of the LUMC up to October 25, 2023.

Fourteen parameters with a known association to HF decompensation were extracted from the CIED sensor data. An overview of these parameters is presented in table 3. In the context of model development, these parameters are referred to as features. Moreover, accounting for the temporal nature of the classification problem, we included a set of temporal features in the models. For S1, S3, impedance, respiratory rate, night heart rate, 24-hour heart rate, heart rate variability (HRV) and activity the difference between a value on day X and the value 14 days prior was calculated and implemented as additional ‘delta’ features (Equation 6). Lastly, previous studies have demonstrated the significance of the rate at which cardiac filling pressures and thoracic impedance change in relation to HF decompensation (32, 58). Therefore, the rate of increase, i.e. the slope, of S3 and thoracic impedance during seven days was calculated and included as feature, as depicted in Equation 7. As a result, twenty-four features were selected from the CIED data for model development.

$$F_{day\,x} = F_{day\,x} - F_{day\,x-14}. \quad (6)$$

$$F_{day\,x} = \frac{F_{day\,x} - F_{day\,x-7}}{F_{day\,x-7}}. \quad (7)$$

Table 3: *An overview of the parameters extracted from Cardiac Implantable Electronic Devices (CIED), the methods used for measurement, the clinical relevance of the parameter in relation to worsening HF, and the direction of the change in case of worsening HF (27, 59, 60).*

Parameter	Measurement	Clinical relevance	Direction of change
Heart sounds	Both are derived from accelerometer		
First heart sound (S1)	data from accelerations through RV	Ventricular contractility	↓
Third heart sound (S3)	lead	Cardiac filling pressures	↑
Thoracic impedance	Impedance between RV lead and pulse generator	Fluid accumulation	↓
Respiratory rate	Respiratory rate, derived from impedance measurements	Dyspnea	↑
Heart rate		Sympathetic control	
Daily heart rate	Mean 24-hour heart rate		↑
Night heart rate	Mean heart rate between midnight and 6 a.m.		↓
Heart rate variability	SD of sinus-to-sinus intervals		
Atrial tachyarrhythmias		Known precipitating factor of	↑
Duration	Hours in ATR mode	HF decompensation	
Heart rate	Mean and maximum		
Count	Number of episodes		
Sleep incline	Angle between torso and the horizontal plane during sleep	Orthopnea	↑
%Ventricular pacing	Percent of beats paced through LV lead	Marker for correction of ventricular dyssynchrony	↓
Activity	Active hours per day, derived from relation between respiration and heart rate	Overall health status and fatigue	↓

AF, atrial fibrillation; AT, atrial tachycardia; ATR, atrial tachy response; LV, left ventricular; RMS, root mean square; RV, right ventricular; SD, standard deviation

3.2 Model development

The twenty-two extracted features serve as the input to the models. Furthermore, based on the alert follow-up data and additional data retrieved from hospital information systems, a daily output was generated that represents the patient's daily classification of HF status as either 'stable HF' or 'unstable HF'. In this context, 'stable HF' indicates that HeartLogic was not in-alert and the patient was not experiencing an episode of HF decompensation undetected by HeartLogic. Contrarily, 'unstable HF' indicates HeartLogic was in-alert, given the alert was assigned as true positive according to the HeartLogic protocol, or the patient experienced HF decompensation not detected by HeartLogic. As a result, in terms of model development, the model is trained to a two-class classification problem.

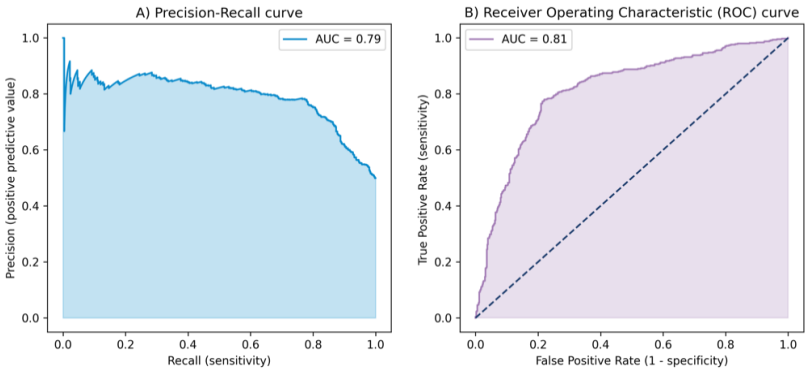
For each patient, two patient-specific classification models were developed. The classifiers considered for model development were XGBoost and SVC. The methods of these models have been described in paragraph 2.2.

Performance evaluation

The XGBoost model classifies a daily observation as 'unstable HF' if the predicted probability is higher than 0.5. Alternatively, the SVC assigns the observations to a class based on the decision boundary. If the observation lies on the 'unstable HF' side of the decision boundary, the observation is classified accordingly. The threshold at which an observation is classified to a class is referred to as the decision threshold. Effectively, this threshold is dynamic. For instance, in XGBoost the decision threshold can be altered so that it only classifies an observation as 'unstable HF' if the predicted probability is higher than 0.6. Similarly, the decision threshold for a SVC can be adjusted to classify an observation as 'unstable HF' only if its distance to the decision boundary exceeds a specified value.

The performance of the models was evaluated through the Area Under the Precision-Recall Curve (AUPRC). The Precision-Recall (PR) curve is a plot of recall, equivalent to sensitivity, against precision, equivalent to positive predictive value, across all decision thresholds. An example of a PR curve is provided in Figure 5. Therefore, the AUPRC provides a robust estimation of the overall model's performance, independent of the decision threshold. This performance metric is typically preferred in scenarios with significant class imbalance as it focuses on the performance of the model to classify the 'positive' (minority) class (40). As a result, a higher AUPRC typically indicates fewer false positives and false negatives.

Figure 5: *Examples of A) Precision-Recall curve, and B) Receiver Operating Characteristic curve.*



Anther metric that was obtained for performance evaluation is the area under the receiver operating characteristic curve (AUROC). The receiver operating characteristic (ROC) curve is a plot of false positive rate, equivalent to 1 – specificity, against sensitivity. Balanced accuracy is another useful performance metric in case of a class imbalanced dataset (61). Given that it represents both the proportion of true positives and proportion of true negatives, as depicted in Equation 8, both classes are considered equally important. Lastly, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were acquired, which were calculated according to Equations 9-12.

$$Balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right). \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN}. \tag{9}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{10}$$

$$Positive\ predictive\ value = \frac{TP}{TP + FP}. \tag{11}$$

$$Negative\ predictive\ value = \frac{TN}{FN + TN}. \tag{12}$$

Process of model training and testing

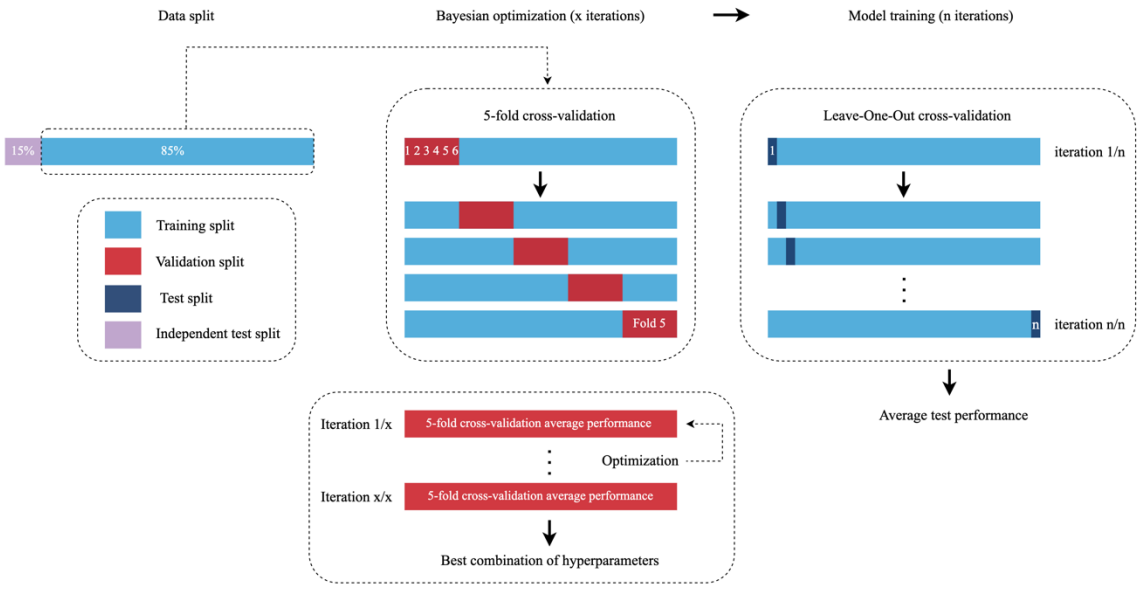
Figure 6 outlines the model development process. The first step is a stratified split of the data in a development dataset used for model training and testing (85%), and a dataset that represents an independent test set (15%). The independent test set contains data that the model does not see during the training process.

Subsequently, the development dataset is used for Bayesian optimization in a 5-fold cross-validation. For each combination of hyperparameters, the data is split in five subsets. Each subset serves as a validation set once. The model is fitted to the remaining four subsets. The performance of a given set of hyperparameters is defined as the average AUPRC of the model on the validation sets. During optimization, the number of iterations indicates the number of hyperparameter combinations that is evaluated. XGBoost optimization was conducted with 50 iterations. Since training a SVC can be computationally expensive, optimization of the SVC was conducted with 10 iterations. The search spaces defined for optimization of the model specific hyperparameters are presented in Table S1 and Table S2 in the Supplementary Material.

The hyperparameters that performed best on the validation sets were implemented in the models. Next, for model training and testing, leave-one-out cross-validation was performed. In each iteration, a single observation was used for testing. The remaining data made up the training set that the model was fitted to. The predictions made on the test observations were used for performance evaluation. Of note, when evaluating performance through leave-one-out cross-validation, the combined test observations are referred to as the development test set.

Finally, both models were fit once more to the entire development set. The fitted models were then applied to the independent test set to retrieve the performance of the model to unseen data. Ultimately, for each patient, the model that resulted in the best performance on the independent test set was selected.

Figure 6: *Schematic presentation of the model development process.*



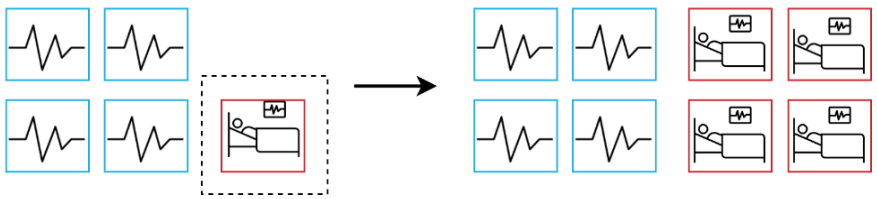
During optimization and model training, the training data is fed into a transformation pipeline. The pipeline is depicted in Figure 7. First, we addressed any missing values in the data. For a given feature, all missing values in the data were imputed with the median value of that feature, derived from the training data. In case all values from a feature were missing, these were all set to 0. Regarding missing data, it is key to understand why data is missing. Missing data may be due to random occurrences, specific causes or structural deficiencies in the data (62). To address this challenge, information was gathered on the nature of the missing data. For each patient, the proportion of missing values was retrieved and potential causes for missingness were explored.

Figure 7: *(Training) data transformation pipeline. Successively, we performed: missing data imputation, minority class oversampling and feature scaling.*

SMOTE, synthetic minority oversampling technique

Hereafter, the SMOTE algorithm was applied to the training data to oversample the minority class to achieve a 1:1 class distribution (Figure 8). As a result, the fitted model is less subjected to bias towards the majority class.

