35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Random Forest For Automatic Assessment Of Heart Failure Severity In A Telemonitoring Scenario

G. Guidi, M. C. Pettenati, R. Miniati and E. Iadanza, *Member IEEE*

***Abstract*— In this study, we describe an automatic classifier of patients with Heart Failure designed for a telemonitoring scenario, improving the results obtained in our previous works. Our previous studies showed that the technique that better processes the heart failure typical telemonitoring-parameters is the Classification Tree. We therefore decided to analyze the data with its direct evolution that is the Random Forest algorithm. The results show an improvement both in accuracy and in limiting critical errors.**

I. Introduction

In our work we are developing an Heart Failure (HF) Computer Decision Support System (CDSS) that, combined with a handy device for the automatic acquisition of a set of clinical parameters, enables the support for telemonitoring functions. The system provides an HF severity/type assessment function using four machine learning techniques: a Neural Network, a Support Vector Machine, a Classification Tree and a Fuzzy Expert System whose rules are produced by a Genetic Algorithm. This system is accurately described in [1][2][3] where it was found that the technique that better processes the HF typical telemonitoring-parameters is Classification And Regression Tree (CART) which also provides human-readable results. CART is also used in other studies that deal with the diagnosis of HF such as [8]. In this paper we investigate the performance improvement obtainable analyzing data with CART's direct evolution that is the Random Forest (RF) algorithm [4]. RF are often used, in the field of HF, to predict death/readmission, to identify risk factor or in general to analyze HF parameters such HRV [5][6][7]. A telemonitoring scenario suitable for proposed CDSS is shown in Figure 1. and requires that the patient is provided of a kit for the automatic acquisition of some parameters on a daily basis. A nurse will periodically visit the patient at home to perform other measurements on a monthly basis (Brain Natriuretic Peptide - BNP).

The HF severity output provided by the CDSS may be useful both in giving prompt advices to non-experts personnel, as well as in filtering patients' data and highlighting to the cardiologists only the worst patients for further analysis.

G. Guidi is with the Department of Information Engineering – University of Florence, Florence, Italy (e-mail: gabriele.guidi@unifi.it).

M.C. Pettenati is with ICON (International Center of Computational Neurophotonics) Foundation, 50019 Florence, Italy

R. Miniati is with the Department of Information Engineering – University of Florence, Via Santa Marta 3 50039 - Florence, Italy

E. Iadanza is with the Department of Information Engineering – University of Florence, Via Santa Marta 3 50039 - Florence, Italy (corresponding author: 0039-3475922874; e-mail: ernesto.iadanza@unifi.it).

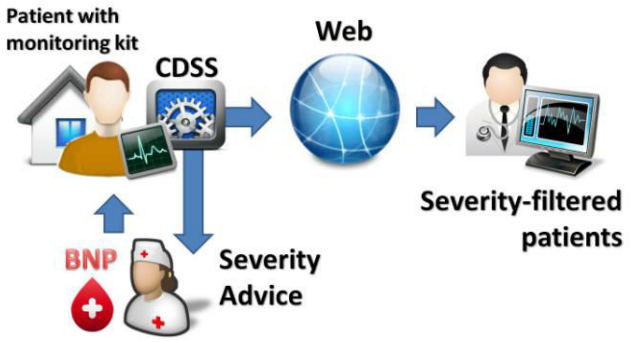

Figure 1. Telemonitoring Scenario of the CDSS

A comparison of the performance of the algorithm CART and Random Forest is shown in Results section and discussed in section IV.

II. Materials and Methods

A. Database

We worked on a de-identified HF patients database, with varying severity degrees, all treated by the Cardiology Department of the Hospital Santa Maria Nuova in Florence, Italy in the period 2001-2008. The database consists of a total of 136 records of 90 patients, including baseline and follow-up data. The demographics characteristic of the patient at the baseline visit are shown in TABLE I.
Variables in database that are used as input of the Machine Learning Techniques are the following:

- Anamnestic data: *Age, Gender, Codified symptomatology* (1: no symptoms and no limitation in ordinary physical activity, 2: mild symptoms, 3: marked limitation in activity due to symptoms, 4: severe limitations)
- Instrumental data*: Weight, Systolic Blood Pressure, Diastolic Blood Pressure, Ejection Fraction (EF), Brain Natriuretic Peptide (BNP), Heart Rate, ECG-Parameter* (atrial fibrillation true/false, left bundle branch block true/false, ventricular tachycardia true/false.

