Intelligence-Based Medicine 5 (2021) 100044

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

HeartPredict algorithm: Machine intelligence for the early detection of heart failure

Habiboulaye Amadou Boubacar [a,*], Mehdi Rahim [a], Gisele Al-Hamoud [b], Spyridon Montesantos [b], Cecile Delval [b], Sylvie Bothorel [b], Juan Fernando Ramirez-Gil [b]

[a] Computational & Data Science, Air Liquide R&D - Innovation Campus Paris, 78350, Les Loges-en-Josas, France
[b] Air Liquide Santé International, 28 Rue d'Arcueil, 94250, Gentilly, France

ARTICLE INFO

Keywords:
Heart failure
Remote patient monitoring
Machine learning

ABSTRACT

Heart failure (HF) is among the leading causes of death. Its prevalence is increasing dramatically causing considerable healthcare costs as well. Remote patient monitoring (RPM) is one of the solutions to enhance patient well-being. Taking advantage of new advances in artificial intelligence and RPM digital platforms, we propose HeartPredict: a novel machine learning algorithm for early detection of HF episodes and associated unplanned hospitalizations. The algorithm relies on a telemonitoring dataset from the largest European clinical study on HF. It uses balanced random forests on significant features extracted from patient weight time series, symptoms and socio-demographics. We benchmark HeartPredict with rules from medical guidelines and state of the art machine learning models. HeartPredict has better performance in terms of sensitivity (72% vs. guidelines: 53%) and specificity (94% vs. guidelines: 84%), with AUROC = 0.8. In addition, we introduce precocity as a new criterion to evaluate the ability of the algorithm to detect early HF risks, and therefore, enabling proactive medical actions. Finally, we show that HeartPredict prediction scores are consistent with HF risk levels, thereby limiting the risk of non-detection for patients.

1. Introduction

Heart Failure (HF) is a pathophysiological state with an abnormal cardiac function [1], such that the heart muscle does not pump blood properly. With high prevalence in elderly people and a significant increase in middle-aged people, heart failure is projected to rise drastically by 46% according to a report of the American Heart Association (AHA) [2].

HF is a leading cause of mortality. It represents a burden of healthcare costs due to high rate of unplanned hospital admissions, long stays at the hospital, and expensive care.

Recent advances in telemonitoring technologies and data science pave the way towards more effective remote patient monitoring (RPM) Herold et al. [3]. These RPM solutions propose personalized care programs [4]. They help to prevent patient health worsening thanks to timely and proactive interventions [5]. Predictive models are key elements to anticipate patient risks, enabling a care continuum from hospital to home and vice-versa. However, there are many challenges that make the success of machine learning (ML) based RPMs deployment difficult. Access to data and sufficient medico-economic evidence are issues that hamper the implementation and the adoption of ML-based RPM tools.

Recently, Air Liquide Healthcare launched an RPM solution for the remote monitoring of patients suffering from heart failure. The workflow includes connected medical devices and a monitoring platform. Remote monitoring relies on three stages:

(1) The patient's vital signs are recorded and transmitted using a connected weight-scale to a remote server.
(2) The measurements are processed by an algorithm that generates alarms in case of abnormal variations.
(3) The alarms are categorized in the monitoring platform where appropriate medical actions are taken. Typical actions include recommendations from a nurse, treatments by a doctor, or hospitalizations.

In this paper, we introduce HeartPredict: a machine learning algorithm to improve the performance of Air Liquide Healthcare RPM

* Corresponding author.
E-mail address: habiboulaye.amadou-boubacar@airliquide.com (H. Amadou Boubacar).

https://doi.org/10.1016/j.ibmed.2021.100044
Received 27 January 2021; Received in revised form 14 September 2021; Accepted 21 October 2021
Available online 28 October 2021

solution. Our goal is to improve early detection of HF episodes, while reducing the number of false alarms. The proposed algorithm uses balanced random forest on patient time series representing weight, symptoms and socio-demographics. We demonstrate the effectiveness of HeartPredict through an intensive evaluation over the OSICAT dataset [6,7], a large study on monitoring for patients with heart failure by tele-cardiology. Before describing the HeartPredict algorithm and the results, preliminary studies of HF telemonitoring will first be presented.

2. Heart failure telemonitoring

Daily monitoring of weight gain is recommended by both the Heart Failure Association of America (HFSA) and the European Society of Cardiology (ESC). Indeed, cardiac decompensation is often warned by fluid retention with the appearance of edema. Many studies report that body weight gain is one of the most important predictors of heart failure [8–10]. For this, clinical experiments rely on rule-based alarm tools. The idea is to compare the weight gain within a range during a time interval. Among the rules of thumb of weight-based HF telemonitoring, McMurray John et al. [11] consider that cardiac patients experiencing an increase of body weight by 2 kg in 3 days should have an increase of their diuretic dose. HFSA guidelines [12] recommend to monitor an increase of weight from 0.9 kg in 1 day, or more than 1.8 kg over a week.

Although weight gain is a good predictor of heart failure, there is no consensus on the choice of the weight variation threshold. Also, it is not clear if monitoring other vital signs can help to increase the accuracy of HF prediction. Some studies use non-invasive indicators such as heart rate, oxygen saturation [13], blood pressure [14] and bio-impedance [10]. Abraham et al. [15] use invasive hemodynamic monitors. Such monitors can improve overall accuracy, but it involves a higher cost with more discomfort for the patient. Recent studies combine all these indicators to generate heart failure alarms. Larburu et al. [13] and Gilliam et al. [16] use body weight and heart rate indicators such as Heart Rate Variability (HRV), Standard Deviation of the (5 min) Average Normal-to-Normal intervals (SDANN). They report an improved performance of HF alarms by adding patient symptoms like fatigue, dyspnoea, edema, cough, fever.

Overall, there are two main categories of algorithms for HF prediction: rule-based approaches and machine learning approaches. We describe in the following sub-sections these two categories.

2.1. Rule-based HF prediction methods

HF alarms are generated from recorded vital signs using rules and adjusted thresholds. Specific rules are implemented to trigger alarms if the variation in body mass exceeds a threshold over a period of time. There is no consensus on the best weight variation threshold, as different works advocate for different thresholds. For example, +2 kg weight gain over 2–3 days as in the European guidelines [17], or +1.36 kg weight gain over 1 day [18]. Rules are often stacked to reduce the risk of missing HF episodes. This results in a high false positive rate. Typically, the following rules from medical guidelines are used as a reference for comparison:

- R1: +2 kg in 5 days.
- R2: ±10 kg in 45 days.
- R3: +3 kg in 2 days.

A more exhaustive list of weight-based rules is reported in Ref. [19] where more than twenty scientific papers are reviewed. In Ref. [9], a patented set of rules called HeartPhone is used to show how an individualized approach can generate more specific alarms. Thresholds are calibrated for each patient at different periods. A dynamic monitoring is done through a moving average over weight measurements. Moving average is sensitive to outliers and missing data which can yield unreliable alarms. This makes it hard to define gold-standard rules for a reliable monitoring of heart failure.

2.2. Machine learning-based HF prediction methods

Binary classification is a suitable setting to address the prediction of HF-related events from time series data. HF prediction is formulated as a classification of time-series of body weight records, where HF-related events are rare. This rises two main challenges: i) How to extract informative characteristics from time series; ii) How to accurately train ML algorithm on unbalanced dataset.

2.2.1. Time-series features

Statistical models on time-series are usually used to forecast future states of a system (e.g. weather, sales, prices, heartbeats, …). We distinguish two approaches:

1. Classification on raw time-series, where time-series are used directly to build a predictive model. For instance, Auto-Regressive (AR) models predict the future from past time-series. This approach requires both significant temporal resolution and data history. It is usually used for regression rather than for classification.
2. Classification on time-series' characteristics, where features are extracted from time-series. Then, a ML model is used to forecast the future outcome or to predict a specific class.

The second approach is more relevant to address our HF detection problem. Indeed, HF telemonitoring consists of only one (sometimes less than one) measurement per day. More informative features are extracted from weight records, symptoms, or other static information (age, sex, …) by using aggregation methods. To better capture the variation of body-weight, various statistics including mean, variance, trend, Hjorth parameters [20] are extracted on a time-varying sliding window. The statistics computed are then aggregated to create the inputs for the predictive model.