Figure 8: *Simplified schematic presentation of synthetic minority oversampling technique. The class imbalance problem (3:1) on the left is handled by oversampling of the minority class to a class distribution of 1:1.*



The final step in the transformation pipeline prior to model training is feature scaling. Numerical features that are on different scales create potential issues for ML models. Most ML models do not perform well when the features are on different scales. Feature scaling transforms the scale of the features to a uniform range. With regards to the models developed in this study, tree-based models like XGBoost do not inherently require feature scaling. Contrarily, a SVC does require feature scaling as it relies on distance-based calculations (40, 62). For consistency, feature scaling was applied to both models. Within the transformation pipeline, two different approaches for feature scaling were studied: standardization and robust scaling. Standardization centers each feature around the mean; each feature has a mean of 0 and a standard deviation (SD) of 1. Similarly, robust scaling centers features around the median. In a dataset with a significant amount of outliers, robust scaling often performs better (43). Within each iteration of the 5-fold cross-validation and the leave-one-out cross-validation, the scaler was fitted to the training data and then subsequently applied to the test data.

$$X_i' = \frac{X_i - mean}{SD} . \tag{12}$$

$$X_i' = \frac{X_i - median}{IQR} . \tag{13}$$

The parameters involved in the data transformation pipeline were treated as hyperparameters. These hyperparameters are applicable to both the XGBoost model and the SVC. The defined search space for the parameters is shown in table 4. The number of neighbors in the SMOTE algorithm and the type of scaler are optimized within the Bayesian optimization process. Of note, in case the number of neighbors *k* exceeded the number of samples *N* within the minority class, the number of neighbors was automatically set to $N - 2$.

Table 4: *Hyperparameters involved in the data processing pipeline with the search space defined for the Bayesian optimization process.*

Hyperparameter	**Search space**
Number of neighbors in SMOTE	3, 4 or 5
Scaler	Standard scaler (Equation 12) or robust scaler (Equation 13)

SMOTE, synthetic minority oversampling technique

Feature importance

To gather more information on what drives the model's predictions, it is valuable to determine the contribution of individual features to the predictive performance of the model. To this end, XGBoost models offer a straightforward attribute that quantifies the importance of each feature. In this study, the feature importance is defined as the gain; the increase in AUPRC brought by a given feature. All feature gains in a model sum up to 1. Therefore, if all features (n = 24) contributed equally to the model's predictions, the gain of each feature would be approximately 0.04. To provide a conclusion on the feature significance throughout all XGBoost models, the average feature contributions were computed across all models with an AUPRC > 0.65.

Finally, to integrate the technical and clinical aspects of this study, one case is discussed to illustrate the development of the XGBoost model, the PR curve and the interpretation of the feature contributions.

3.3 Software and statistical analysis

Model development and statistical analyses were conducted in Python 3.11.6 with the following packages: Imblearn 0.11.0, Matplotlib 3.8.2, NumPy 1.26.3, Pandas 2.1.2, Pyreadstat 1.2.6, Scikit-learn 1.3.2, Scikit-optimize 0.10.1, Statsmodels 0.14.1 and Xgboost 2.0.3.

Baseline demographic and clinical data were retrieved from the hospital patient information systems (EPD-Vision and HiX). Collected data included age, gender, type of CIED, etiology of heart disease, left ventricular ejection fraction, comorbidities, cardiac history, medication and New York Heart Association (NYHA) class. Normally distributed descriptive data are reported as mean ± SD and non-normally distributed data as median ± interquartile range (IQR). Normality testing was performed with a Shapiro-Wilk test. Model performance was compared with Wilcoxon Signed Rank test. Correlation coefficients were calculated using Pearson's r and Spearman's ρ. For all analyses, a P-value of ≤ 0.05 was considered statistically significant.

4

Results

4.1 Patient population

At the moment of inclusion (October 2023), 166 patients were under active follow-up at the LUMC. Boston Scientific provided data of 126 of these patients. In total, 62 of these patients experienced at least one episode of HF decompensation and were therefore included in the study. Table 5 summarizes the baseline patient characteristics. Overall, median age was 69 (61-77) years, 73% of patients were male and the median left ventricular ejection fraction was 34%. The majority of patients had ischemic cardiomyopathy (56%) and NYHA class 1 (44%) or class 2 (26%) HF at the moment of inclusion in the HeartLogic cohort. Moreover, 37 (60%) patients had a cardiac resynchronization therapy with defibrillator (CRT-D) device and 25 (40%) an implantable cardioverter-defibrillator (ICD). At baseline, the prevalence of AF was 45% and of hypertension was 44%.

Table 5: *Baseline patient characteristics (n=62).*

Demographics (n=62)	
Age, median [IQR]	69 [61-77]
Male, n (%)	45 (73)
Years since HF diagnosis, median [IQR]	15 [6-19]
BMI, median [IQR]	27 [24-30]
LVEF, median [IQR]	34 [28-42]
NYHA-class	
- Class 1, n (%)	27 (44)
- Class 2, n (%)	16 (26)
- Class 3, n (%)	11 (18)
- Class 4, n (%)	8 (13)
Etiology of HF	
- Ischemic, n (%)	35 (56)
- Non-ischemic, n (%)	21 (34)
- Congenital heart disease, n (%)	6 (10)
Laboratory values	
eGFR, median [IQR]	71 [53-83]
NT-ProBNP, median [IQR]	541 [42-1413]
Device	
CRT-D, n (%)	37 (60)
- Percentage biventricular pacing, mean ± SD	84 ± 30
ICD, n (%)	25 (40)
Cardiac history	
Valve surgery, n (%)	12 (19)
CABG, n (%)	10 (16)

Comorbidities	
Atrial fibrillation, n (%)	28 (45)
- Paroxysmal, n (%)	24 (39)
- Permanent/long persistent, n (%)	4 (6)
Hypertension, n (%)	27 (44)
COPD, n (%)	3 (5)
Diabetes Mellitus, n (%)	8 (13)
Ischemic CVA/TIA, n (%)	7 (11)
Medical therapy	
Beta-blocker, n (%)	58 (94)
ACE-I/ARB/ARNI, n (%)	61 (98)
- ACE-I, n (%)	29 (47)
- ARB, n (%)	14 (23)
- ARNI, n (%)	18 (29)
MRA, n (%)	32 (52)
Diuretics, n (%)	43 (69)
Ivabradine, n (%)	2 (3)
Digoxin, n (%)	2 (3)

ACE-I, angiotensin-converting enzyme inhibitor; ARB, angiotensin 2 receptor blocker; ARNI, angiotensin receptor neprilysin inhibitor; BMI, Body Mass Index; CABG, coronary artery bypass graft; COPD, chronic obstructive pulmonary disease, CRT, cardiac resynchronization therapy; CVA, cerebral vascular accident; eGFR, estimated glomerular filtration rate; ICD, implantable cardioverter defibrillator; IQR, interquartile range; LVEF, left ventricular ejection fraction; MRA, mineral corticoid inhibitor; NYHA, New York Heart Association; NT-ProBNP, n-terminal pro B-type natriuretic peptide; TIA, transient ischemic attack.

4.2 Alert follow-up and characteristics

Median follow-up duration was 35 months (IQR: 23-49). In total, 187 patient-years of sensor data were collected. Throughout all patients, 153 episodes of HF decompensation occurred during follow-up. This corresponds to 0.82 episodes per patient-year (PPY). Of those, 142 were true positive HeartLogic alerts and 11 were false negative episodes of HF decompensation, not detected by HeartLogic. Most patients (n = 33, 53%) experienced one episode during follow-up. Moreover, 10 patients (15%) experienced 2 episodes, 7 patients (11%) 3 episodes, 4 patients (6%) 4 episodes and 2 patients (3%) 5 episodes. Lastly, 5 patients (13%) experienced 6 episodes or more. One patient experienced a maximum of 13 episodes of HF decompensation. The median duration of an episode was 31 days (IQR: 20-49 days).

Figure S1 and Figure S2 in the Supplementary Material provide a complete overview of the follow-up of all episodes. In summary, of 153 episodes, most (n = 125, 82%) could be managed by the HF care team in the outpatient setting. In most cases (n = 115, 75%), oral diuretic therapy was escalated in response to an episode. In 16 cases (10%), oral HF medication was optimized, either alone or in addition to diuretics. In some cases, these actions were insufficient to recompensate the patient or another form of therapy was needed. These episodes resulted in hospitalizations for HF without intravenous (IV) diuretic therapy (n = 5, 3%), administration of IV diuretics (n = 5, 5%), electrical cardioversion (n = 9, 6%) or an unscheduled outpatient clinic visit (n = 6, 4%).

4.3 Data analysis

Class distribution

The median follow-up of 35 months corresponds to a median number of 1064 days (IQR: 696-1477), or in terms of model development, observations. The median number of observations labeled as 'unstable HF' is 61 (IQR: 28-117), representing 7.2% (IQR: 2.3-13%) of the data. Table S3 in the Supplementary Material outlines the class distribution across all patients.

Missing data

To address the nature of missing data, the proportion of missing samples of each feature was computed for all patients. Of significance, LV pacing, HRV and sleep incline were not available for a number of patients. Specifically, of 21 patients with a CRT-D device, no LV pacing data were available. HRV and sleep incline data were missing in 21 and 36 patients, respectively. As a result, all values were set to 0. With regards to LV pacing, EPD-Vision was accessed to rule out potential issues relevant to the classification of the patient's HF status. Computation of sleep incline requires knowledge of a sleep period specified by the patient. This could be the reason of missing sleep incline. Finally, it was concluded that the missingness was not related to the outcome of the models; the classification of a patient's HF status. All remaining missing feature values were imputed with the median value computed on the training data.

4.4 Hyperparameter optimization

The Bayesian optimization process provided the hyperparameters for both models for all patients. Thereafter hyperparameters were fixed, as presented in Table S4 and Table S5.

4.5 Performance evaluation

After the hyperparameters were fixed, the models were trained and tested through leave-one-out cross-validation on the development dataset and, finally, tested on the independent test set. For all individual models the specified performance metrics were computed. Table 6 compares the average performance of the SVC and the XGBoost models. First of all, the XGBoost models achieved higher AUPRCs in the independent test set, 0.63 ± 0.28 versus 0.57 ± 0.26 ($p < 0.01$). Observed PPV (or precision) of the XGBoost models in the independent test set was 0.57 ± 0.27 and observed sensitivity was 0.65 ± 0.27. Overall, the models exhibit strong AUROC (0.90 ± 0.11 versus 0.90 ± 11, $p = 0.097$), specificity (0.80 ± 0.25 versus 0.94 ± 0.071, $p < 0.01$) and NPV (0.92 ± 0.22 versus 0.94 ± 0.071, $p = 0.90$), with XGBoost models slightly outperforming the SVC models.

Table 6: *Average performance of the SVC and XGBoost models, obtained with leave-one-out cross-validation and the independent test set.*

Performance metric	SVC LOOCV	SVC independent test	XGB LOOCV	XGB independent test	p-value LOOCV	p-value independent test
AUPRC	0.50 ± 0.25	0.57 ± 0.26	0.60 ± 0.25	0.63 ± 0.28	$6.9*10^{-8}$	$1.5*10^{-3}$
AUROC	0.87 ± 0.12	0.90 ± 0.11	0.91 ± 0.076	0.90 ± 0.13	$3.9*10^{-6}$	0.097
Balanced accuracy	0.76 ± 0.12	0.78 ± 0.16	0.79 ± 0.10	0.80 ± 0.13	0.13	0.36
Sensitivity (recall)	0.75 ± 0.24	0.76 ± 0.29	0.64 ± 0.22	0.65 ± 0.27	$1.9*10^{-3}$	$6.1*10^{-3}$
Specificity	0.78 ± 0.25	0.80 ± 0.25	0.93 ± 0.067	0.94 ± 0.071	$5.7*10^{-8}$	$3.3*10^{-8}$
PPV (precision)	0.35 ± 0.23	0.41 ± 0.28	0.52 ± 0.20	0.57 ± 0.27	$1.3*10^{-8}$	$1.3*10^{-5}$
NPV	0.92 ± 0.21	0.92 ± 0.22	0.97 ± 0.032	0.96 ± 0.041	0.27	0.90

Performance metrics are presented as mean ± standard deviation.

AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristic; LOOCV, leave-one-out cross-validation; SVC, Support Vector Classifier; XGBoost, Extreme Gradient Boosting

For all individual patients, one model was selected; the model that resulted in the highest AUPRC when applied to the independent test set. In total, the XGBoost model was selected for 51 patients and the SVC for 11 patients. The performance outcomes of these models are detailed in Table 7. On average, the models demonstrated a moderate AUPRC of 0.61 ± 0.25 on the development test set and 0.61 ± 0.28 on the independent test set. The average balanced accuracy was 0.80 ± 0.094 on the development test set and 0.79 ± 0.13 in the independent test set. Lastly, of significance, the performance metrics obtained with the leave-one-out cross-validation and the independent test set are closely aligned.