At the time of the data collection, the specialist physician provided an HF severity assessment in three levels: 1-*Mild*, 2-*Moderate* and 3-*Severe*, which was stored in the database. Moreover, after 12-24 months from the data collection, the status of each patient in terms of HF type (1-*stable,* 2-*rare exacerbation,* 3-*frequent exacerbation*) was assessed and associated to the corresponding record. These physician evaluations are used as target output in the supervised training process. The system has 12 inputs and 2 three-levels

outputs. Output classes distribution in database are shown in TABLE II. To date, the database is populated with data from measure on outpatients, performed in the Cardiology Department of S. Maria Nuova Hospital in Florence. When the system will be fully operational in the telemonitoring scenario, the database will be automatically fed with data acquired at patient's home using a wearable kit, or a custom homecare system. Some parameters are not, however, easily measurable in an automatic way, as the EF and BNP. These parameters, however, are very significant in assessing the HF severity, as shown in Figure 3. Since EF parameter changes very slowly with Increasing severity of the disease, just one measurement every 6 months by physician is needed. With regard to the BNP instead we intend to equip a nurse with a BNP point of care device for capillary draw and the measurement will be performed monthly.

TABLE I. ANAGRAPHICS OF PATIENTS IN DATABASE

	Age					**Gender**	
	<50	***50-60***	***60-70***	***70-80***	***>80***	***M***	***F***
N° of Patients	8	13	23	44	48	91	45
Total	136					136	

TABLE II. CLASSES DISTRIBUITION

	HF Severity Output		
	Mild HF	***Moderate HF***	***Severe HF***
N° of Patients	51	37	48
	HF Typologies Output		
	Stable	***Freq. Ex.***	***Rare Ex.***
N° of Patients	110	14	12

B. Method

We have trained a Random Forest algorithm using the database described above. RFs are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. According to Breiman [4]: "The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. […] Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. […]

Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree of the forest is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning."

To implement algorithm we used Matlab 7.11.0 and the Random Forest Tool available at http://code.google.com/p/randomforest-matlab.

First, we determine what was the best number of features (*m*) to be used for each tree. We performed various tests obtaining the best performances with *m*=4. As we have 12 inputs, this figure is in line with the suggestion in the literature that states that a well-balanced value for *m* is as shown in (1).

$$m=\sqrt{Number\ of\ Features} \tag{1}$$

Then we assessed the optimal number of trees in the forest. To do this we analyzed the Out Of Bag (OOB) error rate related to the increasing of number of trees. We chose the value of 2000 trees, because, as seen in Figure 2. , after that value the error rate is sufficiently stabilized. Another important parameter on which to operate is the cutoff of each class, which allows you to make sure that each class, in order to win, should get more votes from the various trees. By default, the winner class is simply the one that gets the most votes (equal cutoff for each class) but, in case of unbalanced database, it is very useful to work on the cutoff to balance the chances that each class has to win.

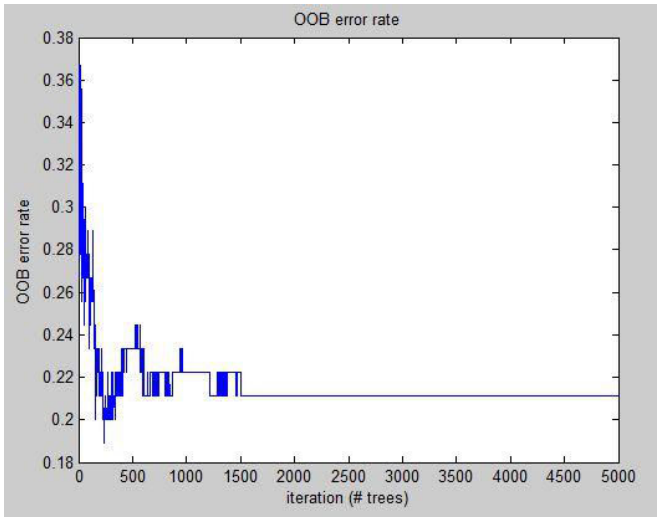

Figure 2. OBB error rate to determine optimal number of trees.