2.2.2. Unbalanced classification

HF-related events occur rarely when monitoring cardiac patients. This causes imbalanced classes where the positive class (HF event) represents a small fraction (e.g. 10%) of the whole dataset. Classification models become biased; for example, a classifier with a constant answer on a dataset with 90%–10% class proportions reaches 90% accuracy, but fails to detect the minority class. Class weighting may not be sufficient when dealing with highly imbalanced datasets. In such settings, oversampling and undersampling approaches can give better classification results. Oversampling generates synthetic samples following the distribution of the minority class to balance the dataset, such as Synthetic Minority Over-sampling Technique (SMOTE) [21], or Adaptive Synthetic sampling (ADASYN) (Haibo [22]. Undersampling removes samples from the majority class. This can be done randomly or with prior information such as TOMEK links [23]. One should note that oversampling could lead to overfitting when dealing with noisy data, while undersampling can be inefficient when the minority class is very small. Finally, undersampling the majority class combined inside an ensemble method is a promising approach (Xu-Ying [24]. This creates a bagged classifier from multiple weak classifiers. Each classifier is trained over iteratively balanced datasets generated with undersampling. The majority class subsamples may overlap or not. This yields a classifier trained on the whole dataset without any synthetic data, and can achieve better accuracy.

2.2.3. Machine learning methods

Recent works propose ML methods for HF detection at early stage. These new approaches can handle multi-dimensional indicators from heterogeneous records. Complex patterns and correlations are extracted from patient profiles and vital sign variation enabling more accurate predictions. Moreover, ML algorithms generate predictions as

continuous scores, often between 0 and 1, that represent how likely an HF alarm is positive or not. Gilliam et al. [16] use logistic regression with feature selection to generate HF alarms. Such a model is easily interpretable. However, it can be less accurate if the relationship between the features and the target is nonlinear.

Larburu et al. [13] compare different ML models (random forests, naive Bayes) to predict HF episodes from body weight variation, heart rate, blood pressure, oxygenation, and symptoms data.

It is hard to compare these methods since they use different datasets, with different number of patients and different follow-up duration. Validation schemes are different and are often within patient, i.e. same patient at different periods in train and test subsets. This makes the generalization of ML models less reliable.

Recent works rely on deep neural networks that are able to ingest raw data. Recurrent Neural Networks are specifically designed for time-series. They use the sequential relationship between timepoints to process the output Rumelhart et al. [25]. For instance, Long-Short Term Memory (LSTM) [26] networks are successfully used for speech recognition [27] and language processing [28]. In Ref. [29], the authors show that LSTM can predict patient diagnosis from electronic health records. However, deep learning models are data hungry in the training stage: they need a large amount of data with a fine temporal resolution.

2.3. Evaluation metrics of HF prediction methods

The evaluation helps to quantify the accuracy of the prediction of HF events. Common ML performance metrics and criteria are used to evaluate RPM tools depending on the expected outcomes. In HF context, we define an HF event as Positive P when a decompensation or urgent hospitalisation occur. Negative N stands for a healthy state when there is no decompensation nor hospitalisation. Accordingly, a True Positive TP is an HF event correctly detected by the algorithm. A False Positive FP is a healthy state incorrectly detected by the algorithm as an HF event. A True Negative TN is a healthy state correctly detected by the algorithm. A False Negative FN is an HF event not detected by the algorithm.

The evaluation metrics for HF prediction are:

2.3.1. Sensitivity

SEN is the true positive rate, i.e. the percentage of positives correctly detected (TP) by the model. It is also called the recall or the proportion of HF events detected by an algorithm.

$$\mathrm{SEN} = 100 \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

2.3.2. Specificity

SPE is the True Negative Rate, i.e. the percentage of negatives correctly detected (TN) by the algorithm, i.e., the proportion of healthy states correctly identified

$$\mathrm{SPE} = 100 \times \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2}$$

2.3.3. False alarms

HF studies report the number of false positives (FP). This number is normalised by the number of patients and their respective follow-up periods which yields a False Alarm rate (FA). There are two ways to compute FA:

$$\mathrm{FA}^{(1)}_{\mathrm{patient\ /\ year}} = \frac{365}{\#\mathrm{patients}} \sum_{\mathrm{patient}} \frac{\mathrm{FP(patient)}}{\mathrm{Follow-up\ \ days(patient)}} \tag{3}$$

$$\mathrm{FA}^{(2)}_{\mathrm{patient\ /\ year}} = 365 \times \frac{\sum_{\mathrm{patient}} \mathrm{FP(patient)}}{\sum_{\mathrm{patient}} \mathrm{Follow-up\ \ days(patient)}} \tag{4}$$

In this paper, we use $\mathrm{FA}^{(2)}$ since it is a more stable indicator, in particular for patients with short follow-up periods.

2.3.4. Area under the ROC curve (AUC)

The area ROC (Receiver Operating Characteristic) curve represents the variation of model sensitivity and specificity according to prediction thresholds. This helps the practitioner to select the threshold that corresponds to the desired trade-off between sensitivity and specificity.

2.4. Literature review

From this literature review on HF telemonitoring, we observe that there is a lack of clarity about the most predictive vital signs to use for HF telemonitoring. However, body weight variation seems to be an important feature commonly used in studies to monitor for HF, since there is a trade-off between simplicity and efficiency [19]. In terms of methods for HF prediction, ML is a promising approach when there is sufficient data for training Larburu et al. [13]. ML algorithms can extract hidden correlations in data and can output probability scores to predict the risk level of HF. Regarding the performance evaluation, the main metrics used in clinical context are sensitivity, specificity, and false alarm rates. To the best of our knowledge, there is no evaluation scheme that assess how early can an algorithm detect a HF event for a patient. Indeed, if an alarm is raised too late, the medical actions would not be efficient to prevent the hospitalisation. We introduce in the following sections such an evaluation.

3. Data: the OSICAT study

Our work relies on data from the clinical trial OSICAT [6,7]. The study involves telemonitoring data collected from patients with chronic HF. A protocol was set to reuse anonymised data of the telemonitored patients after their consent. It was registered at the French institute of health data (INDS, Institut National des Données de Santé).

3.1. Patient description

The patients included in this clinical study satisfy the following inclusion criteria: i) Adults aged over 18 years old; ii) Hospitalised for at least one cardiac decompensation during the last twelve months; iii) Patients with a landline or access to the General Packet Radio Service (GPRS) network at home.

In addition to these criteria, our research work has been focused on patients in the RPM (i.e. intervention) group with a minimum of 1 month follow-up and who are not opposed to use of their data.

Table 1 gives some statistics about the patients used in our work.

The patients were telemonitored up to 18 months between June 2013 and May 2017, as highlighted in Fig. 1. Data were recorded from weight scale and a questionnaire of 8-symptom.

3.2. Patient measurements

Patients measure their body weight daily using a connected weight scale. They send through a tablet the assessment of 8 symptoms: dyspnoea, orthopnea, cough, edema, fatigue, fever, palpitations and weakness. Fig. 2 shows that most of the patients are compliant with the telemonitoring, as they send their records almost everyday.

Table 1
Patient characteristics.

Patient Characteristic	Statistics
Age (mean ± std)	69 ± 12 y.o.
Sex (women/men)	93/279
Follow-up period (mean ± std)	421 ± 166 days
Patients with at least one HF hospitalisation	168 (45%) patients
Number of HF events	412 events
Number of patients	372 patients

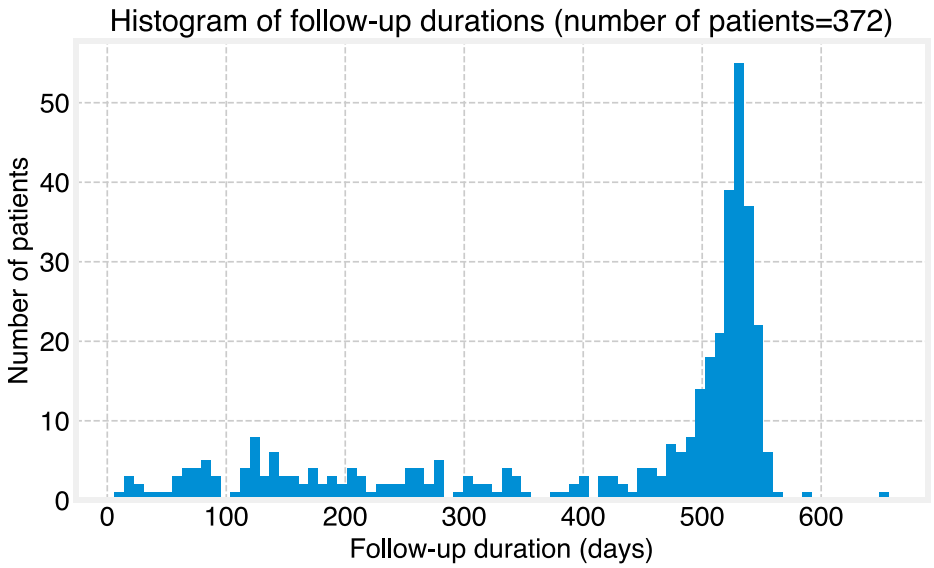

Fig. 1. Patient follow-up duration histogram - The histogram shows the distribution of the number of patients according to the duration of their respective follow-up periods. Most of the patients have more than 500 days of records. This means that the average follow-up duration is around 1.5 year.