Table 7: *Average performance of final models: 52 XGBoost models and 11 SVC models. Performance metrics were obtained from predictions on the development test set with leave-one-out cross-validation and on the independent test set.*

Performance metric	LOOCV	Independent test
AUPRC	0.61 ± 0.25	0.61 ± 0.28
AUROC	0.92 ± 0.064	0.91 ± 0.090
Balanced accuracy	0.80 ± 0.094	0.79 ± 0.13
Sensitivity (recall)	0.68 ± 0.19	0.66 ± 0.26
Specificity	0.92 ± 0.13	0.92 ± 0.14
PPV (precision)	0.52 ± 0.21	0.57 ± 0.28
NPV	0.97 ± 0.033	0.96 ± 0.044

Performance metrics are presented as mean ± standard deviation.

AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristic; LOOCV, leave-one-out cross-validation; SVC, Support Vector Classifier; XGBoost, Extreme Gradient Boosting

The standard deviations in the observed AUPRCs indicate a substantial variation in performance throughout the different models. The variability in performance may be attributed to the class imbalance in the datasets. To explore this, the relations between the number and proportion of samples in the 'unstable HF' class and the

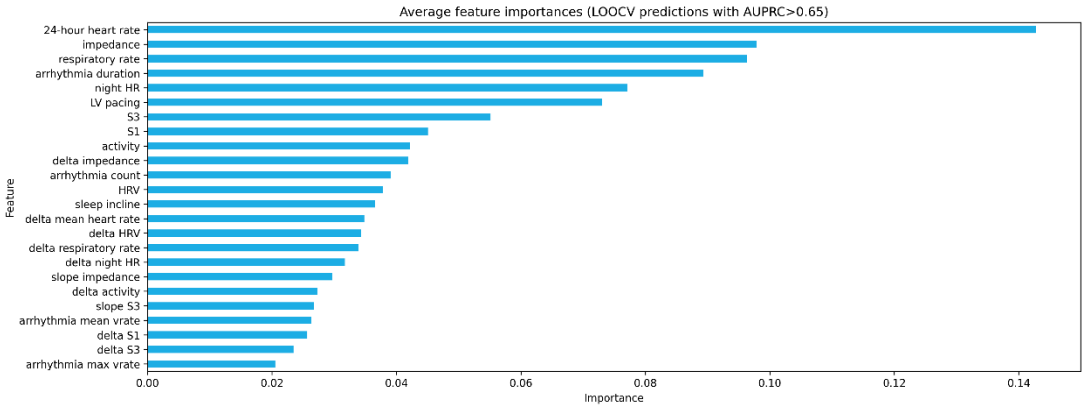
AUPRC were studied, as illustrated in Figure S3. Both the number and proportion of samples in the 'unstable HF' class were positively correlated to the AUPRC. The proportion of samples in the 'unstable HF' class exhibited a moderate but significant correlation to the AUPRC, with Pearson $r = 0.52$, with $p = 1.8*10^{-5}$ and Spearman $\rho = 0.55$, with $p = 3.1*10^{-6}$.

4.6 Feature importance

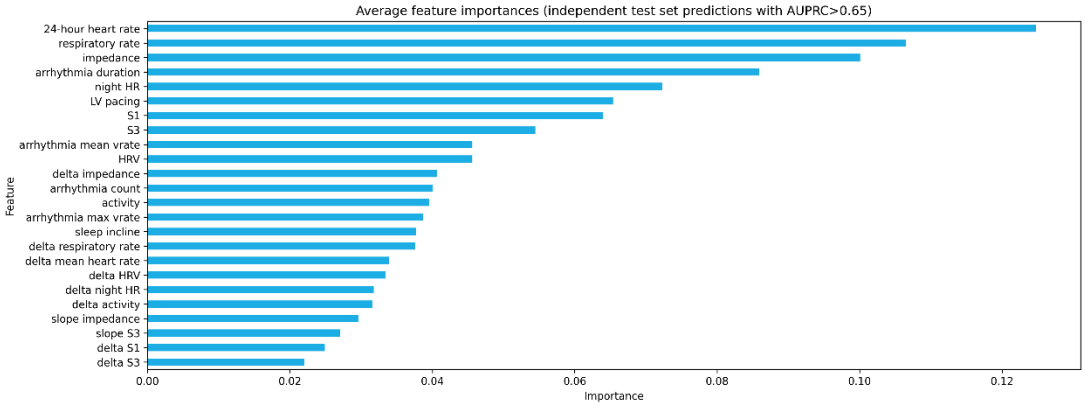
Of all XGBoost models, the average feature contributions were derived, represented by the gain of each feature. Importantly, the average gain was computed only from the subset of models in which the feature was available. Figure 9 presents the average feature contributions of all XGBoost models that resulted in a AUPRC > 0.65. In descending order, the five most significant features in the independent test set predictions were 24-hour heart rate, respiratory rate, impedance, arrhythmia duration and night heart rate.

Figure 9: *Bar plot indicating the average feature contributions in the XGBoost models that resulted in a AUPRC >0.65. Each bar represents a different feature. The length of the bar represents the gain, which is defined as the relative increase in AUPRC brought by a given feature. The sum of all feature gains in a model equals 1. The mean was calculated from the subset of models in which the feature was available.*

A) *Average feature importance from the leave-one-out cross-validation predictions (n = 29 models).*



B) *Average feature importance from the independent test set predictions (n = 34 models).*



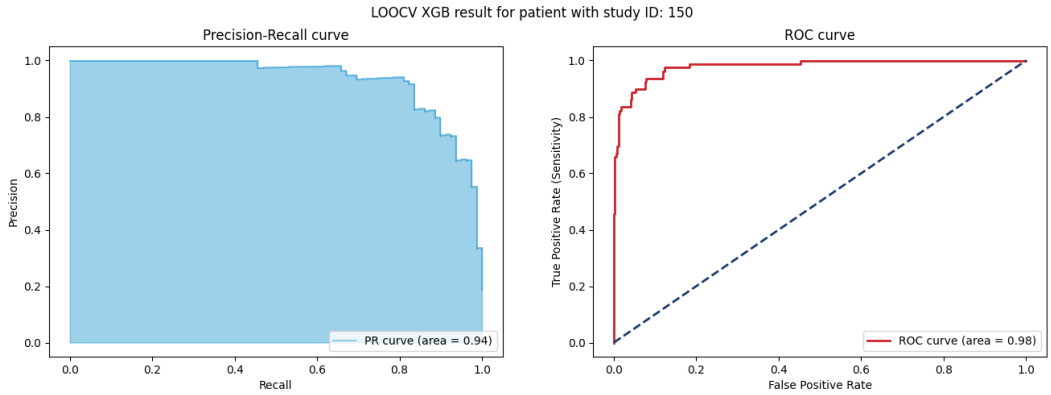
AUPRC, area under precision-recall curve; HR, heart rate, HRV, heart rate variability; LV, left ventricular; S1, first heart sound; S3, third heart sound; XGBoost, Extreme Gradient Boosting

Case: patient 150

Patient 150 experienced one episode of HF decompensation. During this episode, the patient presented with signs and symptoms of HF decompensation, for which oral diuretic therapy was enhanced. Since the symptoms did not resolve, the patient was clinically admitted and received IV diuretic therapy. Thereafter, the patient was discharged. The total episode lasted 93 days. Therefore, for the purpose of model development, 93 observations of 'unstable HF' were at hand. The model development dataset consisted of 423 observations of 'stable HF' and 79 observations of 'unstable HF'. The remainder of 'unstable HF' observations were set aside for the independent test set. The SMOTE algorithm was applied to upsample the minority class to achieve an equal class distribution, resulting in 423 observations in both classes.

Figure 10 presents the PR and ROC curves of the XGBoost model retrieved from the leave-one-out cross-validation. Accordingly, the AUPRC was 0.94 and AUROC was 0.98 on the development test set. On the independent test set, the AUPRC was 0.88 and AUROC 0.97 (Table S9).

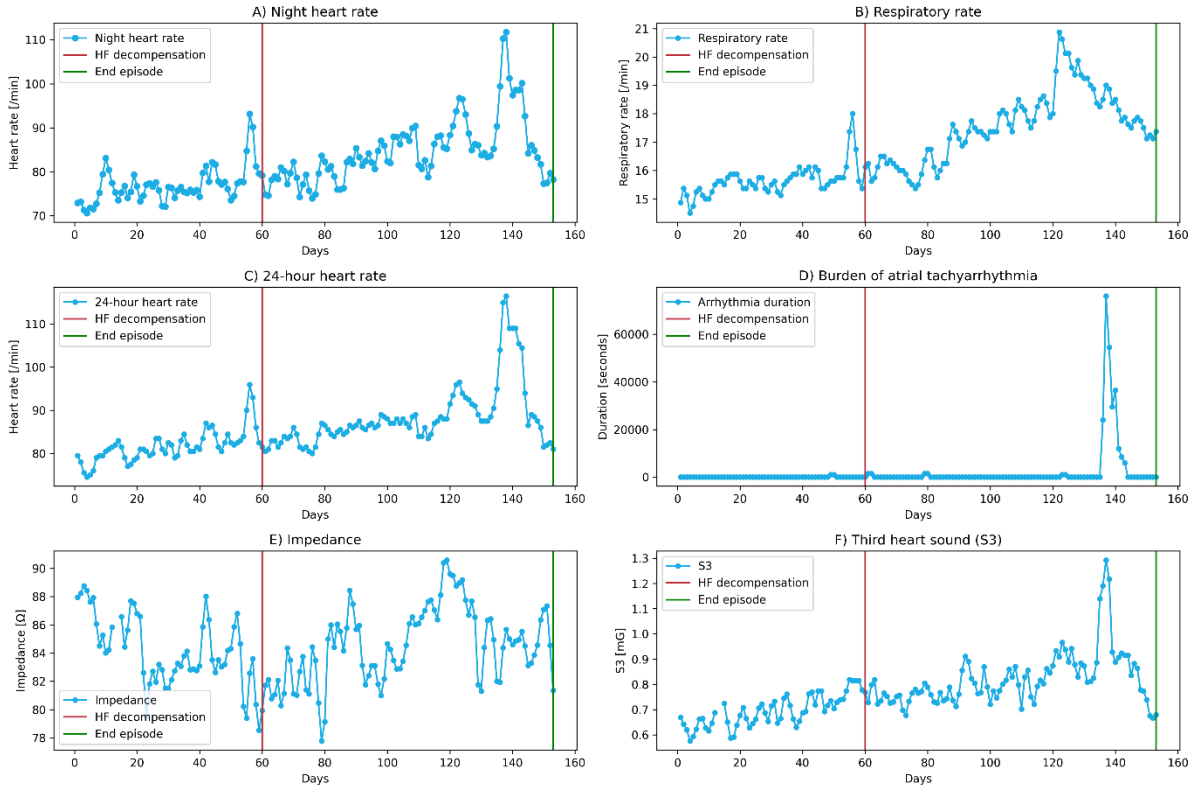
Figure 10: *Precision-recall curve for patient 150*



LOOCV, leave-one-out cross-validation; PR, precision-recall; XGB, Extreme Gradient Boosting

The XGBoost model's average feature gains revealed the most important features. Features with a gain above 0.04 were, provided in descending order: night heart rate, respiratory rate, 24-hour heart rate, arrhythmia duration, impedance, delta impedance and S3. Figure 11 illustrates the trends of the key features from the 60 days preceding the episode to the end of the episode.

