Exploring Machine Learning Algorithms to Identify Heart Failure Patients: the Tuscany Region case study

Silvia Panicacci, Massimiliano Donati, Luca Fanucci
Dept. of Information Engineering
University of Pisa
Pisa, Italy
silvia.panicacci@phd.unipi.it
massimiliano.donati@unipi.it
luca.fanucci@unipi.it

Irene Bellini
Hygiene and Preventive Medicine
ASL Toscana Centro
Prato, Italy
irene.bellini@uslcentro.toscana.it

Francesco Profili, Paolo Francesconi
Osservatorio di Epidemiologia
Agenzia Regionale Sanità, ARS
Florence, Italy
francesco.profili@ars.toscana.it
paolo.francesconi@ars.toscana.it

***Abstract*—Heart failure patients have become an important challenge for the healthcare system, since they represent a medical, social and economic problem. Early heart failure diagnoses can be very useful to improve patients' quality of life and to reduce the resources consumption, but they can be complex for the general practitioners. Data mining and machine learning techniques can really help in this field. The aim of this study is to validate some machine learning models to identify heart failure patients, starting from administrative data, and to make them transparent and interpretable. Despite the lack of clinical data, not available in Italy, but the most employed for the identification of heart failure patients, the results are comparable with the state-of-the-art ones and the models outperform the performances already obtained in Tuscany.**

***Keywords*-heart failure; machine learning; black box;**

I. Introduction

Heart failure (HF) is defined as the incapability of the heart to pump blood around the body properly. Its symptoms can be mutiple, e.g. beathlessness, ankle swelling, fatigue, jugular venous pressure and pulmonary crackles [1].

Heart failure patients represent a real problem for the healthcare system, from a medical, social and economic point of view [2]. According to the European Society of Cardiology (ESC), in fact, 26 millions of adults are affected by HF and about 3.6 million people are diagnosed every year in Europe. 17-45% of them die during the first year, while the others die in the following 5 years [3]. HF is one of the main causes of death all over the world [4]. Moreover, due to the disabling symptoms, it also has the greatest negative impact on quality of life (QoL) with respect to the other major chronic diseases, such as diabetes, arthritis and hypertension [2]. At last, the economic burden on these patients is approximately 1-2% of all the healthcare costs, especially because of the repeated hospitalizations [5].

Unfortunately, there are not some shared guidelines for the general practitioners (GPs) on the identification of heart failure patients, making the diagnoses very difficult. However, the GPs are supported by blood tests, chest radiography, electrocardiography and echocardiography and can use several criteria, e.g. Framingham, Boston and Gothenburg ones, based on the results of the tests [7]. On the other hand, they usually react to patients' symptoms, leading to ineffective and inefficient treatments.

To improve the patients' quality of life and to reduce the resources consumption due to multiple hospital admissions, early HF diagnoses can be significantly useful. Data mining and machine learning (ML) algorithms can really help in the early identification of HF patients, thanks to the great quantity of data. Multiple studies have been done, using different techniques according to the available data. Natural Language Processing (NLP) was used to extract the diagnosis of HF with the Framingham criterion from clinical notes and electronic medical records [8]-[10]. When words were not available, clinical data were employed as inputs for several ML algorithms (e.g. SVMs, random forests, neural networks, decision trees, k-NN) [5]: blood test, heart rate variability (HRV), echocardiography, electrocardiography, chest radiography and physical tests [11], blood pressure, smoking situation, age and sex [12], dyspnea and Pro-Brain Natriuretic Peptides (Pro-BNP) [13], long-term ECG time series [14], demographic, health behaviour, use of care, clinical diagnosis, clinical measures, laboratory data and prescription orders for anti-hypertensive information [15] and heart sound and cardiac reserve characteristics [16]. Only few studies used only administrative data for the identification of HF patients, without the implication of ML, but with statistical analyses or ad-hoc algorithms (e.g. the algorithm already used in Tuscany region for this aim [17], and the method proposed by Shultz et al [18]). They reached very lower performances with respect to the previous ones.