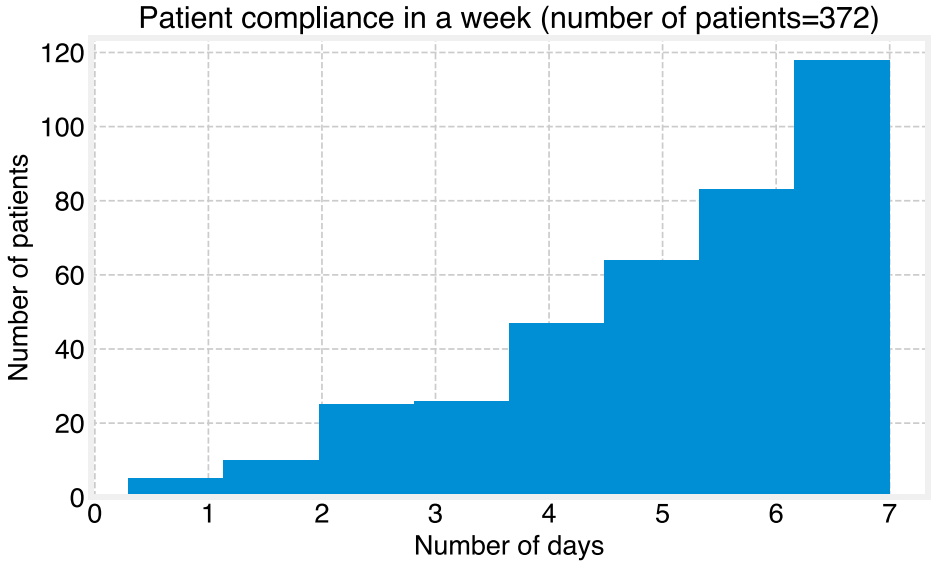

Fig. 2. Patient weekly compliance - The histogram shows the number of patients with daily records over one week. Most of the patients send records more than 4 days per week.

3.3. Heart failure events

Heart failure events are related to any risk of cardiac decompensation. Full description of heart failure etiology can be found in Ref. [7]. To prevent hospitalisation, alarms should be raised before this event occurs. Fig. 3 explains this principle.

We consider two kinds of heart failure events:

3.3.1. Unplanned hospitalisations for heart failure causes

Fig. 4 shows the distribution of the hospitalisations. The dataset includes 412 HF-related hospitalisations of 168 patients.

3.3.2. Cardiac decompensation with medical actions

Physicians advocate for treatments when signs of HF appear. These treatments aim to limit the aggravations of the signs. For example, treatment with diuretics can be advocated.

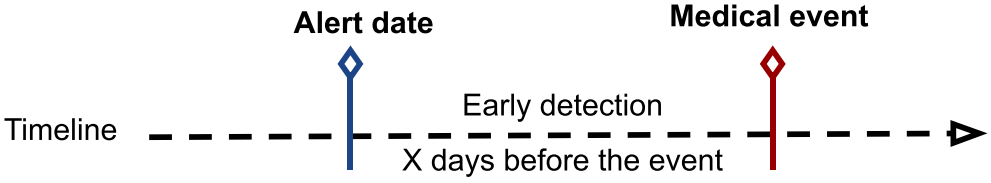

Fig. 3. Timeline of alarms and medical events for remote patient monitoring.

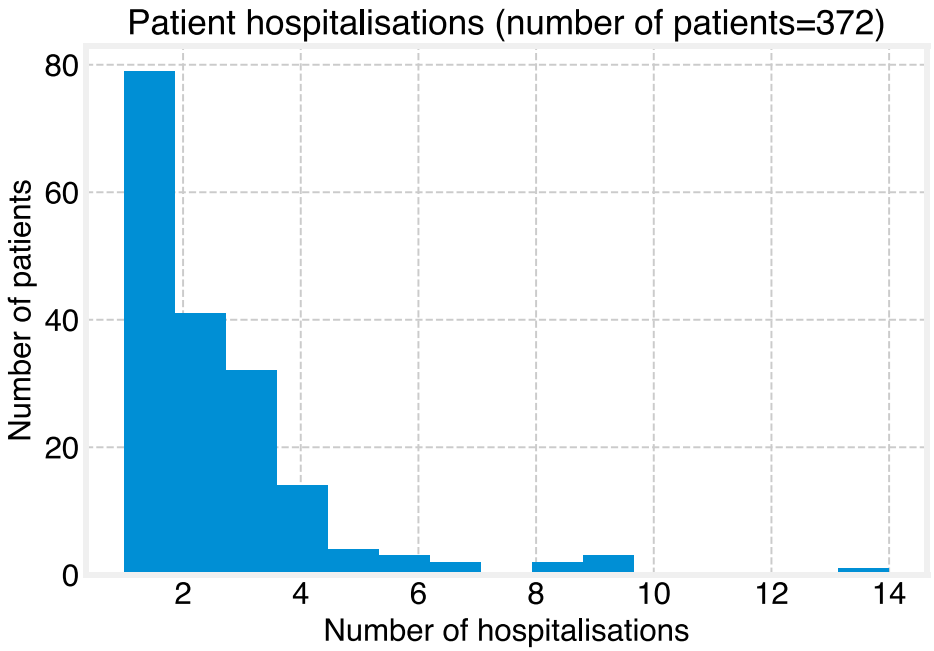

Fig. 4. Patient hospitalisations - The histogram shows the number of patients with one or more hospitalisations. 168 patients had at least one hospitalisation, resulting in 412 hospitalisation events.

3.3.3. Other medical events

The dataset contains logs of the remote interventions of the nurses. Typical follow-up interventions include surveillance, hygiene recommendations. Also, actions from physicians are recorded, such as health checks and non-HF related treatments.

4. HeartPredict algorithm for heart failure prediction

We present in this section HeartPredict algorithm, including the representation of the telemonitoring data, the feature engineering, and the classification. Fig. 5 depicts the whole pipeline of HeartPredict. Our approach relies on patient telemonitoring data that consists of multivariate time series of weight, self-reported symptoms and medical events. Then, time series are aligned with medical events in order to extract relevant features and the target to be predicted (HF events in our case). Finally, a balanced random forests classification model built on these features and targets yields HF risk score to predict future HF episodes.

4.1. Data representation

In our study, the definition of medical events is done with the support of four independent cardiologists, members of OSICAT adjudication committee and managed by the Cardiovascular European Research Center (CERC). There are three categories of events related to patients health state:

- (P0, P1): medical events related to HF decompensations
- (P2, P3, P4): risk of health worsening to be checked (unknown)
- (P5): stable condition i.e. no deterioration of patient health

Table 2 shows all the medical events of the 372 patients included in this study. Medical events are ordered according their severity. Label P0 stands for the most severe medical state (HF-related hospitalisation) while P5 refers to stable health-state.

4.1.1. Target definition

Our goal is to predict events related to HF hospitalisations and HF related treatments. Thus, the target to predict includes labels P0 and P1. In order to take suitable medical actions earlier, the target to predict is

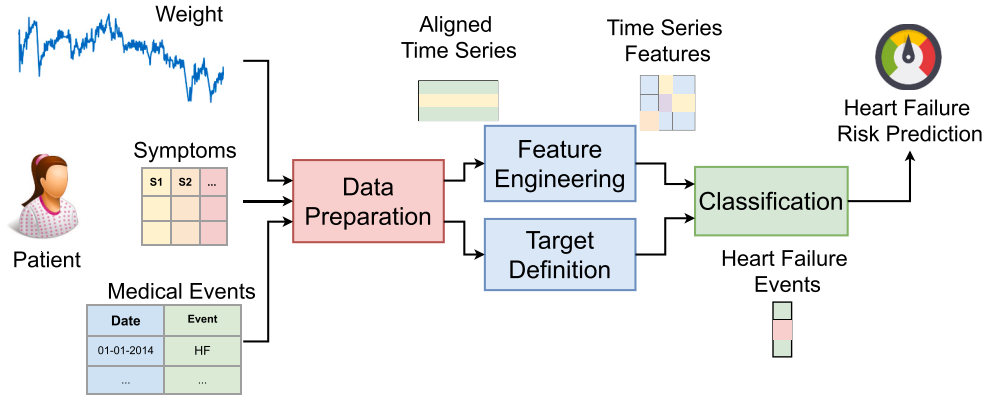

Fig. 5. HeartPredict Algorithm Workflow. Input telemonitoring data consists of multivariate time series of weight, self-reported symptoms and medical events. Time series are aligned to extract relevant features and target. Finally, a classification model predicts HF risk score.

Table 2
Reference Database. HeartPredict is trained to predict HF events (P0, P1).

Event tag	Event description	Nb of events
P0	Urgent HF-related hospitalisation	254
P1	Treatment for HF	158
P2	Other treatment	505
P3	Health check by physicians	474
P4	Surveillance or hygiene by call-center nurses	910
P5	Healthy (stable) state	18146
	Total	20447

positive (target = 1) if the event happens within a time horizon of seven days. Otherwise, the target is negative (target = 0) when there is no HF-related event within a horizon of 7 days.