Figure 11: *Trend of features from 60 days prior to the episode of HF decompensation (day 0) to the end of the episode. Figure A, B and C show an increase in heart rate and respiratory rate, related to increased sympathetic drive. Moreover, increased respiratory rate is associated with dyspnea. Figure D) indicates that the patient developed an atrial tachyarrhythmia during the episode. Figure E) depicts fluctuating impedance values. The general trend shows an initial decrease which is followed by an increase. Possibly, impedance stabilized after diuretic therapy was enhanced. Figure F) illustrates a gradual increase in S3, indicative of elevated filling pressures.*



5

Discussion

5.1 Discussion of results

The present study was a proof-of-concept for the development of personalized machine learning models for the timely identification of HF decompensation. To this aid, two ML classification models were developed for all patients, an XGBoost classifier and a SVC. The average performance of both models was moderate, with AUPRC on the development test set of 0.50 ± 0.25 for the SVC and 0.60 ± 0.25 for the XGBoost ($p = 6.9*10^{-8}$). Similar results were obtained on the independent test set, 0.57 ± 0.26 for the SVC and 0.63 ± 0.28 for the XGBoost ($p = 1.5*10^{-3}$). Whereas the SVC demonstrated higher sensitivity (or recall) on the independent test sets than the XGBoost (SVC: 0.76 ± 0.29, XGB: 0.65 ± 0.27, $p = 6.1*10^{-3}$), the PPV (or precision) of the SVC was significantly lower (SVC: 0.41 ± 0.28, XGB: 0.57 ± 0.27, $p = 1.3*10^{-5}$). Consequently, in general, the XGBoost provides a better precision-recall trade-off, and therefore a more robust classification. After selecting the best model for each patient, the average AUPRC increased slightly to 0.61 ± 0.25 on the development test set and 0.61 ± 0.28 on the independent test set. On average, the models provide high AUROC (0.91 ± 0.090) and a good balanced accuracy (0.79 ± 0.13) on the independent test set. However, a sensitivity of 0.66 ± 0.26 and PPV of 0.57 ± 0.28 demonstrate the limitations of the models to accurately identify ‘unstable HF’ and suggest room for improvement.

The highly comparable results on the development test set and the independent test set suggest that the models were not overfit to the training data. However, the majority of patients ($n = 33$, 53%) experienced one episode of HF decompensation throughout the entire follow-up period. As a consequence, the model is always somewhat overfit to detect that episode of HF decompensation. Therefore, it could be of interest to study the models developed for patients who experienced multiple episodes during follow-up. Specifically, training the model on a single episode and subsequently testing its ability to detect subsequent episodes could be a feasibly strategy. A disadvantage of this approach is that it will exacerbate the class imbalance in the datasets.

Notably, the performance metrics demonstrate a significant variability in the performance of the individual models. Evidently, for some patients the models have very low predictive ability. It is key to understand why the performance of the models demonstrates a substantial variation across different patients. Several potential causes may be identified. Firstly, a significant class imbalance may limit the model’s predictive ability. SMOTE was applied to help mitigate part of the class imbalance problem. Nonetheless, while synthetically created observations are assumed to belong to the minority class, a degree of uncertainty remains whether a synthetic observation is truly a minority class sample (63). To explore the relation between the class imbalance and model performance, the correlation coefficients between the number of ‘unstable HF’ observations, the proportion of ‘unstable HF’ observations and the AUPRCs were computed. Moderate correlation coefficients were found for both the number of ‘unstable HF’ observations (Spearman $\rho = 0.50$, $p = 3.3*10^{-5}$) and for the proportion of ‘unstable HF’

observations (Spearman ρ = 0.55, p = 3.1*10^{-5}). These results suggest that at least part of the variation in model performance may be attributed to the class imbalance problem. Secondly, the number of episodes that a patient experienced could account for part of the observed variability. Thirdly, a reason for varying model performance could be the complex etiology of HF decompensation, which is not captured by the features in all patients. To provide a concise answer to this question, the models that perform poorly should be further studied in detail.

For model development, a total of twenty-four features were selected. The individual feature contributions (gains) to the XGBoost predictions were highest for 24-hour heart rate, respiratory rate, impedance, arrhythmia duration and night heart rate. However, it must be noted that the mean was derived only from the subset of models in which the given feature was available. As a result, models that did not use arrhythmia duration for their predictions, since no AF or atrial tachycardia (AT) occurred, were excluded from the feature ranking. Moreover, data on LV pacing was available for only 16 CRT-D patients. Therefore, the identified significant contribution of arrhythmia duration and LV pacing to the predictions applies only to a specific subset of patients. In addition to the aforementioned features, other significant features with an average contribution > 0.04 to the test set predictions were S1, S3, number of AF/AT episodes, mean ventricular rate during AF/AT, HRV and 'delta' impedance. Interestingly, on average, the derived 'delta' and 'slope' features have relatively low individual feature contributions. This suggests that these features may not introduce significant new information. These features were included in the models to capture the temporal nature of the classification. In future studies, it could be advantageous to investigate additional time-domain features (62, 64).

5.2 Future research

The models developed in this study provide a daily classification of the patient's HF status, as 'unstable' or 'stable'. When considering further improvements of the models, post-optimization techniques could focus on translating the output to a metric more aligned with the clinical practice. For instance, HeartLogic is an alert-based algorithm that computes an index value, issuing an alert when the index surpasses the threshold of 16. Conversely, Medtronic (Minneapolis, Minnesota, USA) devices are employed with the TriageHF algorithm, which generates a monthly risk status (65, 66). This output represents the probability of HF hospitalization within the next 30 days, and is translated to a low, medium or high risk status. Effectively, the aim of post-optimization in this context is to design conditions that should trigger an alert.

Conceptually, enhanced performance of the personalized models could be achieved by including clinical data relevant to HF in the models. Beyond the scope of CIED-derived monitoring, variables that are typically included in predictive HF models are age, gender, systolic blood pressure, laboratory values (sodium, creatinine, hemoglobin, blood urea nitrogen, N-terminal pro B-type natriuretic peptide), diabetes, NYHA class and ejection fraction (67). Of note, another algorithm that has been developed for the purpose of CIED-derived monitoring is HeartInsight, produced by Biotronik SE & Co. KG (Berlin, Germany) (68). Notably, this model is adjusted with a baseline variable, for which the Seattle Heart Failure Model (SHFM) is used (69). SHFM includes gender, age, baseline NYHA, LVEF, medication use, systolic blood pressure, etiology of HF, and multiple laboratory values.

For this proof-of-concept study, a SVC and an XGBoost model were developed. In future research it is interesting to consider other models as well. The models developed in this study rely on data from previous episodes of HF decompensation. Ideally, a model that can predict HF decompensation in patients who have not previously experienced an episode would be developed. This approach aligns with anomaly detection, an unsupervised ML task that is trained with only 'normal' observations (40). Moreover, recent studies with AI models in the field of HF care explored with amongst others neural networks, classification and regression trees, long short-term memory, Markov models and adaptive boosting (28, 70). Specifically, a Markov model might be a feasible approach to consider, given that this model is typically used for problems involving sequential decisions over time (71). Lastly, it may be feasible to consider combining the personalized models with models that are trained on data of all patients, making use of model ensembling techniques (40).

5.3 Study limitations

A number of limitations of this study should be highlighted. Firstly, several features in the dataset contained a significant amount of missing data. Whereas it was concluded that the missing data was not related to the outcome, the cause of the missing data was not determined. Thereafter, all missing data were imputed with the median value, a relatively simple approach to missing data imputation. Potentially, advanced methods such as *k*-nearest neighbors or tree-based models are more effective strategies (62). Additionally, imputation with median values does not take into account the time component of the data.

Secondly, all available features known to be associated with HF decompensation were included in the models. However, many of these features are not exclusive to HF decompensation. Including non-specific features in the model might constrain its predictive ability. For example, heart rate, HRV and atrial arrhythmia data are not uniquely indicative of HF decompensation. A previously developed algorithm, which included these parameters along with several other clinical variables, proved ineffective in accurately predicting HF admissions (72).

Thirdly, Bayesian optimization was applied for the purpose of hyperparameter tuning. The defined search spaces allowed for a broad range in numerical hyperparameters. Consequently, there is a variety in selected hyperparameters throughout the different models. Moreover, the minimal constraints applied to the search space may have resulted in combinations of hyperparameters that enable overfitting. Furthermore, given the high computational burden of training a SVC, the SVC models were optimized with simply ten iterations and only two kernel types were allowed. Therefore, it is possible that the optimal hyperparameters for SVC were not identified.

Finally, the models were trained on episodes of HF decompensation, as detected by the HeartLogic algorithm. With this method, the models are to a certain degree trained to mimic HeartLogic's predictions. The HeartLogic guided care path adapted in the LUMC is able to identify episodes of HF decompensation with a sensitivity of 79-90% and a specificity of 89% (22, 23). The episodes that were undetected by HeartLogic were also included in this study. Nonetheless, specifying when an episode of decompensation exactly starts or ends is complex. As the mathematical methods behind HeartLogic are unknown, the reasoning behind the start or end date are unclear. In addition, in this study, the complex nature of these episodes was simplified to a two-class classification problem.

All observations from start to end of the were classified as ‘unstable HF’ and no distinction in severity of the episodes was made.

6

Conclusion

In summary, CIED-derived personalized models were developed for the detection of upcoming HF decompensation. For each patient, two machine learning models were developed: an XGBoost model and a SVC model. The model with better performance in terms of AUPRC was selected for further evaluation. In general, the XGBoost models demonstrated superior performance to the SVC models.

Overall, the selected models exhibit a moderate performance in classifying patients’ HF status. The high predictive performance obtained in a subset of patients advocates for further development of personalized models. However, the significant observed variation in performance highlight the need for further investigation on potential causes of low performance in order to improve robustness of the models.

References

1. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021;42(36):3599-726.
2. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GMC, Coats AJS. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res. 2023;118(17):3272-87.
3. VZinfo.nl. Hartfalen | Zorguitgaven. RIVM:Bilthoven 2022 [Available from: https://www.vzinfo.nl/hartfalen/zorguitgaven.
4. (RIVM) RvVeM. Hartfalen: epidemiologie, risicofactoren en toekomst. 2012.
5. Ministerie van Volksgezondheid WeS. Intergraal Zorg Akkoord, Samen werken aan gezonde zorg. https://www.rijksoverheid.nl/2022.
6. Setoguchi S, Stevenson LW, Schneeweiss S. Repeated hospitalizations predict mortality in the community population with heart failure. Am Heart J. 2007;154(2):260-6.
7. Koehler F, Koehler K, Deckwart O, et al. Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. Lancet. 2018;392(10152):1047-57.
8. Inglis SC, Clark RA, McAlister FA, et al. Structured telephone support or telemonitoring programmes for patients with chronic heart failure. Cochrane Database Syst Rev. 2010(8):CD007228.
9. Lin MH, Yuan WL, Huang TC, Zhang HF, Mai JT, Wang JF. Clinical effectiveness of telemedicine for chronic heart failure: a systematic review and meta-analysis. J Investig Med. 2017;65(5):899-911.
10. Thorvaldsen T, Benson L, Dahlström U, Edner M, Lund LH. Use of evidence-based therapy and survival in heart failure in Sweden 2003-2012. Eur J Heart Fail. 2016;18(5):503-11.
11. Feijen M, Egorova AD, Kuijken T, Bootsma M, Schalij MJ, van Erven L. One-Year Mortality in Patients Undergoing an Implantable Cardioverter Defibrillator or Cardiac Resynchronization Therapy Pulse Generator Replacement: Identifying Patients at Risk. J Clin Med. 2023;12(17).
12. Adamson PB, Smith AL, Abraham WT, et al. Continuous autonomic assessment in patients with symptomatic heart failure: prognostic value of heart rate variability measured by an implanted cardiac resynchronization device. Circulation. 2004;110(16):2389-94.
13. Adamson PB. Pathophysiology of the transition from chronic compensated and acute decompensated heart failure: new insights from continuous monitoring devices. Curr Heart Fail Rep. 2009;6(4):287-92.
14. Vegh EM, Kandala J, Orencole M, et al. Device-measured physical activity versus six-minute walk test as a predictor of reverse remodeling and outcome after cardiac resynchronization therapy for heart failure. Am J Cardiol. 2014;113(9):1523-8.
15. Conraads VM, Tavazzi L, Santini M, et al. Sensitivity and positive predictive value of implantable intrathoracic impedance monitoring as a predictor of heart failure hospitalizations: the SENSE-HF trial. Eur Heart J. 2011;32(18):2266-73.
16. Maier SKG, Paule S, Jung W, et al. Evaluation of thoracic impedance trends for implant-based remote monitoring in heart failure patients - Results from the (J-)HomeCARE-II Study. J Electrocardiol. 2019;53:100-8.