The aim of this study is to validate some machine learning algorithms for the identification of HF patients, using administrative data and reaching performances comparable with the models including also clinical data.

To let the GPs trust these new models, an analysis of the involved variables has been done, underlying their different weight for the predictions.

II. Methods

In Italy, clinical data are not available. On the contrary, all the residents produce some digital tracks every time they receive any public or private health services. Data coming from the hospitals (i.e. diagnoses and procedures), from the outpatients (i.e. assistive, diagnostic and rehabilitation performances), from the pharmacies (e.g. prescribed drugs, etc.) and data regarding the exemptions (both for income or diagnosis) compose the administrative flows. They are usually almost complete, because the Regional Health System requires a complete reporting used for governance purpose and to rank providers' performance.

In this study, the data available in the mARSupio database of the Agenzia Regionale Sanità (ARS) in Florence, Tuscany, Italy [19], where the Tuscan administrative data are collected, were used as inputs for the ML algorithms, to solve the binary classification problem of identification of heart failure patients. In mARSupio, patients' privacy is protected according to the Italian law [20], in fact every patient is identified by an univocal identification code (IDUNI).

In order to implement supervised algorithms, 11 Tuscan doctors provided a list of their patients surely affected by HF, as part of a specific heart failure pathway project. The assumption was that all their other patients were not affected by HF. The population involved in the study was then composed by 14 616 Tuscan residents, 347 of whom (2.37%) were affected by HF and 14 269 (97.63%) were not. So, HF prevalence was coherent with other studies, saying that it is of about 2-3% [3].

For each patient involved in the study, the medical history from 1^{st} January 2010 (data before 2010 are not so confident) to 1^{st} January 2018 was rebuilt. In particular, the considered features were similar to the ones used in Tuscany for the identification of complex patients [21], with some few differences for cardiac problems:

- the number of admissions and the number of days of hospitalizations for each Aggregated Clinical Code (ACC) [22] in the considered period represented two features of the final dataset, except for the ACCs regarding cardiomyopathies, secondary hypertension and congestive heart failure, for which the number of admissions and the number of days of hospitalizations were considered for each ICD9 code separately. Diagnoses provided then 740 variables (280 ACCs + 90 ICD9 codes);
- the number of the procedures done in the considered period for each ACC were counted, except for the ACCs regarding heart valve procedures, coronary heart bypass grafts, percutaneous transluminal coronary angioplasties, coronary thrombolyses, diagnostic cardiac catheterizations, insertions and replacements of cardiac pacemaker and other heart surgeries, where the number of the procedures were counted for each ICD9 code. Procedures provided then 380 attributes (224 ACCs + 156 ICD9 codes);
- the number of the performances done in the considered period for each group (made ad-hoc) represented one feature, except for the groups considering instrumental cardiology performances, cardiac medical examinations and cardiac drugs laboratory exams, for which all the codes were considered separately. The number of features provided by this category was 90 (73 groups + 17 codes);
- the number of drugs taken in the considered period for each ATC3 code (third level of Anatomical Therapeutic Chemical classification system [23], which indicates the therapeutic/pharmacological subgroup of the drug) was considered as a variable. Drugs provided 271 features (271 ATC3 codes);
- the combination of the date of release and of expiration for each group of exemptions was considered as a categorical variable with three values (never released, expired in the period 2010-2017, not expired yet), except for exemptions concerning cardiovascular diseases, where each ICD9 code was considered separately. The exemptions provided 31 attributes (27 groups + 4 ICD9 codes);
- age at 1^{st} January 2018 and gender were considered, for a total of 2 features.

The total number of features used as input for the ML algorithms was then 1 514, as shown in Table I.

The target variable was dichotomous, distinguishing between patients with HF and patients without HF, according to the lists made by the GPs. The problem was thus a supervised binary classification one, considering HF patients as the positive class.

The final dataset was then composed by 14 616 rows (a row for each patient) and 1 515 columns (1 514 input features

Table I
Summary of the features of the dataset.