4.1.2. Feature engineering

This step defines the inputs of the ML model. For each patient, medical records (time-series) are segmented into sub-periods of N consecutive days. In this study, we set N = 55 to ensure the benchmark of all the medical guideline rules (e.g. rule R2 in as mentioned in section 2.1). Each sub-period followed by a medical event is then tagged positively (target = 1). Sub-periods corresponding to the stable state are tagged negatively (target = 0). We apply a lag from 7 to 14 days between each patient consecutive sub-periods (Fig. 6). This avoids any overlap between stable and degradation sub-periods or any contamination between medical events.

Following the data representation steps, a reference dataset is constructed from patients daily tele-transmitted weight and symptoms records. The first day of each interval is the beginning of the telemonitoring in the corresponding sub-period whilst the last day corresponds to the day of a medical event if occurred. The dataset is then well formatted for training ML algorithms with target representation for classification setting:

- Y = 1 stands for HF decompensations from medical events P0 and P1.
- Y = 0 stands for the healthy conditions corresponding to stable state of patients (P6) or no proven worsening (P2, P3, P4, P5).

Another data representation option would be to include uncertain situations (P2, P3, P4) as positive targets. Our experiments have shown that this increases the number of false alarms considerably. This implies a considerable workload of nurses without a significant improvement in the prediction of HF-related events. In section 6.1, we show that the proposed predictive algorithm yields predictions that are close to the medical event levels.

4.2. Performance evaluation

An essential part of this work is to ensure that the model gives reliable and generalisable results. To evaluate ML models, different performance metrics are used. In section 2.3, we introduced the evaluation metrics commonly used in healthcare. We use sensitivity, specificity and false alarms to assess the performance of our predictive models. These quantitative metrics emphasize a trade-off between detecting HF-related events correctly and minimizing HF-related false alarm rates. In addition, we introduce a new indicator, called Precocity of detection. This indicator evaluates how earlier an algorithm detects HF-related events. It corresponds to the number of days the first alarm is raised before the actual HF-related event. For this, we evaluate the performance of algorithms shifting the date of detection from 0 to 7 days before each HF event in the testing set.

4.3. Validation schemes

We evaluate the model through two validation schemes:

- Held-out test set validation; where the reference dataset is split into a train subset (75% of the whole dataset) on which the model is built. The test subset (25% of the whole dataset) is used as a ground-truth on which the overall performance is measured.
- K-fold cross-validation; here, the initial train subset (75% of the whole dataset) is split with a stratification, in order to have homogeneous subsets, resulting in subsets having the same proportions of the target. As shown in Fig. 7, we group the samples (features) by patient during the split. This means that patients of the test subset are different from those in the train subset. Such an approach ensures a better generalisation of the fitted classifier.

5. Results and discussion

We present in this section the evaluation of our model called HeartPredict. First, we benchmark various ML models in order to select the best one. Then, we compare our ML approach to rules from medical guidelines. Finally, we assess the relevance of the prediction scores produced by our model.

5.1. Benchmark of classifiers

We compare the performance several ML algorithms for HF prediction. Table 3 presents the AUC scores as results of on validation data. Ensemble-based models achieve better performance for HF prediction compared to linear models. An adequate trade-off would be to keep

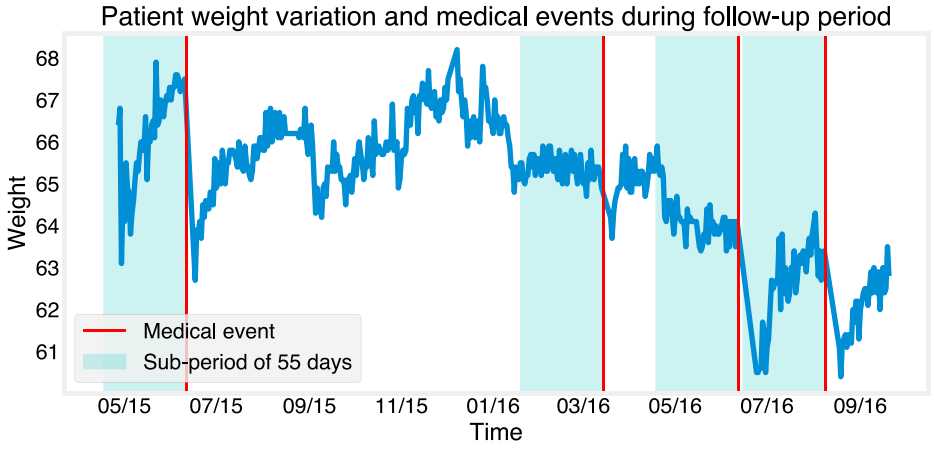

Fig. 6. Patient follow-up splitting - The whole time-serie is equally divided into sub-periods of 55 days (in blue). The last day of each period corresponds to a medical event (red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

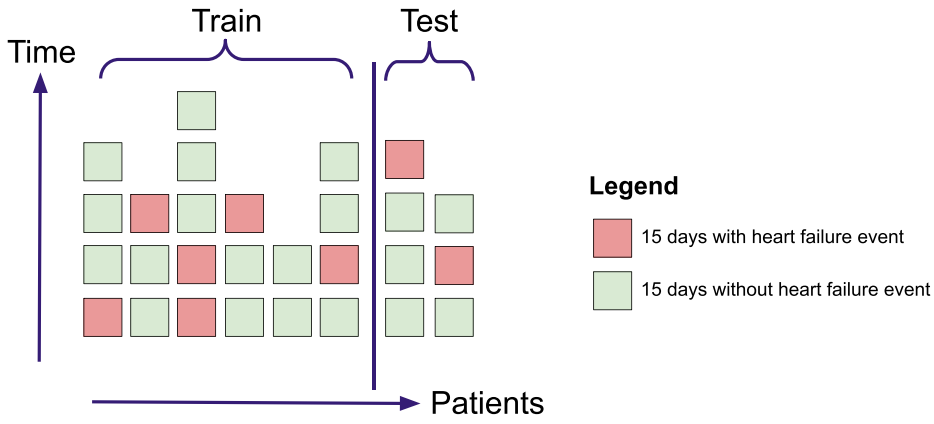

Fig. 7. Cross-validation scheme - In order to perform a generalisable model selection, we split the dataset into train/test folds by keeping the proportions of the target (stratification), and by isolating patients between train and test. Here, each patient (rows) has multiple time-series (columns).

Table 3
Benchmark of classifiers in terms of ROC AUC. Ensemble models (Gradient Boosting, Balanced random forests) outperform linear classifiers and kernel SVM. Balanced random forests yield well-calibrated predictions' probabilities.

Model	Cross-val. AUC (mean ± std)	AUC on test set
Linear SVM	0.72 ± 0.03	0.75
RBF Kernel SVM	0.73 ± 0.03	0.75
Polynomial Kernel SVM	0.69 ± 0.04	0.71
Logistic Regression	0.73 ± 0.03	0.72
Nearest Neighbours	0.65 ± 0.02	0.68
Decision Tree	0.70 ± 0.04	0.70
Neural Net	0.77 ± 0.01	0.78
Naive Bayes	0.70 ± 0.01	0.70
Gradient Boosting Classifier	0.80 ± 0.03	0.82
Balanced Random Forests	0.80 ± 0.03	0.82

interpretable and accurate model, in particular in the healthcare domain. Indeed, transparent rules can be derived from the decision tree algorithm. However, such classifiers have low performances in terms of sensitivity and specificity. From an operational stand point, this means that there is a higher risk of missing HF events without lowering nurse workload. From this comparative study, the Balanced Random Forests classifier outperforms all the other classifiers without requiring fine-tuning. Only 200 trees are used to aggregate weak estimators. The strategy of resampling data in the bootstrap samples is used to balance the positive and negative classes in the Random Forests and perform with well-calibrated predictions' probabilities.

5.2. HeartPredict performance

HeartPredict algorithm, implemented with balanced random forests, is compared to alarming rules derived from medical guidelines in the literature. Fig. 8 shows the ROC curve of HeartPredict computed using the probabilities' outputs on validation data. The sensitivity and specificity of guidelines rules are also reported corresponding to a fixed decision point. We also show sensitivity and specificity scores from the literature, although these results come from different datasets.

HeartPredict achieves better performance compared to algorithms from the state-of-the art on heart failure detection. It is slightly better than the HeartPhone [9] algorithm. However, this comparison is provided for information only, since we do not have access to the dataset used in the related studies. More importantly, HeartPredict surpasses any guideline rule found in the literature, on the same dataset. Indeed, Table 4 shows that for a similar sensitivity (SEN) level (53%), the number of false alarms is decreased by 51% (guidelines: 9.4 to HeartPredict: 4.6 false alarms per patient per year). For similar specificity (SPE) (84%), HF prediction is improved by 19 points of sensitivity (Guidelines: 53% vs. HeartPredict: 72%). That means a gain of 19% of HF prediction. Further analysis in cross-validation setting confirms this observation. Fig. 9 shows sensitivity and specificity variations over 10 cross-validation folds. We observe that HeartPredict outperforms significantly rules from guidelines performance. HF prediction is improved by 13 points of sensitivity in average (Guidelines: 56% vs. HeartPredict: 69%).