17. Forleo GB, Santini L, Campoli M, et al. Long-term monitoring of respiratory rate in patients with heart failure: the Multiparametric Heart Failure Evaluation in Implantable Cardioverter-Defibrillator Patients (MULTITUDE-HF) study. J Interv Card Electrophysiol. 2015;43(2):135-44.
18. Scholte NTB, Gürgöze MT, Aydin D, et al. Telemonitoring for heart failure: a meta-analysis. Eur Heart J. 2023;44(31):2911-26.
19. Zito A, Princi G, Romiti GF, et al. Device-based remote monitoring strategies for congestion-guided management of patients with heart failure: a systematic review and meta-analysis. Eur J Heart Fail. 2022;24(12):2333-41.
20. Boehmer JP, Hariharan R, Devecchi FG, et al. A Multisensor Algorithm Predicts Heart Failure Events in Patients With Implanted Devices: Results From the MultiSENSE Study. JACC Heart Fail. 2017;5(3):216-25.
21. de Juan Bagudá J, Gavira Gómez JJ, Pachón Iglesias M, et al. Remote heart failure management using the HeartLogic algorithm. RE-HEART registry. Rev Esp Cardiol (Engl Ed). 2022;75(9):709-16.
22. Treskes RW, Beles M, Caputo ML, et al. Clinical and economic impact of HeartLogic™ compared with standard care in heart failure patients. ESC Heart Fail. 2021;8(2):1541-51.
23. Feijen M, Egorova AD, Treskes RW, et al. Performance of a HeartLogic(TM) Based Care Path in the Management of a Real-World Chronic Heart Failure Population. Front Cardiovasc Med. 2022;9:883873.
24. Capucci A, Santini L, Favale S, et al. Preliminary experience with the multisensor HeartLogic algorithm for heart failure monitoring: a retrospective case series report. ESC Heart Fail. 2019;6(2):308-18.
25. Santobuono VE, Favale S, D'Onofrio A, et al. Performance of a multisensor implantable defibrillator algorithm for heart failure monitoring related to co-morbidities. ESC Heart Fail. 2023;10(4):2469-78.
26. Feijen M, Beles M, Tan YZ, et al. Fewer Worsening Heart Failure Events With HeartLogic on top of Standard Care: a Propensity-Matched Cohort Analysis. J Card Fail. 2023.
27. Boehmer JP, Wariar R, Zhang Y, et al. Rationale and Design of the Multisensor Chronic

Evaluations in Ambulatory Heart Failure Patients

(MultiSENSE) Study. The Journal of Innovations in Cardiac Rhythm Management. 2015;6:2137–43.
28. Gautam N, Ghanta SN, Mueller J, et al. Artificial Intelligence, Wearables and Remote Monitoring for Heart Failure: Current and Future Applications. Diagnostics (Basel). 2022;12(12).
29. Triposkiadis F, Butler J, Abboud FM, et al. The continuous heart failure spectrum: moving beyond an ejection fraction classification. Eur Heart J. 2019;40(26):2155-63.
30. Greene SJ, Bauersachs J, Brugts JJ, et al. Worsening Heart Failure: Nomenclature, Epidemiology, and Future Directions: JACC Review Topic of the Week. J Am Coll Cardiol. 2023;81(4):413-24.
31. Pandhi P, Ter Maaten JM, Anker SD, et al. Pathophysiologic Processes and Novel Biomarkers Associated With Congestion in Heart Failure. JACC Heart Fail. 2022;10(9):623-32.
32. Zile MR, Bennett TD, St John Sutton M, et al. Transition from chronic compensated to acute decompensated heart failure: pathophysiological insights obtained from continuous monitoring of intracardiac pressures. Circulation. 2008;118(14):1433-41.
33. Siejko KZ, Thakur PH, Maile K, Patangay A, Olivari MT. Feasibility of heart sounds measurements from an accelerometer within an ICD pulse generator. Pacing Clin Electrophysiol. 2013;36(3):334-46.

34. Thakur PH, An Q, Swanson L, Zhang Y, Gardner RS. Haemodynamic monitoring of cardiac status using heart sounds from an implanted cardiac device. ESC Heart Fail. 2017;4(4):605-13.
35. Mehta NJ, Khan IA. Third heart sound: genesis and clinical importance. Int J Cardiol. 2004;97(2):183-6.
36. Calò L, Capucci A, Santini L, et al. ICD-measured heart sounds and their correlation with echocardiographic indexes of systolic and diastolic function. J Interv Card Electrophysiol. 2020;58(1):95-101.
37. Abraham WT, Compton S, Haas G, et al. Intrathoracic impedance vs daily weight monitoring for predicting worsening heart failure events: results of the Fluid Accumulation Status Trial (FAST). Congest Heart Fail. 2011;17(2):51-5.
38. Yu CM, Wang L, Chau E, et al. Intrathoracic impedance monitoring in patients with heart failure: correlation with fluid status and feasibility of early warning preceding hospitalization. Circulation. 2005;112(6):841-8.
39. Carlisle MA, Fudim M, DeVore AD, Piccini JP. Heart Failure and Atrial Fibrillation, Like Fire and Fury. JACC Heart Fail. 2019;7(6):447-56.
40. Géron A. Hands-On Machine Learning with Scikit-Learn & TensorFlow. 1 ed. United States O'Reilly Media Inc; 2017.
41. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. 2 ed. Casella G, Fienberg S, Olkin I, editors: Springer Science+Business Media; 2013.
42. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
44. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001:1189-232.
45. Introduction to Boosted Trees xgboost developers 2022 [82d846bb:[Available from: https://xgboost.readthedocs.io/en/stable/index.html#.
46. Rodriguez C. The Notorious XGBoost: Towards Data Science 2023 [Available from: https://towardsdatascience.com/the-notorious-xgboost-c7f7adc4c183.
47. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011;2(3):1-27.
48. Maimon O, Rokach L. Data mining and knowledge discovery handbook: Springer; 2005.
49. Singh N. Soft Margin SVM: Exploring Slack Variables, the 'C' Parameter, and Flexibility: AI Mind; 2023 [Available from: https://pub.aimind.so/soft-margin-svm-exploring-slack-variables-the-c-parameter-and-flexibility-1555f4834ecc.
50. Dioşan L, Rogozan A, Pecuchet J-P. Improving classification performance of support vector machine by genetically optimising kernel shape and hyper-parameters. Applied Intelligence. 2012;36:280-94.
51. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent data analysis. 2002;6(5):429-49.
52. Suresh T, Brijet Z, Subha TD. Imbalanced medical disease dataset classification using enhanced generative adversarial network. Comput Methods Biomech Biomed Engin. 2023;26(14):1702-18.

53. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.
54. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE. 2015;104(1):148-75.
55. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Advances in neural information processing systems. 2011;24.
56. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems. 2012;25.
57. skopt.BayesSearchCV: schikit-optimize contributors [Available from: https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html.
58. Vanderheyden M, Houben R, Verstreken S, et al. Continuous monitoring of intrathoracic impedance and right ventricular pressures in patients with heart failure. Circ Heart Fail. 2010;3(3):370-7.
59. HeartLogic Alert Management Guide Marlborough; 2023.
60. Cao M, Stolen CM, Ahmed R, et al. Small decreases in biventricular pacing percentages are associated with multiple metrics of worsening heart failure as measured from a cardiac resynchronization therapy defibrillator. Int J Cardiol. 2021;335:73-9.
61. Brodersen KH, Ong CS, Stephan KE, Buhmann JM, editors. The balanced accuracy and its posterior distribution. 2010 20th international conference on pattern recognition; 2010: IEEE.
62. Kuhn M, Johnson, K. Feature Engineering and Selection: A Practical Approach for Predictive Models Boca Raton: Taylor & Francis Group; 2019.
63. Alkhawaldeh IM, Albalkhi I, Naswhan AJ. Challenges and limitations of synthetic minority oversampling techniques in machine learning. World J Methodol. 2023;13(5):373-8.
64. Isler Y, Narin A, Ozer M, Perc M. Multi-stage classification of congestive heart failure based on short-term heart rate variability. Chaos, Solitons & Fractals. 2019;118:145-51.
65. Cowie MR, Sarkar S, Koehler J, et al. Development and validation of an integrated diagnostic algorithm derived from parameters monitored in implantable devices for identifying patients at risk for heart failure hospitalization in an ambulatory setting. Eur Heart J. 2013;34(31):2472-80.
66. Virani SA, Sharma V, McCann M, Koehler J, Tsang B, Zieroth S. Prospective evaluation of integrated device diagnostics for heart failure management: results of the TRIAGE-HF study. ESC Heart Fail. 2018;5(5):809-17.
67. Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. JACC Heart Fail. 2014;2(5):429-36.
68. D'Onofrio A, Solimene F, Calò L, et al. Combining home monitoring temporal trends from implanted defibrillators and baseline patient risk profile to predict heart failure hospitalizations: results from the SELENE HF study. Europace. 2022;24(2):234-44.
69. Levy WC, Mozaffarian D, Linker DT, et al. The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation. 2006;113(11):1424-33.
70. Khan MS, Arshad MS, Greene SJ, et al. Artificial intelligence and heart failure: A state-of-the-art review. Eur J Heart Fail. 2023;25(9):1507-25.

71. Data MITC, Komorowski M, Raffa J. Markov Models and Cost Effectiveness Analysis: Applications in Medical Research. Secondary Analysis of Electronic Health Records. Cham (CH): Springer

Copyright 2016, The Author(s). 2016. p. 351-67.

72. Gilliam F, Ewald GA, Sweeney RJ. Feasibility of automated heart failure decompensation detection using remote patient monitoring: results from the decompensation detection study. Journal of Innovations in Cardiac Rhythm Management. 2012;3:1-10.

1. Hyperparameter search spaces for Bayesian optimization

The hyperparameter search spaces for Bayesian optimization indicate the allowed ranges within which the hyperparameters can be explored.

Table S1: *Hyperparameter search space for optimization of the SVC*

Hyperparameter	Search space	Scale
C	0.001 – 1000	Real
Gamma	0.0001 – 1	Real
Kernel	Linear or radial basis function	Categorical
Class weight	Balanced, 1:5, 1:10 or 1:15	Categorical

SVC, Support Vector Classifier

Table S2: *Hyperparameter search space for optimization of the XGBoost model*

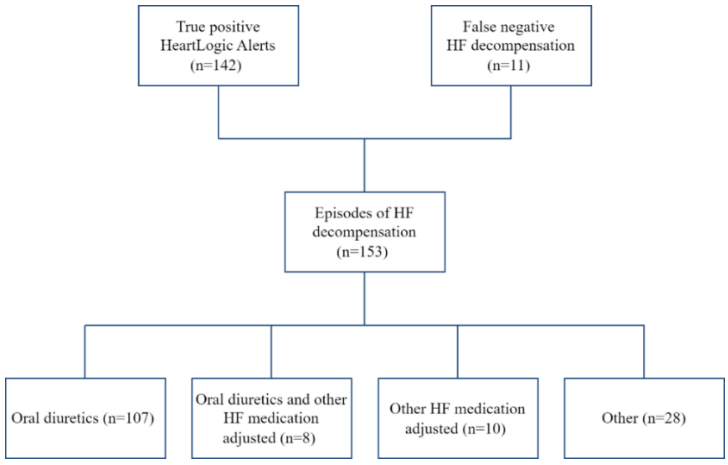
Hyperparameter	Search space	Scale
Learning rate	0.01 – 0.2	Real
Maximum depth	3 – 10	Integer
Minimum child weight	1 – 10	Integer
Subsample	0.5 – 1.0	Real
Colsample by tree	0.5 – 1.0	Real
Number of estimators	50 – 300	Integer
Scale position weight	1 – 10	Real

XGBoost, Extreme Gradient Boosting

2. Alert follow-up

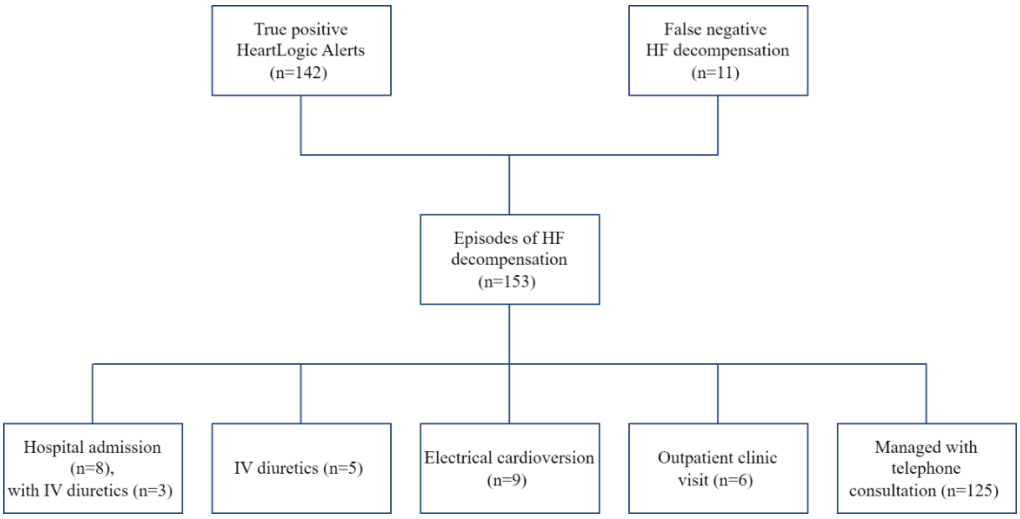
Figure S1 and Figure S2 provide a schematic overview of the clinical course of the episodes of HF decompensation that were used to derive the outputs of the models; the daily classification of a patient's HF status as 'unstable' or 'stable'.