As shown in TABLE II. , while data to perform severity assessment are balanced, as regards the HF Type the database is strongly asymmetric: the most part of patients follow in the "stable" class rather than in the other two. So we set a cut off vector as shown in TABLE III. Reducing the cutoff value makes a class an easy-winner.

TABLE III. RANDOM FOREST PARAMETER - CUTOFF

	HF Severity Output		
	Mild HF	***Moderate HF***	***Severe HF***
Cutoff	30	30	40
	HF Typologies Output		
	Stable	***Freq. Ex.***	***Rare Ex.***
Cutoff	50	20	30

Tuning the cutoff value we were able to reduce to a minimum the number of critical errors. The RF, as a result of its operation, also provides the importance of each variable. This is very useful because it makes the results more human-readable, by highlighting which inputs are indispensable.

C. Performance Measurement

In order to measure and compare the performance of the RF and CART methods, we adopted both 10 fold cross validation and holdout methods. Notice that for the RF a cross validation is not strictly necessary due to its internal bootstrap. In order to better exploit our data, we made the assumption of considering each record of the database, i.e. "follow-up information," as if it were a patient. In this way we have considered to have a database composed of 136 different patients each with a single follow-up. This assumption is justified by the fact that the follow up are for a large period of time (1-2-3 years) and the parametric situation and health of the patient was changed so as to justify the approximation described. In this article, from now on, we will refer to "patient" meaning a "database record". During the cross validation process we have taken precautions so that follow-ups of the same patient are grouped within the same fold thus our assumption does not affect the independence of the folds.

In both methods, a person-independent scheme was used, that is, all the records of the same patients were in the same fold or in the same subset (training / testing subset).

So we semi-randomly divided the 136 patient in 10 folds each containing 13-14 patient for cross validation method and in a training set of 92 patients and a test set of 44 patients (test-set=1/3 train-set=2/3) for holdout method. In the Results section is reported the test set accuracy (generalization capability) for each machine learning technology. We used multiclass accuracy formula (2) according with [9] (TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative). For each method the number of "critical errors" committed, meaning the classification of a *severe* HF patient as *mild* and vice versa has been assessed. Our primary goal is to minimize these critical errors, in order to minimize false alarms or missed activations of appropriate care interventions.

$$Accuracy = \sum_{i=1}^{N^\circ class} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \qquad (2)$$

III. RESULTS

Cross validation and holdout results are shown in TABLE IV. , TABLE V. , TABLE VI. And TABLE VII. Variable Importance in Holdout test is shown in Figure 3. and Figure 4.

TABLE IV. PERFORMANCE IN HF SEVERITY ASSESSMENT

	Cross Valid. Method Performance	
	Average Accuracy	***N° of Critical Errors***
RF	83.30%	1/99
CART	81.79%	2/99

TABLE V. PERFORMANCE IN HF SEVERITY ASSESSMENT

	Holdout Method Performance	
	Average Accuracy	***N° of Critical Errors***
RF	89.39%	0/35
CART	87.88	0/35

TABLE VI. PERFORMANCE IN HF TYPE ASSESSMENT

	Cross Valid. Method Performance	
	Average Accuracy	***N° of Critical Errors***
RF	85.68%	5/122
CART	87.58%	9/122

TABLE VII. PERFORMANCE IN HF TYPE ASSESSMENT

	HoldOut Method Performance	
	Average Accuracy	***N° of Critical Errors***
RF	86.36	4/39
CART	87.88	4/39