We did not keep socio-demographics in HeartPredict. Indeed, although socio-demographic indicators (age, gender, …) could have a potential impact on HF prediction as mentioned in Ref. [30], our experiments show that these indicators do not improve HeartPredict performance. Performance difference between HeartPredict trained on weight only and with symptoms is not significant. This does not imply that symptoms do not bring valuable information. Our experimentation suggests that the information provided by important symptoms (e. g. edema) is already captured by features extracted from weight. Home care programs could then increase patient experience by simplifying the measurement protocol. For example, the self-reporting of symptoms can be adapted. In this perspective, our research team propose a smart questionnaire which minimizes the interaction using Reinforcement Learning techniques [31].

5.3. Heart failure detection precocity

In addition to the sensitivity and the false alarm rate at day 0 of the HF event, the performance of HeartPredict is compared to rules from guidelines up to 7 days before the HF event. For this purpose, the alarms are computed each day from 7 to 0 days before the event. Fig. 10 shows the sensitivity and the false alarm rate according to all time windows from 0 to 7 days. For example, at 3 days before the event, we look if there is an alarm in the window $[-7, -3]$ days before the event. The results show a better early detection of medical events with HeartPredict. Rules from guidelines detection sensitivity decreases drastically compared to HeartPredict when varying the horizon of prediction, while false alarm rates variation for HeartPredict is slower.

There is a lack of evidence regarding the best choice of prediction window in the state of the art. Our algorithm shows a good capability of prediction in this window with more than 50% of the events detected 3 days before they occur. Even if such anticipation is important for enabling effective medical actions, the choice of optimal prediction window still remains an open question.

5.4. Optimal threshold selection

HeartPredict outputs probability scores related to cardiac decompensation. Alarms are generated according to a defined threshold. The ROC curve in Fig. 8 shows that there is an interval in which HeartPredict

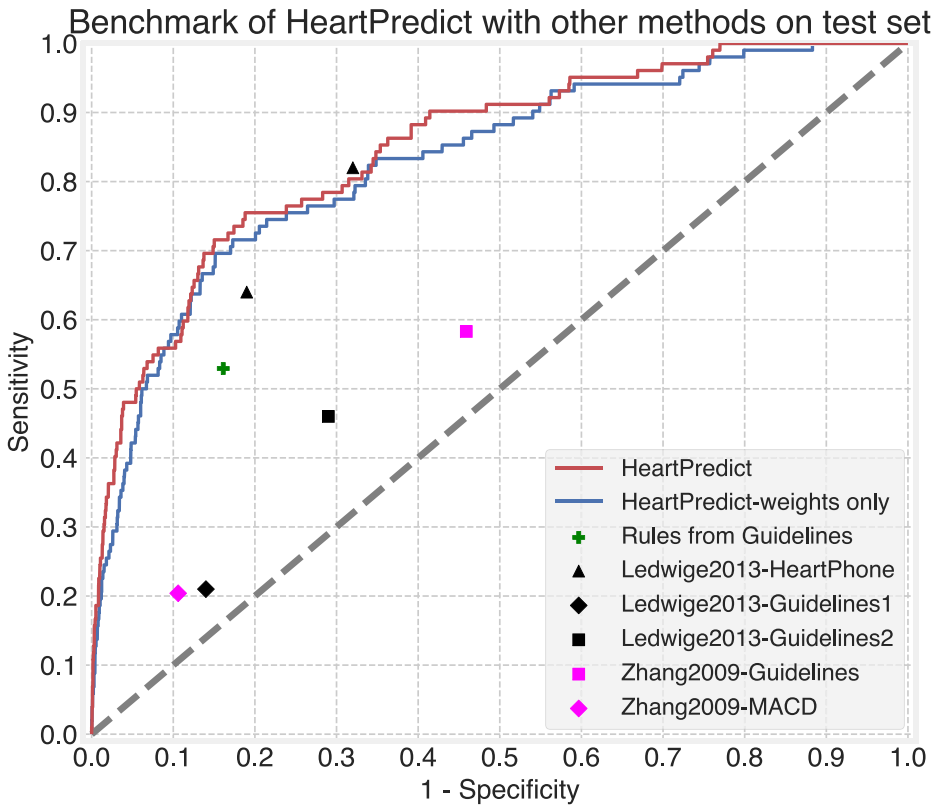

Fig. 8. HeartPredict evaluation with ROC curve on held out test set - The red (weights + questions) and the blue (weights only) curves show better sensitivity and specificity compared to rules and state of the art studies. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Performance results (up to 7 days) on test set of HeartPredict with guidelines at equivalent sensitivity and specificity thresholds. Rules from guidelines are the state of the art rules of thumb. HeartPredict W is trained with weight records only. HeartPredict WQ is trained weight and symptoms records.

Method	Sensitivity	Specificity	False Alarms
Rules from guidelines	53%	84%	9.4
HeartPredict WQ@SEN = 53%	53%	94%	4.6
HeartPredict WQ@SPE = 84%	72%	84%	9.8
HeartPredict WQ@SEN = 63%	63%	88%	7.5
HeartPredict W@SEN = 53%	53%	90%	5.3
HeartPredict W@SPE = 84%	68%	84%	9.9
HeartPredict W@SEN = 63%	63%	86%	7.9

performance is better than rules from guidelines in both sensitivity and specificity. If the decision threshold is selected at the upper bound (0.69), we have similar sensitivities of between rules and HeartPredict (53%) but HeartPredict specificity is better (94% vs 84%). One the other hand, if the decision threshold is selected at the lower bound (0.62), we have similar specificity of rules and HeartPredict (84%) but HeartPredict sensitivity is better (72% vs 53%). The range is selected such that model sensitivity and specificity are better than rules from guidelines. Fig. 11 shows that the optimal range is from 0.64 to 0.69.

6. HeartPredict complementary analyses

6.1. Analysis of predictive scores for HF detection

As mentioned in section 3, Table 2 describes the 3 categories of situations characterizing the patient's health conditions (from P0 to P5).

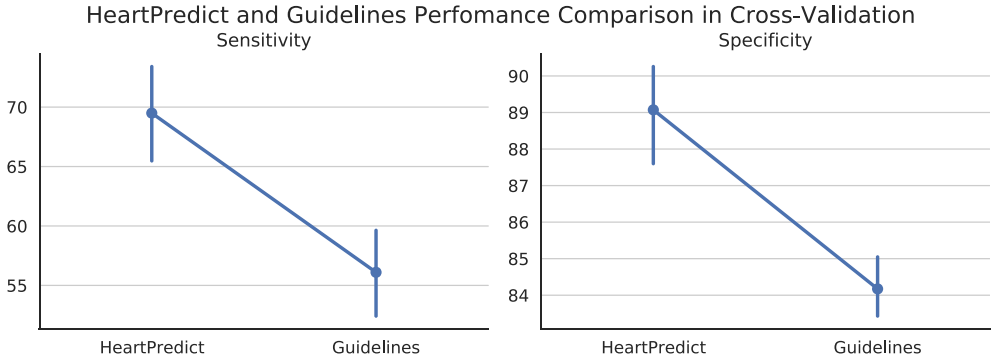

Fig. 9. Cross-validation comparison of HeartPredict and rules from guidelines. Results shows a higher sensitivity of HeartPredict for a fixed specificity (left), and a higher specificity for a fixed sensitivity (right).