Figure S1: *Schematic overview of the clinical actions taken in response to episodes of HF decompensation.*



HF, heart failure

Figure S2: *Schematic overview of management of episodes of HF decompensation.*



HF, heart failure; IV, intravenous

3. Class distribution

Table S3 presents the class distribution in the data.

Table S3: *Class distribution in data*

Study Id	Total number of observations (n)	Total number of unstable HF observations (n)	Proportion of observations in unstable HF class (%)	Number of unstable HF observations in development dataset (n)	Number of unstable HF samples in independent test set (n)
4	2137	10	0.468	8	2
9	2106	10	0.475	8	2
10	2051	163	7.95	139	24
11	1944	328	16.9	279	49
12	2082	20	0.961	17	3
20	1823	33	1.81	28	5
33	1085	89	8.20	76	13
35	1547	111	7.18	94	17
37	1316	86	6.53	73	13
38	1386	120	8.66	102	18
39	1528	64	4.19	54	10
42	1345	4	0.297	3	1
47	996	38	3.82	32	6
50	1508	604	40.1	513	91
51	1575	16	1.02	14	2
55	1095	119	10.9	101	18
56	1575	187	11.9	159	28
57	1400	22	1.57	19	3

58	1697	92	5.42	78	14
62	1480	526	35.5	447	79
69	1644	32	1.95	27	5
72	1289	40	3.10	34	6
78	1684	106	6.29	90	16
81	1340	270	20.1	230	40
86	1469	106	7.22	90	16
89	1367	34	2.49	29	5
90	1358	31	2.28	26	5
94	1361	98	7.20	83	15
95	1613	515	31.9	438	77
96	1281	191	14.9	162	29
101	1329	32	2.41	27	5
102	1219	126	10.3	107	19
109	1043	64	6.14	54	10
114	990	35	3.54	30	5
115	994	135	13.6	115	20
116	955	17	1.78	14	3
118	945	252	26.7	214	38
122	858	100	11.7	85	15
123	822	84	10.2	71	13
124	832	206	24.8	175	31
125	945	106	11.2	90	16
126	768	13	1.69	11	2
127	809	27	3.34	23	4
128	792	80	10.1	68	12
129	696	20	2.87	17	3
131	678	11	1.62	9	2
132	781	17	2.18	14	3
138	697	51	7.32	43	8
140	669	121	18.1	103	18
143	586	11	1.88	9	2
145	517	34	6.58	29	5
146	503	50	9.94	42	8
150	498	93	18.7	79	14
151	499	127	25.5	108	19
152	472	55	11.7	47	8
153	443	20	4.51	17	3
157	298	104	34.9	88	16
164	619	58	9.37	49	9
167	601	10	1.66	8	2
175	94	23	24.5	19	4
176	129	34	26.4	29	5
179	112	32	28.6	27	5

4. Hyperparameters

Table S4 and Table S5 present the results of the Bayesian hyperparameter optimization. These tables provide the hyperparameters that resulted in the highest AUPRC in the validation sets within the 5-fold CV.

Table S4: *Overview of selected hyperparameters for the SVC models.*

Study Id	Scaler	k neighbours	C	Gamma*	Kernel	Class weight
4	RobustScaler()	4	0.683	-	linear	{0: 1, 1: 5}
9	RobustScaler()	4	0.609	$7.75*10^{-3}$	rbf	{0: 1, 1: 5}
10	RobustScaler()	4	10.9	-	linear	balanced
11	StandardScaler()	4	7.42	$1.60*10^{-4}$	rbf	None
12	RobustScaler()	4	82.0	0.0680	rbf	None
20	RobustScaler()	4	3.47	0.0110	rbf	{0: 1, 1: 15}
33	StandardScaler()	4	$1.04*10^{-3}$	-	linear	{0: 1, 1: 10}
35	StandardScaler()	4	939	$1.36*10^{-3}$	rbf	balanced
37	StandardScaler()	4	1.74	-	linear	{0: 1, 1: 5}
38	StandardScaler()	4	80.5	-	linear	balanced
39	StandardScaler()	3	0.628	-	linear	{0: 1, 1: 5}
42	RobustScaler()	1	0.399	0.239	rbf	{0: 1, 1: 10}
47	StandardScaler()	5	200	-	linear	{0: 1, 1: 5}
50	StandardScaler()	4	599	$5.22*10^{-3}$	rbf	{0: 1, 1: 10}
51	StandardScaler()	4	45.1	0.105	rbf	{0: 1, 1: 10}
55	StandardScaler()	3	$9.72*10^{-3}$	0.624	rbf	{0: 1, 1: 5}
56	StandardScaler()	4	1.71	-	linear	{0: 1, 1: 5}
57	RobustScaler()	4	6.09	$8.96*10^{-4}$	rbf	{0: 1, 1: 5}
58	RobustScaler()	5	109	-	linear	None
62	StandardScaler()	4	296	0.0334	rbf	{0: 1, 1: 15}
69	RobustScaler()	4	5.37	-	linear	{0: 1, 1: 5}
72	StandardScaler()	4	82.4	-	linear	None
78	StandardScaler()	4	0.647	-	linear	{0: 1, 1: 10}
81	StandardScaler()	4	0.396	-	linear	balanced
86	StandardScaler()	4	0.323	-	linear	{0: 1, 1: 10}
89	StandardScaler()	3	6.68	$6.28*10^{-4}$	rbf	balanced
90	RobustScaler()	4	0.667	-	linear	{0: 1, 1: 10}
94	RobustScaler()	4	719	$4.25*10^{-4}$	rbf	{0: 1, 1: 10}
95	StandardScaler()	5	25.6	-	linear	balanced
96	RobustScaler()	4	13.2	-	linear	{0: 1, 1: 5}
101	StandardScaler()	3	122	$6.25*10^{-4}$	rbf	{0: 1, 1: 10}
102	StandardScaler()	5	0.583	-	linear	balanced
109	StandardScaler()	3	0.0701	-	linear	balanced
114	RobustScaler()	3	174	$1.02*10^{-4}$	rbf	{0: 1, 1: 5}
115	RobustScaler()	3	4.26	$4.79*10^{-3}$	rbf	balanced
116	StandardScaler()	4	21.3	$6.33*10^{-3}$	rbf	balanced
118	RobustScaler()	5	7.60	$5.53*10^{-4}$	rbf	balanced
122	StandardScaler()	5	$2.15*10^{-3}$	-	linear	None

123	RobustScaler()	5	5.41	-	linear	{0: 1, 1: 5}
124	StandardScaler()	3	0.243	9.45	rbf	balanced
125	StandardScaler()	5	0.736	0.142	rbf	None
126	StandardScaler()	5	0.304	0.0747	rbf	{0: 1, 1: 5}
127	StandardScaler()	3	0.0998	-	linear	None
128	StandardScaler()	5	$1.66*10^{-3}$	0.11	rbf	{0: 1, 1: 10}
129	StandardScaler()	4	7.08	-	linear	balanced
131	StandardScaler()	4	22.7	$1.79*10^{-4}$	rbf	{0: 1, 1: 10}
132	RobustScaler()	4	0.0394	$1.65*10^{-4}$	rbf	balanced
138	RobustScaler()	4	2.23	$5.40*10^{-3}$	rbf	{0: 1, 1: 10}
140	StandardScaler()	3	97.3	0.338	rbf	balanced
143	StandardScaler()	4	158	0.0244	rbf	balanced
145	RobustScaler()	3	0.334	-	linear	None
146	RobustScaler()	5	0.061	-	linear	{0: 1, 1: 10}
150	RobustScaler()	4	0.0337	-	linear	{0: 1, 1: 10}
151	RobustScaler()	4	1.38	0.0168	rbf	{0: 1, 1: 10}
152	RobustScaler()	5	1.11	-	linear	None
153	StandardScaler()	4	95.2	$4.96*10^{-3}$	rbf	{0: 1, 1: 10}
157	StandardScaler()	5	951	-	linear	None
164	StandardScaler()	5	1.62	0.0188	rbf	balanced
167	RobustScaler()	4	0.075	0.583	rbf	{0: 1, 1: 15}
175	StandardScaler()	4	315	$8.55*10^{-4}$	rbf	None
176	StandardScaler()	3	$6.18*10^{-3}$	-	linear	{0: 1, 1: 10}
179	StandardScaler()	3	1.39	-	linear	{0: 1, 1: 10}

SVC, Support Vector Classifier

*Only applicable to non-linear kernels

Table S5: *Overview of selected hyperparameters for XGBoost models.*

Study Id	**Scaler**	**k neigh-bors**	**Learn-ing rate**	**Maximum depth**	**Minimum child weight**	**Subsample**	**Colsample by tree**	**Number of estimators**	**Scale position weight**
4	StandardScaler()	3	0.200	3	1	0.500	0.699	300	1.00
9	RobustScaler()	5	0.107	9	8	0.992	0.511	247	1.00
10	RobustScaler()	3	0.130	3	1	0.500	1.00	300	1.00
11	RobustScaler()	3	0.200	3	10	1.00	0.804	300	10.0
12	RobustScaler()	5	0.0958	10	1	0.500	0.500	202	1.00
20	RobustScaler()	5	0.200	10	10	0.500	1.00	240	10.0
33	RobustScaler()	3	0.200	5	1	0.661	0.500	50	1.00
35	RobustScaler()	5	0.0367	10	1	1.00	0.500	247	1.00
37	StandardScaler()	5	0.200	9	7	1.00	1.00	267	10.0
38	StandardScaler()	4	0.0828	3	1	1.00	0.500	300	1.00
39	RobustScaler()	3	0.188	10	1	1.00	0.662	285	1.21
42	RobustScaler()	1	0.017	3	2	1.00	0.756	74	4.99
47	StandardScaler()	5	0.200	3	10	0.500	1.00	300	10.0
50	StandardScaler()	5	0.056	10	1	0.500	1.00	300	10.0