Class	Number
Diagnoses	740
Procedures	380
Performances	90
Drugs	271
Exemptions	31
Personal	2
Total	1 514

+ 1 output variable).

The whole dataset was divided in training and test set, the 80% and the 20% of the original one, respectively.

Since the prevalence of the two sets was the same of the original one (2.37%), the training set was preliminarly balanced using the Synthetic Minority Over-Sampling Technique (SMOTE) [24], which consists in over-sampling the minority class and under-sampling (or over-sampling) the majority one to achieve the desired prevalence and number of samples in both the classes. In this way, the dataset does not become too small. The new training set was then composed by 139 278 samples, 69 639 per class, with a prevalence of 50%. On the contrary, the test set, was not modified, because the performances have to be evaluated on a real sample of the Tuscan population.

On the new balanced training set, a features selection process was executed, in order to reduce the dimension of the dataset and to delete features without relevance for the output. The Boruta algorithm [25] was chosen for this aim. This is a wrapper method, which adds randomness to random forests, performing a top-down search for the most predictive attributes, iteratively deleting the variables whose importance is significantly less than their importance when randomly shuffled. The Boruta algorithm confirmed a group of 572 features, selected as follows:

- 311 of 740 (about 42%) variables regarding diagnoses;
- 97 of 380 (almost 25.5%) attributes for procedures;
- 54 of 90 (60%) features of the class of performances;
- 98 of 271 (about 36.2%) variables for drugs;
- 10 of 31 (almost 32.3%) attributes for exemptions;
- age and gender.

Random Forest (RF) [26] and LASSO [27] were selected as ML algorithms: the first one has usually good performances in classification problems, but it is hard to understand; the second one has usually slightly worse results, but it is more interpretable [28]. Both were trained in 10-fold cross-validation with grid search (to perform also the tuning of the parameters) using different training sets. All the training sets had the 572 features selected before by the Boruta algorithm, but they differed in prevalence. The 10 training sets, obtained simply under-sampling the negative class or with SMOTE, are described in Table II.

All the analyses were done on a Linux server with 64 GB of RAM, using a program written in R language.

Table II
SUMMARY OF THE TRAINING SETS USED TO TRAIN THE MODELS, WITH THE TECHNIQUES USED TO DERIVE THEM, THE ACHIEVED PREVALENCE AND THE FINAL NUMBER OF ROWS.

	Technique	**Prevalence**	**# Rows**
Original	Nothing	2.37%	11 694
Bal10	Under-sampling the negative class	10%	2 780
Bal20	Under-sampling the negative class	19.6%	1 419
Bal25	Under-sampling the negative class	25%	1 112
Smote28	SMOTE	28%	6 950
Bal33	Under-sampling the negative class	33.33%	834
Smote38	SMOTE	38.46%	3 614
Smote42	SMOTE	42.86%	1 946
Bal50	Under-sampling the negative class	50%	556
Smote50	SMOTE	50%	1 112

III. RESULTS

The results were evaluated on the test set, with the original prevalence. Because of the great unbalance of the classes, accuracy, which is the index of correctness in a classification system, cannot be used, to avoid the accuracy paradox [29]. For this reason, F1-Score [30], which gives the same importance to Sensitivity (SE) and Positive Predictive Value (PPV), and F2-Score [31], which weights more SE than PPV, were used to compare and evaluate the models. They are defined as follows:

$$F1Score = 2 \cdot \frac{PPV \cdot SE}{PPV + SE} \tag{1}$$

$$F2Score = 5 \cdot \frac{PPV \cdot SE}{4 \cdot PPV + SE} \tag{2}$$

The results achieved by all the models trained with the different training sets in terms of these two metrics are shown in Table III. It is possible to observe that RF behaves better than LASSO, no matter the set used to train the models. In addition, the performances are different not

Table III
RESULTS ACHIEVED BY ALL THE MODELS IN TERMS OF F1-SCORE AND F2-SCORE.