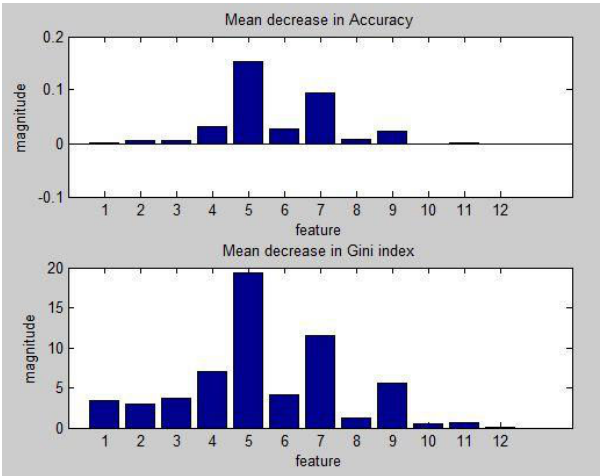

Figure 3. Mean decrease in Accuracy and in Gini Index in Hold Out severity assesment Test. Legend: 1-*Syst. Blood Pressure*, 2-*Diast. Blood Pressure*, 3-*Heart Rate*, 4-*Weight*, 5-*BNP*, 6-*Sympthoms*, 7-*EF*, 8-*Gender*, 9-*Age*, 10-*Atrial Fibr.*, 11-*Bundle Bloc*, 12-*Venticular Tachicardia*.

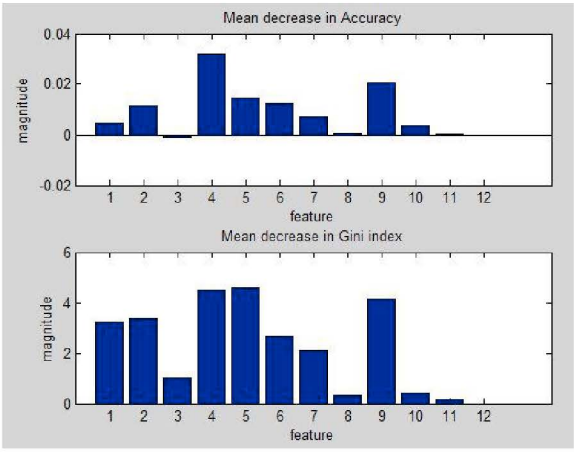

Figure 4. Mean decrease in Accuracy and in Gini Index in holdout HF type assessment test. Legend: 1-*Syst. Blood Pressure*, 2-*Diast. Blood Pressure*, 3-*Heart Rate*, 4-*Weight*, 5-*BNP*, 6-*Sympthoms*, 7-*EF*, 8-*Gender*, 9-*Age*, 10-*Atrial Fibr.*, 11-*Bundle Bloc*, 12-*Venticular Tachicardia*.

These results are in agreement with what was found in our previous studies [1][2][3]. Indeed, as shown in Figure 5. split variables of CART (after pruning) in severity assessment are precisely BNP, EF and Weight.