51	RobustScaler()	3	0.200	10	1	0.500	1.00	300	10.0
55	StandardScaler()	3	0.186	5	3	1.00	0.500	50	10.0
56	RobustScaler()	5	0.200	10	3	0.500	0.501	300	8.97
57	RobustScaler()	5	0.0783	3	10	1.00	1.00	300	1.74
58	StandardScaler()	3	0.0948	10	2	1.00	1.00	135	1.00
62	StandardScaler()	5	0.0255	10	3	1.00	0.500	300	10.0
69	RobustScaler()	5	0.200	8	1	0.724	0.884	50	10.0
72	RobustScaler()	3	0.0187	3	1	1.00	1.00	300	10.0
78	StandardScaler()	5	0.010	4	1	1.00	0.500	210	4.18
81	StandardScaler()	3	0.010	9	1	0.800	0.500	300	1.00
86	RobustScaler()	4	0.200	3	9	0.500	0.624	193	10.0
89	RobustScaler()	4	0.192	10	7	0.848	0.522	297	9.94
90	RobustScaler()	3	0.200	10	10	1.00	0.500	50	10.0
94	StandardScaler()	5	0.200	3	10	1.00	1.00	300	4.07
95	StandardScaler()	4	0.0525	9	1	0.984	0.606	263	9.44
96	RobustScaler()	3	0.101	7	1	0.500	1.00	300	1.00
101	RobustScaler()	3	0.200	3	1	0.500	0.500	300	1.00
102	RobustScaler()	3	0.0511	10	1	0.500	1.00	204	1.00
109	RobustScaler()	4	0.0757	9	1	1.00	0.500	144	5.25
114	StandardScaler()	5	0.200	3	4	0.500	0.500	194	1.00
115	StandardScaler()	3	0.0949	10	1	0.719	0.865	300	10.0
116	RobustScaler()	3	0.200	3	1	0.500	0.500	300	10.0
118	RobustScaler()	3	0.200	10	1	1.00	0.500	50	1.00
122	RobustScaler()	3	0.0704	10	9	1.00	0.634	300	7.91
123	RobustScaler()	4	0.0593	10	1	0.500	0.500	300	2.95
124	RobustScaler()	3	0.102	3	1	1.00	0.500	173	10.0
125	StandardScaler()	5	0.0944	9	1	0.729	0.703	300	10.0
126	RobustScaler()	5	0.0398	7	5	0.750	0.818	300	10.0
127	RobustScaler()	5	0.192	6	8	0.536	0.617	242	1.41
128	RobustScaler()	3	0.0798	8	1	0.500	1.00	233	1.00
129	RobustScaler()	5	0.200	10	10	1.00	0.923	300	10.0
131	StandardScaler()	5	0.200	10	1	0.500	0.500	300	10.0
132	RobustScaler()	3	0.200	4	5	0.886	0.901	103	1.40
138	RobustScaler()	5	0.184	3	8	0.999	0.500	158	10.0
140	RobustScaler()	3	0.0262	10	1	0.500	1.00	208	10.0
143	StandardScaler()	3	0.108	3	1	0.500	0.500	285	10.0
145	RobustScaler()	5	0.200	10	1	0.500	0.505	239	10.0
146	RobustScaler()	5	0.200	10	1	0.500	0.500	300	1.00
150	RobustScaler()	5	0.165	5	1	0.899	0.918	278	6.84
151	StandardScaler()	4	0.0553	10	10	0.500	1.00	300	10.0
152	StandardScaler()	4	0.0776	5	1	0.500	0.500	300	7.66
153	StandardScaler()	5	0.010	5	1	1.00	0.500	300	10.0
157	StandardScaler()	4	0.196	8	1	0.852	0.500	50	6.32
164	RobustScaler()	3	0.0636	3	1	0.500	1.00	300	1.00
167	RobustScaler()	5	0.121	8	1	1.00	0.500	50	5.12
175	StandardScaler()	5	0.158	4	1	0.569	0.643	219	6.68
176	RobustScaler()	5	0.200	10	3	0.583	0.961	300	10.0

179	RobustScaler()	5	0.200	3	3	1.00	0.500	50	10.0

XGBoost, Extreme Gradient Boosting

5. Performance evaluation, results obtained with leave-one-out cross-validation

Table S6 and Table S7 provide the performance evaluation of the models, obtained with leave-one-out cross-validation.

Table S6: *Performance of the SVC models for all patients with leave-one-out cross-validation.*

Study Id	AUPRC	AUROC	Balanced accuracy	Sensitivity (recall)	Specificity	PPV (precision)	NPV
4	0.160	0.980	0.624	0.25	0.997	0.286	0.997
9	0.0496	0.915	0.836	0.75	0.923	0.0417	0.999
10	0.479	0.867	0.787	0.763	0.81	0.259	0.975
11	0.447	0.799	0.73	0.692	0.768	0.378	0.925
12	0.581	0.987	0.846	0.706	0.987	0.343	0.997
20	0.118	0.88	0.758	0.607	0.909	0.109	0.992
33	0.863	0.986	0.889	0.987	0.791	0.298	0.999
35	0.646	0.891	0.828	0.734	0.923	0.423	0.978
37	0.324	0.904	0.837	0.973	0.702	0.186	0.997
38	0.381	0.867	0.799	0.863	0.735	0.236	0.983
39	0.284	0.844	0.754	0.833	0.675	0.100	0.989
42	0.00848	0.745	0.652	0.333	0.971	0.0294	0.998
47	0.68	0.956	0.891	0.844	0.937	0.346	0.993
50	0.883	0.908	0.812	0.910	0.714	0.680	0.923
51	0.0695	0.801	0.499	0.00	0.998	0.00	0.990
55	0.557	0.932	0.500	1.00	0.00	0.109	0.00
56	0.455	0.844	0.682	0.987	0.377	0.176	0.996
57	0.238	0.852	0.769	0.684	0.853	0.0703	0.994
58	0.662	0.926	0.864	0.795	0.934	0.408	0.988
62	0.822	0.862	0.786	0.734	0.837	0.713	0.851
69	0.632	0.987	0.861	0.741	0.98	0.426	0.995
72	0.276	0.901	0.864	0.824	0.905	0.217	0.994
78	1.00	1.00	0.999	1.00	0.998	0.968	1.00
81	0.439	0.769	0.697	0.704	0.689	0.364	0.902
86	0.350	0.864	0.756	0.933	0.579	0.147	0.991
89	0.825	0.994	0.982	1.00	0.964	0.414	1.00
90	0.739	0.991	0.899	0.808	0.989	0.636	0.996
94	0.434	0.87	0.783	0.867	0.698	0.182	0.986
95	0.780	0.868	0.787	0.795	0.779	0.628	0.890
96	0.517	0.886	0.793	0.951	0.635	0.313	0.987
101	0.869	0.996	0.921	0.852	0.99	0.676	0.996
102	0.539	0.814	0.775	0.692	0.859	0.361	0.960
109	0.550	0.851	0.809	0.759	0.859	0.259	0.982
114	0.248	0.948	0.922	0.967	0.878	0.227	0.999

115	0.593	0.897	0.826	0.791	0.861	0.474	0.963
116	0.424	0.854	0.725	0.5	0.950	0.149	0.991
118	0.640	0.825	0.738	0.734	0.742	0.508	0.885
122	0.686	0.916	0.833	0.894	0.772	0.341	0.982
123	0.726	0.953	0.848	0.986	0.710	0.278	0.998
124	0.863	0.941	0.858	0.794	0.921	0.768	0.932
125	0.546	0.913	0.701	0.456	0.947	0.519	0.932
126	0.278	0.869	0.627	0.273	0.981	0.200	0.987
127	0.396	0.961	0.884	0.826	0.941	0.328	0.994
128	0.307	0.798	0.500	1.00	0.00	0.101	0.00
129	0.0987	0.684	0.577	0.529	0.625	0.0402	0.978
131	0.116	0.807	0.708	0.667	0.75	0.0405	0.993
132	0.0656	0.690	0.505	1.00	0.0108	0.0213	1.00
138	0.572	0.929	0.852	0.977	0.727	0.219	0.998
140	0.520	0.815	0.661	0.388	0.933	0.563	0.873
143	0.127	0.890	0.606	0.222	0.99	0.286	0.986
145	0.683	0.931	0.877	0.828	0.927	0.444	0.987
146	0.344	0.878	0.752	1.00	0.504	0.18	1.00
150	0.898	0.960	0.892	0.987	0.797	0.527	0.996
151	0.510	0.773	0.670	0.972	0.367	0.344	0.975
152	0.433	0.810	0.781	0.745	0.816	0.35	0.96
153	0.334	0.896	0.636	0.294	0.978	0.385	0.967
157	0.669	0.784	0.731	0.716	0.745	0.6	0.831
164	0.562	0.954	0.887	0.878	0.897	0.467	0.986
167	0.133	0.138	0.562	0.125	1.00	1.00	0.986
175	0.644	0.845	0.724	0.632	0.817	0.522	0.875
176	0.597	0.846	0.500	1.00	0.00	0.266	0.00
179	0.819	0.927	0.845	0.778	0.912	0.778	0.912

AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristic; PPV, positive predictive value; NPV, negative predictive value; SVC, Support Vector Classifier

Table S7: *Performance of XGBoost models for all patients with leave-one-out cross-validation.*

Study Id	AUPRC	AUROC	Balanced accuracy	Sensitivity (recall)	Specificity	PPV (precision)	NPV
4	0.195	0.974	0.686	0.375	0.997	0.375	0.997
9	0.0341	0.887	0.496	0.00	0.992	0.00	0.995
10	0.729	0.962	0.797	0.619	0.974	0.677	0.967
11	0.559	0.85	0.761	0.677	0.844	0.469	0.928
12	0.86	0.998	0.852	0.706	0.998	0.750	0.997
20	0.270	0.873	0.717	0.464	0.97	0.220	0.990
33	0.956	0.995	0.955	0.921	0.988	0.875	0.993
35	0.736	0.953	0.775	0.574	0.975	0.643	0.967
37	0.410	0.927	0.757	0.589	0.925	0.355	0.970
38	0.479	0.895	0.693	0.431	0.955	0.478	0.947
39	0.377	0.804	0.676	0.37	0.982	0.476	0.973

42	0.00314	0.613	0.468	0.00	0.937	0.00	0.997
47	0.611	0.976	0.887	0.812	0.962	0.456	0.992
50	0.904	0.928	0.848	0.899	0.798	0.748	0.922
51	0.230	0.852	0.604	0.214	0.993	0.250	0.992
55	0.877	0.977	0.942	0.950	0.934	0.636	0.994
56	0.656	0.906	0.778	0.623	0.933	0.556	0.948
57	0.163	0.841	0.675	0.368	0.982	0.250	0.990
58	0.747	0.900	0.845	0.705	0.985	0.724	0.983
62	0.884	0.903	0.821	0.861	0.781	0.684	0.911
69	0.726	0.985	0.863	0.741	0.986	0.513	0.995
72	0.549	0.951	0.856	0.765	0.948	0.321	0.992
78	0.998	1.00	0.998	1.00	0.996	0.938	1.00
81	0.602	0.815	0.682	0.422	0.942	0.647	0.866
86	0.664	0.938	0.845	0.767	0.923	0.437	0.981
89	0.879	0.995	0.892	0.793	0.99	0.676	0.995
90	0.832	0.995	0.915	0.846	0.983	0.537	0.996
94	0.459	0.864	0.740	0.554	0.926	0.368	0.964
95	0.922	0.953	0.872	0.881	0.863	0.751	0.939
96	0.585	0.877	0.743	0.568	0.919	0.551	0.924
101	0.727	0.985	0.828	0.667	0.989	0.600	0.992
102	0.594	0.888	0.730	0.495	0.966	0.624	0.943
109	0.703	0.887	0.807	0.630	0.984	0.723	0.976
114	0.593	0.971	0.789	0.600	0.978	0.500	0.985
115	0.755	0.938	0.858	0.783	0.934	0.652	0.965
116	0.392	0.960	0.707	0.429	0.985	0.333	0.990
118	0.726	0.874	0.768	0.64	0.896	0.692	0.873
122	0.711	0.919	0.824	0.729	0.919	0.544	0.963
123	0.978	0.997	0.969	0.958	0.981	0.85	0.995
124	0.925	0.968	0.896	0.909	0.883	0.719	0.967
125	0.758	0.957	0.855	0.767	0.942	0.627	0.970
126	0.147	0.853	0.755	0.545	0.964	0.207	0.992
127	0.438	0.944	0.767	0.565	0.968	0.382	0.985
128	0.589	0.89	0.735	0.515	0.955	0.565	0.946
129	0.344	0.895	0.771	0.588	0.953	0.270	0.987
131	0.239	0.953	0.717	0.444	0.989	0.400	0.991
132	0.0343	0.614	0.547	0.143	0.951	0.0588	0.981
138	0.674	0.966	0.892	0.837	0.947	0.554	0.987
140	0.579	0.861	0.781	0.796	0.766	0.429	0.944
143	0.324	0.974	0.717	0.444	0.99	0.444	0.990
145	0.485	0.937	0.834	0.724	0.944	0.477	0.980
146	0.581	0.935	0.724	0.500	0.948	0.512	0.946
150	0.935	0.979	0.915	0.873	0.956	0.821	0.971
151	0.514	0.785	0.707	0.741	0.674	0.437	0.884
152	0.616	0.871	0.768	0.617	0.918	0.500	0.948
153	0.531	0.94	0.822	0.706	0.939	0.353	0.985
157	0.910	0.948	0.867	0.898	0.836	0.745	0.939
164	0.593	0.954	0.840	0.735	0.945	0.581	0.972

167	0.632	0.881	0.871	0.750	0.992	0.600	0.996
175	0.700	0.917	0.802	0.737	0.867	0.636	0.912
176	0.774	0.884	0.803	0.793	0.812	0.605	0.915
179	0.775	0.909	0.860	0.926	0.794	0.641	0.964

AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristic; PPV, positive predictive value; NPV, negative predictive value; XGBoost, Extreme Gradient Boosting

6. Performance evaluation, results obtained with independent test set

Table S8 and Table S9 present the performance evaluation of both models. The models are trained on the development dataset and tested on the independent test set.