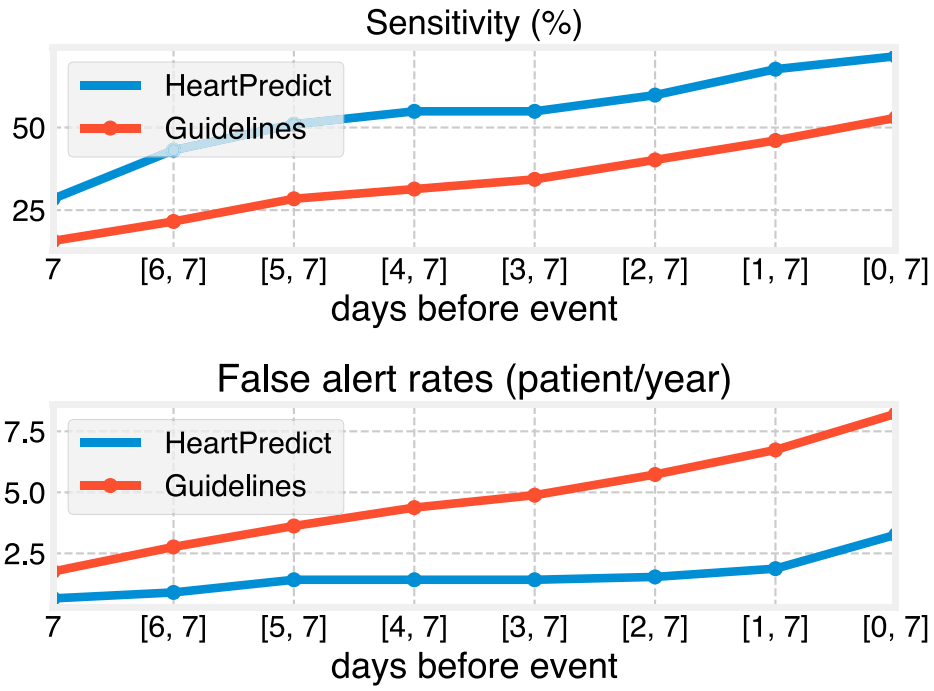

Fig. 10. HeartPredict early detection evaluation on test set - For a similar sensitivity, HeartPredict (in blue) has less false alarms compared to rules (in red) in 0–7 days horizon. Also, for a similar specificity, HeartPredict has a better detection of heart failure events in 0–7 days horizon. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

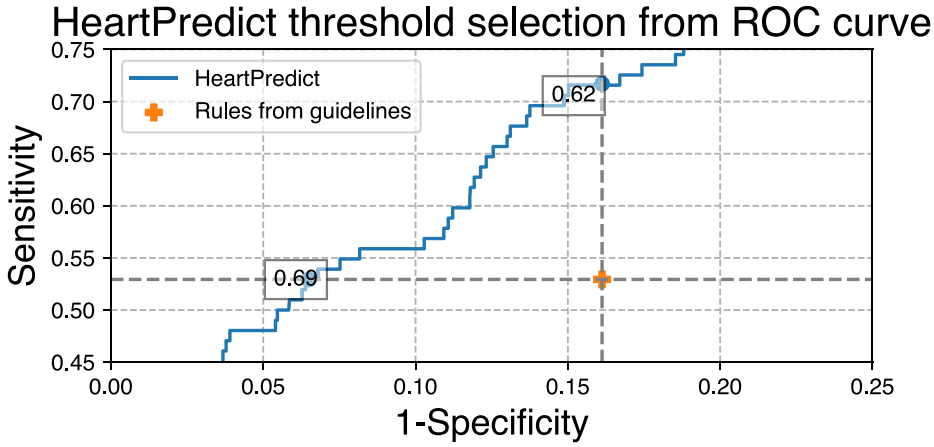

Fig. 11. HeartPredict threshold selection - The blue curve represents the ROC curve while the orange cross represents the sensitivity and specificity of the rules from guidelines. The optimal threshold is taken from the range that model sensitivity and specificity are better than those from guidelines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

HeartPredict model is a binary classifier that provides probabilities. Continuous scores (between 0 and 1) represent the likelihood of an episode of HF occurring in the next seven days. In this section, we analyse the predictions of HeartPredict for each health condition. The aim is to assess the consistency of the predictions regarding the risk for patients. For this, we study how the algorithm behaves in each health situation. Fig. 12 shows the distributions of the predictions of the validation subset for each health condition:

- (P0, P1): The red curve represents the positive target (HF hospitalisation and HF related treatment): the median prediction score = 0.81, with 25th, 75th percentiles [0.69, 0.91]
- (P5): The green curve represents the healthy state when target = 0: the median prediction score = 0.32, with 25th, 75th percentiles [0.20, 0.5]
- (P2, P3, P4): The gray curve represents intermediate states where the risk of health worsening is not confirmed but could be checked: the median prediction score = 0.66, with 25th, 75th percentiles [0.5, 0.79]

The results show that HeartPredict captures discriminating patterns that distinguish states of HF (red) from stable states (green). The intermediate state corresponding to unknown conditions (gray) is characterized by a large distribution of the prediction scores. The median score (0.66) and the distribution shape of the gray area indicate that these (unknown) conditions are more likely to be attributed to risk conditions and thus seek medical follow-up. We observe that the algorithm is conservative. It favors a high sensitivity which is rather recommended to avoid risks for patients in uncertain situations. The desired sensitivity level could be adjusted using a suitable threshold on the probability. A

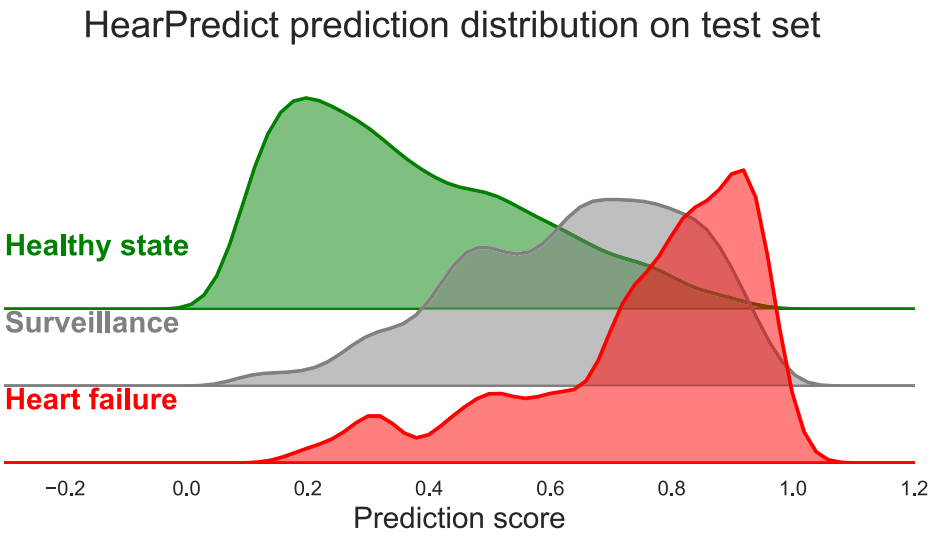

Fig. 12. HeartPredict prediction distribution on the validation set. The distribution of HF predictions is concentrated towards one, while healthy state predictions are concentrated towards zero. The distribution of surveillance events is skewed towards high prediction scores, suggesting the conservative tendency of HeartPredict.

parameter of the algorithm makes it possible to adjust the desired sensitivity to a suitable threshold on the probability.

Moreover, we assess a complementary analysis on the distribution of the prediction scores. Fig. 13 shows a prediction matrix representation of the 6 labels defined in Table 2, according to regular intervals of the prediction scores. Most of the emerging HF hospitalizations (P0) are predominantly predicted in the range of scores [0.8, 1]. This means that there is more confidence in the ability of the algorithm to detect HF. In the same way, the scores of predictions of healthy states (P5) are concentrated in [0, 0.4] which is coherent with the purposes of the classifier. These detailed results also confirm the tendency of the algorithm to consider unknown conditions as probable heart failures.

6.2. *HeartPredict algorithm interpretability*

There is a rising demand for interpretability of AI-based algorithms in real world problems, mainly in healthcare applications. The objective is to have a better understanding of why a machine learning algorithm gives a prediction for more transparent and confident decision support. Moreover, the interpretability is important for: i) users to trust A.I-based algorithms for accountability regarding their actions; ii) validators during their audits in order to give their approval; iii) developers and data scientists to diagnose potential issues for model improvement. Interpretability can then be considered as a proxy for human understanding of A.I models and decisions. Some algorithms by design has transparent models (e.g. Decision Trees) and others (e.g. Deep Learning) needs a downstream post-hoc model to interpret their predictions. HeartPredict is developed with a Random Forests, an ensemble decision trees algorithm considered as “black box” or a “gray box”, which predictions could be interpreted using a model-agnostic module. We use two methods for model-agnostic interpretability:

LIME [32] local surrogate works by taking a point and its black box prediction. A new dataset is generated with local variations around this point of interest. Then, an interpretable algorithm (e. g. decision trees,