Algorithm	**Training Set**	**F1-Score (%)**	**F2-Score (%)**
RF	Original	64.23	53.49
	Bal10	66.01	77.19
	Bal20	51.09	70.63
	Bal25	45.72	66.49
	Smote28	74.22	82.76
	Bal33	39.02	60.66
	Smote38	57.22	73.3
	Smote42	44.33	65.22
	Bal50	28	49.17
	Smote50	35.88	57.9
LASSO	Original	34.73	27.32
	Bal10	46.74	48.4
	Bal20	46.77	57.43
	Bal25	44.39	60.7
	Smote28	53.36	64.76
	Bal33	39.44	56.97
	Smote38	49.47	66.86
	Smote42	45.65	64.33
	Bal50	29.02	49.43
	Smote50	33.89	54.31

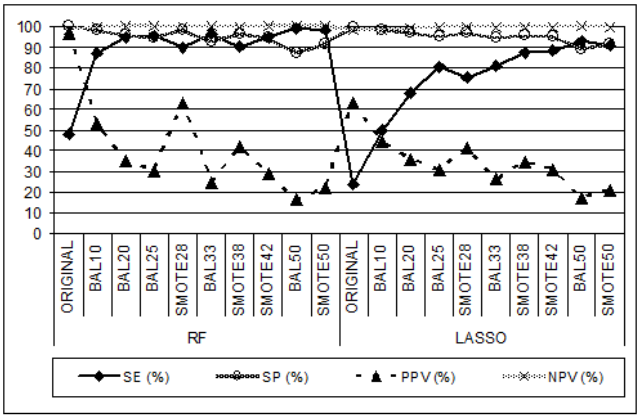

Figure 1. Comparison of the performances of the models trained with different training sets.

only with different prevalences, but also when the same prevalence is obtained in different ways (e.g. Bal50 and Smote50 get different results). The results achieved in terms of the "classical" metrics (Sensitivity, Specificity, Positive Predictive Value and Negative Predictive Value) are shown in Fig. 1. The best model in terms of the two golden metrics is RF with 1 000 trees, mtry = 39 and sample size = 0.632, trained with the Smote28 training set.

IV. Discussion

The Boruta algorithm selected 572 variables of 1 514, reducing the number of input columns of more than 62%.

The fact that age and gender are in the final set of input features confirms their importance: heart failure occurs more in old people and more in men than in women [32].

As regard the medical variables, the proportion of each group of attributes with respect to the total number of features slightly changes after the features selection stage. In particular, the weight of the procedures really decreases (from 25.1% to 16.96%), in favour of the performances, whose weight almost doubles (from 5.94% to 9.44%). Also the incidence of the diagnoses increases, passing from 48.88% to 54.37%. The weight of drugs and exemptions, on the contrary, remains almost the same.

Some of the selected variables are related to heart problems, e.g. the number of admissions and the number of days of hospitalizations for cardiomyopathies, secondary hypertension and congestive heart failure, the number of coronary heart bypass grafts, percutaneous transluminal coronary angioplasties, coronary thrombolyses, diagnostic cardiac catheterizations, insertions and replacements of cardiac pacemaker and other heart surgeries, the number of instrumental cardiology performances and cardiac medical examinations and the number of cardiovascular drugs taken. On the contrary, the exemptions for cardiovascular diseases are excluded. But most of the selected features are related to other problems, e.g. tumours, respiratory problems or nervous system diseases. This means that people with heart failure are usually affected by several comorbidities and so it can be wrong looking only at heart problems for the identification of the target population.