Table S8: *Performance of SVC models on the independent test set.*

Study Id	AUPRC	AUROC	Balanced accuracy	Sensitivity (recall)	Specificity	PPV (precision)	NPV
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	0.0141	0.791	0.459	0.00	0.917	0.00	0.993
10	0.608	0.932	0.875	0.958	0.792	0.280	0.996
11	0.407	0.759	0.721	0.694	0.749	0.358	0.924
12	0.428	0.995	0.995	1.00	0.990	0.500	1.00
20	0.173	0.791	0.763	0.600	0.926	0.130	0.992
33	0.829	0.975	0.883	1.00	0.767	0.271	1.00
35	0.572	0.94	0.883	0.882	0.884	0.375	0.99
37	0.465	0.927	0.857	1.00	0.714	0.197	1.00
38	0.359	0.854	0.763	0.778	0.747	0.226	0.973
39	0.361	0.939	0.868	1.00	0.736	0.147	1.00
42	0.0556	0.96	0.485	0.00	0.970	0.00	0.995
47	0.391	0.969	0.962	1.00	0.924	0.353	1.00
50	0.881	0.891	0.809	0.824	0.794	0.728	0.871
51	0.792	0.998	0.750	0.500	1.00	1.00	0.996
55	0.744	0.974	0.500	1.00	0.00	0.109	0.00
56	0.379	0.857	0.763	0.929	0.598	0.236	0.984
57	0.112	0.847	0.742	0.667	0.816	0.0500	0.994
58	0.765	0.977	0.917	0.929	0.905	0.361	0.995
62	0.841	0.856	0.797	0.734	0.860	0.744	0.854
69	0.627	0.981	0.890	0.800	0.979	0.444	0.996
72	0.453	0.979	0.944	1.00	0.888	0.222	1.00
78	1.00	1.00	1.00	1.00	1.00	1.00	1.00
81	0.489	0.782	0.695	0.750	0.640	0.341	0.912
86	0.318	0.883	0.749	0.938	0.561	0.143	0.991
89	0.509	0.990	0.988	1.00	0.975	0.500	1.00
90	0.963	0.999	0.997	1.00	0.995	0.833	1.00
94	0.435	0.909	0.811	0.933	0.689	0.192	0.992
95	0.809	0.893	0.810	0.857	0.764	0.629	0.920
96	0.450	0.881	0.820	1.00	0.640	0.330	1.00

101	0.860	0.997	0.997	1.00	0.995	0.833	1.00
102	0.365	0.653	0.631	0.421	0.841	0.235	0.926
109	0.513	0.957	0.959	1.00	0.918	0.455	1.00
114	0.233	0.950	0.931	1.00	0.861	0.200	1.00
115	0.535	0.888	0.798	0.750	0.846	0.429	0.957
116	0.549	0.981	0.663	0.333	0.993	0.5.00	0.986
118	0.677	0.841	0.739	0.737	0.740	0.509	0.885
122	0.522	0.857	0.773	0.800	0.746	0.293	0.966
123	0.911	0.988	0.926	0.923	0.928	0.600	0.990
124	0.834	0.932	0.807	0.710	0.904	0.710	0.904
125	0.68	0.925	0.590	0.188	0.992	0.750	0.906
126	0.551	0.943	0.750	0.500	1.00	1.00	0.991
127	0.243	0.936	0.854	0.750	0.958	0.375	0.991
128	0.410	0.860	0.500	1.00	0.00	0.101	0.00
129	0.428	0.984	0.966	1.00	0.931	0.300	1.00
131	0.532	0.890	0.710	0.500	0.920	0.111	0.989
132	0.0314	0.507	0.517	1.00	0.0348	0.0263	1.00
138	0.518	0.948	0.840	1.00	0.680	0.205	1.00
140	0.609	0.871	0.621	0.278	0.964	0.625	0.860
143	1.00	1.00	0.500	0.00	1.00	0.00	0.977
145	0.854	0.967	0.859	0.800	0.918	0.400	0.985
146	0.636	0.912	0.812	0.875	0.750	0.292	0.981
150	0.810	0.945	0.877	1.00	0.754	0.483	1.00
151	0.386	0.688	0.679	0.947	0.411	0.353	0.958
152	0.126	0.536	0.443	0.125	0.762	0.0625	0.873
153	0.656	0.979	0.659	0.333	0.984	0.500	0.969
157	0.54	0.681	0.609	0.562	0.655	0.474	0.731
164	0.444	0.937	0.897	0.889	0.905	0.500	0.987
167	1.00	1.00	1.00	1.00	1.00	1.00	1.00
175	0.871	0.955	0.705	0.500	0.909	0.667	0.833
176	0.918	0.973	0.500	1.00	0.00	0.250	0.00
179	0.86	0.950	0.758	0.6	0.917	0.750	0.846

AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristic; PPV, positive predictive value; NPV, negative predictive value; SVC, Support Vector Classifier

Table S9: *Performance of XGBoost models on the independent test set.*

Study Id	AUPRC	AUROC	Balanced accuracy	Sensitivity (recall)	Specificity	PPV (precision)	NPV
4	0.551	0.980	0.750	0.500	1.00	1.00	0.997
9	0.0251	0.839	0.498	0.00	0.997	0.00	0.994
10	0.854	0.967	0.885	0.792	0.979	0.760	0.982
11	0.534	0.842	0.758	0.694	0.823	0.442	0.930
12	0.655	0.995	0.832	0.667	0.997	0.667	0.997
20	0.372	0.884	0.683	0.400	0.967	0.182	0.989
33	0.958	0.996	0.846	0.692	1.00	1.00	0.974

35	0.653	0.949	0.769	0.588	0.949	0.476	0.967
37	0.443	0.949	0.811	0.692	0.930	0.409	0.977
38	0.470	0.900	0.704	0.444	0.963	0.533	0.948
39	0.530	0.87	0.695	0.400	0.991	0.667	0.973
42	0.00275	0.112	0.493	0.00	0.985	0.00	0.995
47	0.769	0.987	0.892	0.833	0.951	0.417	0.993
50	0.930	0.939	0.892	0.879	0.904	0.860	0.918
51	0.633	0.991	0.989	1.00	0.979	0.286	1.00
55	0.879	0.984	0.924	0.889	0.959	0.727	0.986
56	0.490	0.883	0.794	0.679	0.909	0.500	0.955
57	0.0933	0.804	0.657	0.333	0.981	0.200	0.990
58	0.886	0.99	0.853	0.714	0.992	0.833	0.984
62	0.891	0.895	0.802	0.835	0.769	0.667	0.894
69	0.339	0.978	0.792	0.600	0.983	0.429	0.992
72	0.803	0.988	0.973	1.00	0.947	0.375	1.00
78	1.00	1.00	0.998	1.00	0.996	0.941	1.00
81	0.746	0.89	0.719	0.475	0.963	0.760	0.881
86	0.672	0.932	0.894	0.875	0.912	0.438	0.989
89	0.758	0.984	0.888	0.800	0.975	0.444	0.995
90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
94	0.338	0.847	0.630	0.333	0.926	0.263	0.946
95	0.948	0.964	0.906	0.896	0.915	0.831	0.950
96	0.539	0.837	0.679	0.414	0.945	0.571	0.901
101	0.821	0.99	0.897	0.800	0.995	0.800	0.995
102	0.479	0.845	0.672	0.368	0.976	0.636	0.930
109	0.786	0.939	0.847	0.700	0.993	0.875	0.980
114	0.226	0.943	0.665	0.400	0.931	0.167	0.978
115	0.719	0.942	0.848	0.750	0.946	0.682	0.961
116	0.794	0.991	0.823	0.667	0.979	0.400	0.993
118	0.761	0.858	0.811	0.737	0.885	0.700	0.902
122	0.595	0.913	0.821	0.800	0.842	0.400	0.970
123	1.00	1.00	1.00	1.00	1.00	1.00	1.00
124	0.881	0.942	0.893	0.871	0.915	0.771	0.956
125	0.685	0.930	0.699	0.438	0.960	0.583	0.931
126	1.00	1.00	0.991	1.00	0.982	0.500	1.00
127	0.394	0.773	0.862	0.750	0.975	0.500	0.991
128	0.552	0.877	0.685	0.417	0.953	0.500	0.936
129	0.683	0.843	0.833	0.667	1.00	1.00	0.990
131	0.138	0.715	0.745	0.500	0.990	0.500	0.990
132	0.0893	0.658	0.491	0.00	0.983	0.00	0.974
138	0.869	0.99	0.959	1.00	0.918	0.500	1.00
140	0.623	0.866	0.789	0.722	0.855	0.520	0.934
143	0.0567	0.820	0.500	0.00	1.00	0.00	0.977
145	0.733	0.984	0.966	1.00	0.932	0.500	1.00
146	0.445	0.914	0.846	0.75	0.941	0.600	0.970
150	0.884	0.966	0.86	0.786	0.934	0.733	0.950
151	0.522	0.703	0.636	0.737	0.536	0.350	0.857

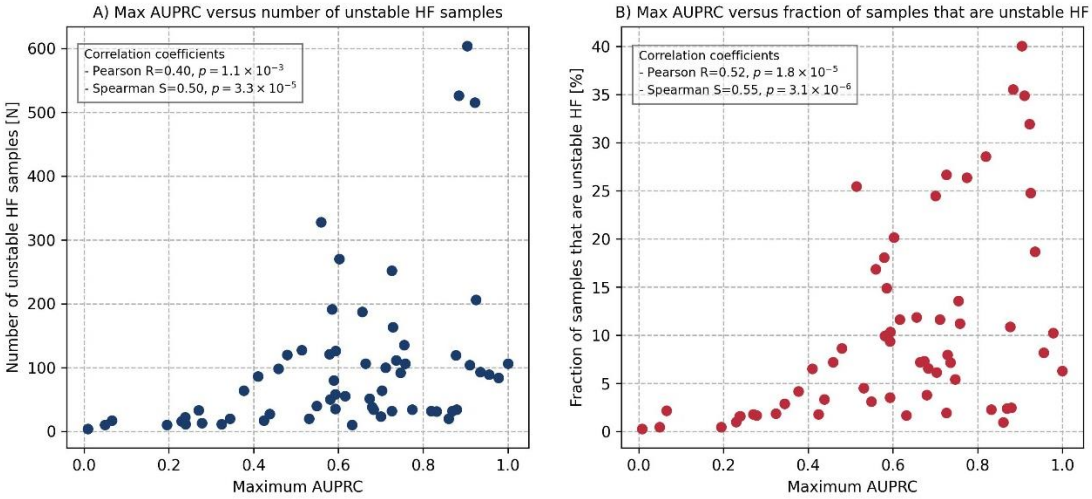
152	0.354	0.760	0.773	0.625	0.921	0.500	0.951
153	0.156	0.885	0.635	0.333	0.938	0.200	0.968
157	0.797	0.841	0.744	0.625	0.862	0.714	0.806
164	0.339	0.902	0.625	0.333	0.917	0.300	0.928
167	1.00	1.00	1.00	1.00	1.00	1.00	1.00
175	1.00	1.00	1.00	1.00	1.00	1.00	1.00
176	0.861	0.933	0.767	0.6	0.933	0.750	0.875
179	0.878	0.883	0.858	0.8	0.917	0.800	0.917

AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristic; PPV, positive predictive value; NPV, negative predictive value; XGBoost, Extreme Gradient Boosting

7. Relation between model performance and class distribution

Figure S3 illustrates the relation between the performance of the models and the class distribution in the dataset.

Figure S3: *Relation between model performance and class distribution in the dataset. The maximum AUPRC is defined as the highest AUPRC achieved through leave-one-out cross-validation, either with the XGBoost model or the SVC.*



AUPRC, area under precision-recall curve; HF, heart failure; SVC, Support Vector Classifier; XGBoost, Extreme Gradient Boosting