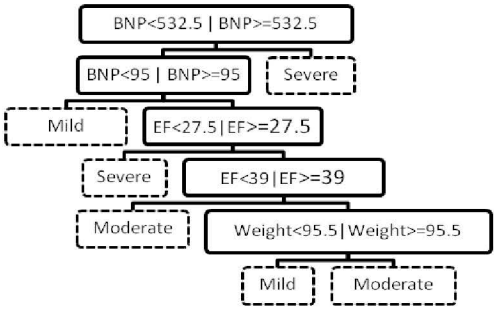

Figure 5. Severity assesment in pruned CART

IV. Discussion And Conclusion

In this paper we described an improvement of a CDSS for the analysis of heart failure in a telemonitoring scenario. In our previous work we used several machine learning methods, and it emerged that the best technique was the CART. Here we used the Random Forest algorithm, that combines many classification trees using bootstrap techniques, to evaluate whether this could increase or not the already good performance of a single CART. In the results section it can be seen that the performance of RF trained using our current database are almost the same of the CART. Working on some parameters of the random forest we were able to obtain slightly better results in severity assessment (moving from a 81.79% CART accuracy to a 83.30% RF accuracy) and in reducing critical errors (improving of about 1% as shown in TABLE IV.). The best reduction in critical errors was obtained in HF type assessment (CART 9, RF 5). This may be due to the higher settings that can be performed on RF and its propensity to deal with unbalanced data [10]. Our results in severity assessment are good if compared with other studies that assess HF severity. In [11] Yang et al. combined two Support Vector Machines (SVM) to classify HF patients in three groups. (74.4% global accuracy, 78.8% - 87.5% - 65.6% accuracy to classify healthy - HF prone - HF patients respectively).

In [8] Pecchia et al. used decision tree techniques to classify patients in three groups of severity (healthy, moderate, severe) using Heart Rate Variability measurements. (HF vs normal subject: 96% accuracy - severe vs moderate: 79.3% accuracy).

Our results are not directly comparable with some HF binary classifiers that distinguish healthy from HF patients (for example [12] and [13]), since these studies have just two output classes (healthy vs diseased) that are obviously more easily separable than our three output classes ('mild disease', 'moderate disease' and 'severe disease').

References

[1] G. Guidi, M. C. Pettenati, R. Miniati, and E. Iadanza, "Heart Failure analysis Dashboard for patient's remote monitoring combining multiple artificial intelligence technologies," in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012, pp. 2210–2213.

[2] G. Guidi, E. Iadanza, and M. C. Pettenati, "Heart Failure Artificial Computer Aided Diagnosis Telecare System Using Various Artificial Intelligence Techniques". In: CONGRESSO NAZIONALE DI BIOINGEGNERIA 2012. ATTI. Roma, 2012, BOLOGNA: Pàtron, ISSN: 978-88-555. .

[3] G. Guidi, E. Iadanza, M. C. Pettenati, M. Milli, F. Pavone, and G. Biffi Gentili, "Heart Failure Artificial Intelligence-based Computer Aided Diagnosis Telecare System," ICOST 2012, Lecture Notes in Computer Science, vol. 7251, pp. 278–281, 2012.

[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] E. Hsich, E. Z. Gorodeski, E. H. Blackstone, H. Ishwaran, and M. S. Lauer, "Identifying important risk factors for survival in patient with systolic heart failure using random survival forests.," Circulation. Cardiovascular quality and outcomes, vol. 4, no. 1, pp. 39–45, Jan. 2011.

[6] A. Jovic and N. Bogunovic, "Random Forest-Based Classification of Heart Rate Variability Signals by Using Combinations of Linear and Nonlinear Features," in XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010, vol. 29, P. Bamidis and N. Pallikarakis, Eds. Springer Berlin Heidelberg, 2010, pp. 29–32.

[7] A. G. Au, F. a McAlister, J. a Bakal, J. Ezekowitz, P. Kaul, and C. van Walraven, "Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization.," American heart journal, vol. 164, no. 3, pp. 365–72, Sep. 2012.

[8] L. Pecchia, P. Melillo, and M. Bracale, "Remote health monitoring of heart failure with data mining via CART method on HRV features.," IEEE transactions on bio-medical engineering, vol. 58, no. 3, pp. 800–4, Mar. 2011.

[9] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing and Management, vol. 45, no. 4, pp. 427–437, 2009.

[10] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," no. 1999, pp. 1–12.

[11] G. Yang, Y. Ren, Q. Pan, and G. Ning, "A heart failure diagnosis model based on support vector machine," IEEE International Conference on Biomedical Engineering and Informatics, no. Bmei, pp. 1105–1108, 2010.

[12] F. S. Gharehchopoghi and Z. A. Khalifelu, "Neural Network Application in Diagnosis of Patient: A Case Study," Computer Networks and Information Technology (ICCNIT), 2011 International Conference on, pp. 245–249, 2011.

[13] C. Wang, "SVD and SVM based approach for Congestive Heart Failure Detection from ECG Signal," Computers and Industrial Engineering (CIE), 2010 40th International Conference on, pp. 1–5, 2010.