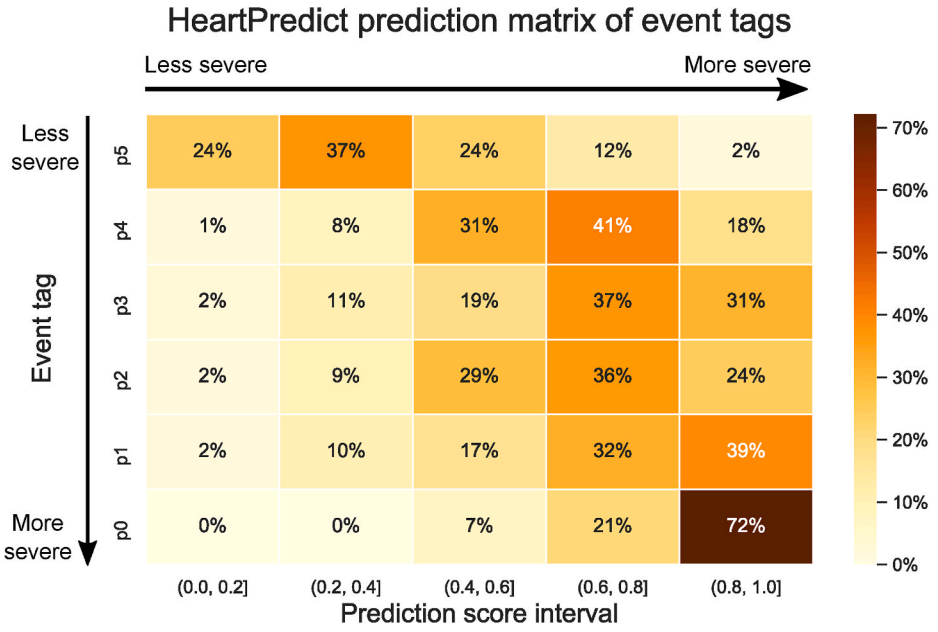

Fig. 13. Matrix representation of the test set predictions. The 6 health state levels are represented according to regular intervals of the prediction scores.

logistic regression, …) is trained using the generated dataset and their corresponding "black box" predictions used as targets. The resulting interpretable model can then be considered as a good local approximation of the "black box" algorithm at this point of interest.

Anchors [33] model explanation works by considering some perturbation distribution at the point of interest. Then, it finds some features that are sufficient to anchor the prediction using a rule i.e. a set of predicates. Afterwards, the algorithm verifies if any counterfactual instance is not found by changing the other features. Finally, the resulting predictions on the features represent a rule interpreting locally the "black box" prediction at this point.

Some examples of simple rules interpreting HeartPredict predictions using Anchors approach are illustrated in Fig. 14.

This interpretability work is carried out for educational purposes to facilitate exchanges and adoption of the algorithm by health professionals.

7. Conclusion and future work

We introduced HeartPredict, a new machine learning algorithm for early detection of heart failure episodes and related complications. HeartPredict relies on a consistent data set (n = 385), resulting from a large clinical study. The algorithm uses home care data in real patient monitoring setting. Measurements are recorded during 18 months follow-up whilst the medical events (labels) are collected from hospitals and other care centers. We rely on a thorough medical expertise to set the HF prediction methodology. This allowed us to define a consistent stratification of medical events, the prediction horizon window, and the time series observation window. Several tests have been carried out to identify the most relevant characteristics sensitive to HF episodes, using patients' weight and symptoms time-series, in addition to socio-demographics.

An extensive assessment of HeartPredict algorithm is performed. Unlike other studies, we use patient out-of-sample validation. Such an assessment is more reliable, since it prevents over-fitting, and ensures a better generalisation on new patients. HeartPredict achieves better sensitivity and specificity compared to rules from guidelines and other machine learning algorithms (Tables 3 and 4).

Besides sensitivity and specificity, we propose a new performance indicator about precocity, i.e. how early can an algorithm detect HF events. This is useful for triggering proactive and effective medical actions, in order to cope with sudden and often fatal events as HF. Within this context, HeartPredict can detect HF events earlier. The algorithm detects more than 50% of HF events 3 days before they occur. Our experiments show also that HeartPredict prediction scores are consistent with patient health condition. Indeed, the distributions of prediction scores allow to establish HF risk stratification levels. Such a feature can help care professionals to optimise resources by prioritizing patients with higher risk of HF episode.

The main limitation regarding HeartPredict, as any other ML algorithm, is the lack of theoretical guarantees of statistical performance. The performance is defined on empirical data sampled from a theoretical distribution. Another limitation is the difficulty to have a quantitative description of the medical events, that could ensure a more accurate analysis of HF episodes. Also, HeartPredict relies on a study with patients that have more than one month of follow-up. This can lead to a participation bias as only more compliant patients are included in the analysis. Such bias should be monitored in order to estimate its impact.

Future works call for a better integration of model explainability in order to facilitate adoption from the practitioners. This includes developing adapted tools for time series explanations such as shapelet [34].

In perspective, a prospective pilot with care professionals at Air Liquide Home Care Services is being designed to assess the algorithm. The pilot relies in an evaluation phase in "silent mode", by monitoring the behaviour of the algorithm without using its outputs for medical decision-making. This will help to safely confirm HeartPredict performance. It will also help to collect operational feedback and more data to enhance the model.

After running the pilot, it is important to identify and to implement key monitoring indicators in order to detect ML algorithm drift or data distribution deviations. Other developments can include the evaluation of the impact of complementary vital signs (heart rate, blood pressure, …), and the implementation of a personalized care plan recommendation system depending on the patient risk score produced by the algorithm.

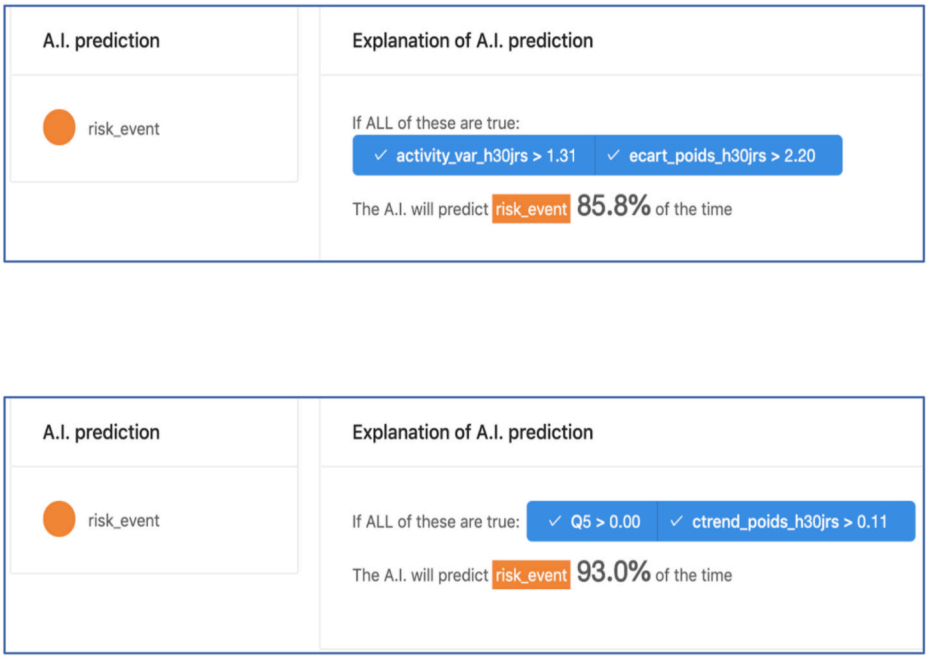

Fig. 14. Interpretability of HeartPredict algorithm predictions at 2 examples

Declaration of competing interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

[1] Remme W. Guidelines for the diagnosis and treatment of chronic heart failure. Eur Heart J 2001;22:1527–60. https://doi.org/10.1053/euhj.2001.2783. 10.1053/euhj.2001.2783.

[2] Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, de Ferranti SD, Floyd J, Fornage M, Gillespie C, Isasi CR, Jiménez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, Mackey RH, Matsushita K, Mozaffarian D, Mussolino ME, Nasir K, Neumar RW, Palaniappan L, Pandey DK, Thiagarajan RR, Reeves MJ, Ritchey M, Rodriguez CJ, Roth GA, Rosamond WD, Sasson C, Towfighi A, Tsao CW, Turner MB, Virani SS, Voeks JH, Willey JZ, Wilkins JT, Wu JH, Alger HM, Wong SS, Muntner P. Heart disease and stroke statistics—2017 update: a report from the american heart association. Circulation 2017;135. https://doi.org/10.1161/cir.0000000000000485. 10.1161/cir.0000000000000485.

[3] Herold R, van den Berg N, Dörr M, Hoffmann W. Telemedical care and monitoring for patients with chronic heart failure has a positive effect on survival. Health Serv Res 2017;53:532–55. https://doi.org/10.1111/1475-6773.12661. 10.1111/1475-6773.12661.

[4] Gensini GF, Alderighi C, Rasoini R, Mazzanti M, Casolo G. Value of telemonitoring and telemedicine in heart failure management. Card Fail Rev 2017;3:1. https://doi.org/10.15420/cfr.2017:6:2. 10.15420/cfr.2017:6:2.

[5] Inglis SC, Clark RA, McAlister FA, Stewart S, Cleland JG. Which components of heart failure programmes are effective? a systematic review and meta-analysis of the outcomes of structured telephone support or telemonitoring as the primary component of chronic heart failure management in 8323 patients: abridged coc. Eur J Heart Fail 2011;13:1028–40. https://doi.org/10.1093/eurjhf/hfr039. URL, https://doi.org/10.1093/eurjhf/hfr039.