According to the presented results, the algorithm already used in Tuscany for the identification of heart failure patients (MaCro) [17] is facing a performance problem. It only uses codes of diagnoses for HF, cardiomyopathies or hypertension and exemptions for HF for the identification, considering nor drugs, procedures and performances nor diagnoses or exemptions for any other disease [33]. As a consequence, its results are not so good: both SE and PPV do not overcome the 50% (using the test set of this study to evaluate the performances). On the contrary, the tested ML algorithms trained with different training sets, except for the one with the original prevalence, achieve a very high SE (higher than 70%), even if the PPV is usually lower than 50% (Fig. 1). The only exception is represented by RF trained with the set with the 28% of prevalence obtained with SMOTE (Smote28): both SE and PPV are really greater than the ones achieved by the MaCro algorithm, while specificity (SP) and negative predictive value (NPV) remain almost unchanged, near 100%. Table IV shows the performances of RF compared with MaCro algorithm. RF outperforms MaCro in terms of the two golden metrics.

In light of the results achieved by the RF model, the employment of such a method for the HF patients identification can have a great impact on the improvement of the healthcare service. Only very few HF patients, in fact, are not correctly identified by the model, leading to early diagnoses for most of people and causing both the improvement in QoL and the reduction of the resources consumption due to repeated and avoidable hospitalizations. Also the false positives are few, meaning that not many people are considered ill by the algorithm even if they are healthy. Because of their little number, their wrong diagnosis can be corrected by the GPs in the following, who can exploit also laboratory tests and clinical data for a more precise evaluation.

However, to let the GPs trust and adopt this new model based on RF, the method cannot be a "black box". RF is

Table IV
Performances of MaCro (the algorithm already used in Tuscany) and Random Forest trained with the SMOTE28 set.

	MaCro	**RF SMOTE28**
SE (%)	50.72	89.63
SP (%)	98.39	98.74
PPV (%)	43.35	63.34
NPV (%)	98.79	99.75
F1-Score (%)	46.75	74.22
F2-Score (%)	49.05	82.76

an ensemble of trees (1 000 in this case). Every tree is built with a random subset of variables (39 in this case) and with a random subset of training samples (the 0.632% in this case). The final prediction is given by the "simple voting" of all the trees [26], [34]. Classification and Regression Trees (CART) are used as base-learners. CART is a decision tree, a sequential procedure that classifies a given input, splitting in a binary way the "best" feature at each node. The "best" feature is the one that maximizes the reduction of impurity, i.e. minimizes the Gini index [35]. The importance of each attribute A_k for every single tree is calculated as the difference between the error rate of the predictions obtained from the original data and the ones obtained with the random permutation of A_k. Averaging among importance measures for individual trees gives the importance of A_k for the RF. Table V shows the top 10 variables for the problem of identification of HF patients, according to the following metrics:

- *Accuracy Decrease* is the mean decrease of accuracy after the attribute is randomly permuted (the greater it is, the more important the variable is);
- *Gini Decrease* is the mean decrease of the Gini index by splitting on the feature (the greater it is, the more the purity of each node increases);
- *# Trees* is the total number of trees in which a split on the attribute occurs;
- *# Roots* is the total number of trees where the attribute is used for splitting at the root (i.e. the attribute maximizes the reduction of impurity).

Fig. 2 shows the distribution of minimal depth and its mean for the most significant variables, where minimal depth is the length of the path from the root to the node where the attribute is used for splitting (the smaller it is, the more relevant the feature is). Table VI shows the meaning of the acronyms of the features shown in Table V and Fig. 2.

The 10 most important variables, both in terms of the decrease of accuracy and of the mean minimal depth, include only age, drugs and performances, not necessarily related to the cardiac system. All the diagnoses, the procedures and the exemptions are not usually employed for splits, meaning that they do not discriminate well HF patients from non-HF ones. This explains the better performances of RF than MaCro algorithm and than Shultz's best one [18], the two models built with only administrative data. In addition, these results are comparable with the ones achieved by the models using also the clinical data, especially in terms of sensitivity.

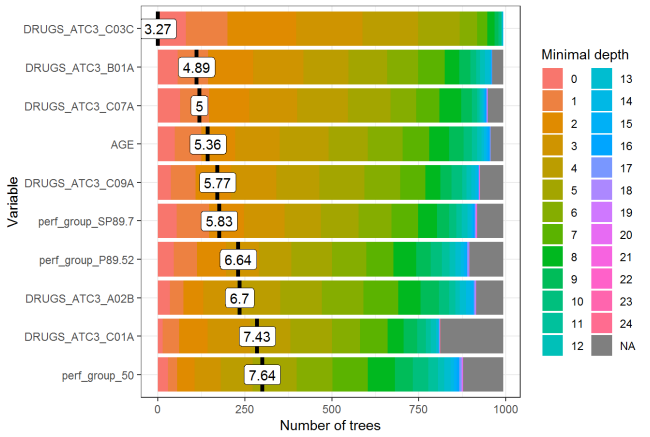

Figure 2. Distribution of minimal depth and its mean (marked by a vertical bar) for RF trained with Smote28.

V. Conclusion

Heart failure is one of the top causes of death all over the world, it generates disabling symptoms, making life very difficult, and the treatments for heart failure patients cover a great part of all the healthcare expenses. Identifying as soon as possible these patients becomes then crucial, in order to improve their QoL and to reduce the resources consumption.

Table V
Top 10 variables used by RF trained with Smote28, in order of importance (decreasing order of *Accuracy Decrease*).

Variable	**Accuracy Decrease (%)**	**Gini Decrease**	**# Trees**	**# Roots**
DRUGS_ATC3_C03C	0.063	198.25	993	81
DRUGS_ATC3_C07A	0.031	101.97	949	63
DRUGS_ATC3_B01A	0.021	111.4	962	57
DRUGS_ATC3_C09A	0.018	67.04	926	37
PERF_GROUP_SP89.7	0.016	96.76	918	54
DRUGS_ATC3_C01A	0.15	36.11	813	14
PERF_GROUP_79	0.014	21.87	829	10
AGE	0.013	79.38	957	49
DRUGS_ATC3_A02B	0.012	43.09	915	34
PERF_GROUP_P89.52	0.011	72.22	896	46

Table VI
Description of the most important features of the dataset.

Variable	**Description**
AGE	Age on 1^{st} January
PERF_GROUP_50	Number of clinical chemistry laboratory exams
PERF_GROUP_79	Number of other performances
PERF_GROUP_P89.52	Number of electrocardiograms
PERF_GROUP_SP89.7	Number of cardiological visits
DRUGS_ATC3_A02B	Number of drugs for peptic ulcer and gastro-oesophageal reflux
DRUGS_ATC3_B01A	Number of antithrombotic agents
DRUGS_ATC3_C01A	Number of cardiac glycosides
DRUGS_ATC3_C03C	Number of high-ceiling diuretics
DRUGS_ATC3_C07A	Number of beta blocking agents
DRUGS_ATC3_C09A	Number of plain ace inhibitors

This paper presents the performances evaluation of several machine learning methods for the identification of heart failure patients, starting from administrative data, the only ones available in Italy, and the explanation of the best model (random forest trained with a set with the 28% of prevalence), to make it transparent and interpretable. This model outperforms the algorithm currently used in Tuscany for the same aim, increasing both sensitivity and positive predictive value and considering at most features related to drugs and performances.

REFERENCES

[1] Heart Failure, https://www.nhs.uk/conditions/heart-failure/

[2] José Lopez-Sendon, "The heart failure epidemic", Medicographia, 2012.

[3] European Society of Cardiology, "Heart failure: Preventing disease and death worldwide", 2015, http://www.escardio.org/communities/HFA/Documents/whfa-whitepaper.pdf.

[4] Ahmed J. Aljaaf, Dhiya Al-Jumeily, Abir Jaafar Hussain, Tom A. Dawson, P. Fergus and M. Al-Jumaily, "Predicting the likelihood of heart failure with a multi level risk assessment using decision tree", Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), IEEE, April 2015.

[5] Evanthia E. Tripoliti, Theofilos G. Papadopoulos, Georgia S. Karanasiou, Katerina K. Naka and Dimitrios I. Fotiadis, "Heath failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques", Computer and Structural Biotechnology Journal, 2017, vol. 15, pp. 26-47.