[6] Bendelac H, Pathak A, Molinier L, Ruidavets J-B, Mayère A, Berry M, Delmas C, Roncalli J, Galinier M. Optimization of ambulatory monitoring of patients with heart failure using telecardiology (osicat). Eur. Res. Telemed. /La Recherche Européenne en Télémédecine 2014;3:161–7.

[7] Galinier M, Roubille F, Berdague P, Brierre G, Cantie P, Dary P, Ferradou J-M, Fondard O, Labarre JP, Mansourati J, Picard F, Ricci J-E, Salvat M, Tartière L, Ruidavets J-B, Bongard V, Delval C, Lancman G, Pasche H, Ramirez-Gil JF, A P. Telemonitoring versus standard care in heart failure: a randomised multicentre trial. Eur J Heart Fail 2020;22:985–94. https://doi.org/10.1002/ejhf.1906. 10.1002/ejhf.1906.

[8] Zhang J, Goode KM, Cuddihy PE, Cleland JG, on behalf of the TEN-HMS Investigators. Predicting hospitalization due to worsening heart failure using daily weight measurement: analysis of the Trans-European Network-Home-Care Management System (TEN-HMS) study. Eur J Heart Fail 2009;11:420–7. https://doi.org/10.1093/eurjhf/hfp033. http://doi.org/10.1093/eurjhf/hfp033.

[9] Ledwidge MT, O'Hanlon R, Lalor L, Travers B, Edwards N, Kelly D, Voon V, McDonald KM. Can individualized weight monitoring using the HeartPhone algorithm improve sensitivity for clinical deterioration of heart failure? Eur J Heart Fail 2013;15:447–55. https://doi.org/10.1093/eurjhf/hfs186. URL, http://doi.org/10.1093/eurjhf/hfs186.

[10] Gyllensten IC, Bonomi AG, Goode KM, Reiter H, Habetha J, Amft O, Cleland JG. Early indication of decompensated heart failure in patients on home-telemonitoring: a comparison of prediction algorithms based on daily weight and noninvasive transthoracic bio-impedance. JMIR Med Inform 2016;4:e3. https://doi.org/10.2196/medinform.4842.

[11] McMurray John J, Adamopoulos S, Anker Stefan D. Esc. guidelines for the diagnosis and treat¬ ment of acute and chronic heart failure. Eur Heart J 2012;33:1787–847.

[12] Lindenfeld J, Albert NM, Boehmer JP, Collins SP, Ezekowitz JA, Givertz MM, Katz SD, Klapholz M, Moser DK, Rogers JG, et al. Hfsa 2010 comprehensive heart failure practice guideline. J Card Fail 2010;16:e1.

[13] Larburu N, Artetxe A, Escolar V, Lozano A, Kerexeta J. Artificial intelligence to prevent mobile heart failure patients decompensation in real time: monitoring-based predictive model. Mobile Inf Syst 2018:1–11. 2018, https://www.hindawi.com/journals/misy/2018/1546210/. 10.1155/2018/1546210.

[14] Adamson PB, Zile MR, Cho YK, Bennett TD, Bourge RC, Aaron MF, Aranda JM, Abraham WT, Kueffer FJ, Taepke RT. Hemodynamic factors associated with acute decompensated heart failure: Part 2—use in automated detection. J Card Fail 2011;17:366–73. https://linkinghub.elsevier.com/retrieve/pii/S1071916411000261. 10.1016/j.cardfail.2011.01.011.

[15] Abraham WT, Compton S, Haas G, Foreman B, Canby RC, Fishel R, McRae S, Toledo GB, Sarkar S, Hettrick DA, On behalf of the FAST Study Investigators. Intrathoracic impedance vs daily weight monitoring for predicting worsening heart failure events: results of the fluid accumulation status trial (FAST): predicting worsening heart failure events. Congest Heart Fail 2011;17:51–5. https://doi.org/10.1111/j.1751-7133.2011.00220.x. URL, http://doi.org/10.1111/j.1751-7133.2011.00220.x.

[16] Gilliam FR, Ewald GA, SWEENEY RJ. Feasibility of automated heart failure decompensation detection using remote patient monitoring: results from the decompensation detection study. Innovat Cardiac Rhythm Manage 2012;3:1–10.

[17] Dickstein K, , Authors/Task Force Members, Cohen-Solal A, Filippatos G, McMurray JJ, Ponikowski P, Poole-Wilson PA, Strömberg A, van Veldhuisen DJ, Atar D, Hoes AW, Keren A, Mebazaa A, Nieminen M, Priori SG, Swedberg K, Vahanian A, , ESC Committee for Practice Guidelines (CPG), Camm J, De Caterina R, Dean V, Dickstein K, Filippatos G, Funck-Brentano C, Hellemans I, Kristensen SD, McGregor K, Sechtem U, Silber S, Tendera M, Widimsky P, Zamorano JL, Tendera M, Document Reviewers, Auricchio A, Bax J, Böhm M, Corrà U, della Bella P, Elliott PM, Follath F, Gheorghiade M, Hasin Y, Hernborg A, Jaarsma T, Komajda M, Kornowski R, Piepoli M, Prendergast B, Tavazzi L, Vachiery J-L, Verheugt FWA, Zamorano JL, Zannad F. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the task force for the diagnosis and treatment of acute and chronic heart failure 2008 of the European Society of Cardiology. Developed in collaboration. Eur J Heart Fail 2008;10:933–89. https://doi.org/10.1016/j.ejheart.2008.08.005. URL, http://doi.org/10.1016/j.ejheart.2008.08.005.

[18] Chaudhry SI, Mattera JA, Curtis JP, Spertus JA, Herrin J, Lin Z, Phillips CO, Hodshon BV, Cooper LS, Krumholz HM. Telemonitoring in patients with heart failure. N Engl J Med 2010;363:2301–9. http://www.nejm.org/doi/abs/10.1056/NEJMoa1010029. 10.1056/NEJMoa1010029.

[19] Brons M, Koudstaal S, Asselbergs FW. Algorithms used in telemonitoring programmes for patients with chronic heart failure: a systematic review. Eur J Cardiovasc Nurs 2018;17:580–8. http://journals.sagepub.com/doi/10.1177/1474515118786838. 10.1177/1474515118786838.

[20] Hjorth B. EEG analysis based on time domain properties. Electroencephalogr Clin Neurophysiol 1970;29:306–10. https://doi.org/10.1016/0013-4694(70)90143-4. 10.1016/0013-4694(70)90143-4.

[21] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. jair 2002;16:321–57. https://www.jair.org/index.php/jair/article/view/10302. 10.1613/jair.953.

[22] He Haibo, Yang Bai, Garcia EA, Li Shutao. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International joint conference on neural networks (IEEE world congress on computational intelligence). Hong Kong, China: IEEE; 2008. p. 1322–8. http://ieeexplore.ieee.org/document/4633969/. 10.1109/IJCNN.2008.4633969.

[23] Tomek I. Two modifications of CNN. In: IEEE Trans. Syst., Man, Cybern., SMC-6; 1976. p. 769–72. http://ieeexplore.ieee.org/document/4309452/. 10.1109/TSMC.1976.4309452.

[24] Liu Xu-Ying, Wu Jianxin, Zhou Zhi-Hua. Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst., Man, Cybern. B 2009;39:539–50. http://ieeexplore.ieee.org/document/4717268/. 10.1109/TSMCB.2008.2007853.

[25] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6. http://www.nature.com/articles/323533a0. 10.1038/323533a0.

[26] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80. : http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735. 10.1162/neco.1997.9.8.1735.

[27] Graves A, Jaitly N, Mohamed A-r. Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE workshop on automatic speech recognition and understanding. IEEE; 2013. p. 273–8. Olomouc, Czech Republic, http://ieeexplore.ieee.org/document/6707742/. 10.1109/ASRU.2013.6707742.

[28] Sundermeyer M, Ney H, Schluter R. From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. Audio Speech Lang. Process 2015;23:517–29. http://ieeexplore.ieee.org/document/7050391/. 10.1109/TASLP.2015.2400218.

[29] Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. 2017. arXiv:1511.03677 [cs], http://arxiv.org/abs/1511.03677. ArXiv: 1511.03677.

[30] Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Comput Struct Biotechnol J 2017;15:26–47. https://doi.org/10.1016/j.csbj.2016.11.001. 10.1016/j.csbj.2016.11.001.

[31] Logé F, Le Pennec E, Amadou-Boubacar H. Adaptive predictive questionnaire by approximate dynamic-programming. In: UCAI 20 : workshop on user-centered artificial intelligence; 2020.

[32] Ribeiro MT, Singh S, Guestrin C. " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135–44.

[33] Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: AAAI conference on artificial intelligence (AAAI); 2018.

[34] Hills J, Lines J, Baranauskas E, Mapp J, Bagnall A. Classification of time series by shapelet transformation. Data Min Knowl Discov 2014;28:851–81.