[6] John J. V. McMurray and Simon Stewart, "Heart failure: epidemiology, aetiology and prognosis of heart failure", Heart (British Cardiac Society), 2000, vol. 83, no. 5, pp. 596-602.

[7] Mauro Di Bari, Claudia Pozzi, Maria Chiara Cavallini, Francesca Innocenti, Giorgio Baldereschi, Walter De Alfieri, Enrico Antonini, Riccardo Pini, Giulio Masotti and Niccolò Marchionni, "The diagnosis of heart failure in the community. Comparative validation of four sets of criteria in unselected older adults: the ICARe Dicomano Study", Journal of the American College of cardiology, 2004.

[8] Serguei Pakhomov, Susan A. Weston, Steven J. Jacobsen, Christopher G. Chute, Ryan Meverden and Veronique L. Roger, "Electronic medical records for clinical research: application to the identification of heart failure", The American Journal of Managed Care, 2007, vol. 13, no. 6, pp. 281-288.

[9] Saul Blecker, Stuart D. Katz, Leora I. Horwitz Gilad Kuperman, Hannah Park, Alex Gold and David Sontag, "Comparison of approaches for heart failure case identification from electronic health record data", JAMA Cardiol, 2016, vol. 1, no. 9, pp. 1014-1020.

[10] Roy J. Byrd, Steven R. Steinhubl, Jimeng Sun, Shahram Ebadollahi and Walter F. Stewart, "Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records", International Journal of Medical Informatics, 2014, vol. 83, no. 12, pp. 983-992.

[11] Guiqiu Yang, Yinzi Ren, Qing Pan, Gangmin Ning, Shijin Gong, Guolong Cai, Zhaocai Zhang, Li li and Jing Yan, "A heart failure diagnosis model based on support vector machine", Third International Conference on Biomedical Engineering anf Informatics, 2010.

[12] Farhad Solemanian Gharehchopogh and Zeynab A. Khalifelu, "Neural network application in diagnosis of patient: a case study", 2011.

[13] Chang Sik Son, Yoon Nyun Kim, Hyung Seop Kim, Hyoung Seob Park and Min Soo Kim, "Decision-making model or early diagnosis of congestive heart failure using rough set and decision tree approaches", Journal of Biomedical Informatics, 2012, vol. 45, no. 5, pp. 999-1008.

[14] Zerina Masetic and Abdulhamit Subasi, "Congestive heart failure detection using random forest classifier", Computer Methods and Programs in Biomedicine, 2016, vol. 130, pp. 54-65.

[15] Jionglin Wu, Jason Roy and Walter F. Stewart, "Prediction modeling using EHR data challenges, strategies and a comparison of machine learning approaches", Medical care, 2010, vol. 48, pp. 106-113.

[16] Yineng Zheng, Xingming Guo, Jian Qin, Shouzhong Xiao, "Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics", Computer Methods and Programs in Biomedicine, 2015, vol. 122, no. 3, pp. 372-383.

[17] MaCro database, https://proter.ars.toscana.it/macro.

[18] Susan E. Schultz, Deanna M. Rothwell and Karen Tu, "Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records", Chronic diseases and Injuries in Canada, 2013.

[19] mARSupio database, https://www.ars.toscana.it/marsupio/database/.

[20] Italian Law, no. 675/1996, Tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali, [Protection of persons and other subjects with regard to personal data processing], http://www.garanteprivacy.it/web/guest/home/docweb/-/docwebdisplay/ docweb/28335.

[21] Silvia Panicacci, Massimiliano Donati, Luca Fanucci, Irene Bellini, Francesco Profili and Paolo Francesconi, "Population health management exploiting machine learning algorithms to identify high-risk patients", 31^{st} International Symposium on Computer-Based Medical Systems, 2018.

[22] Aggregated Clinical Codes (ACC), http://www.salute.gov.it/imgs/C_17_pubblicazioni_1006_allegato.pdf.

[23] Anatomical Therapeutic Chemical classification system (ATC), https://www.whocc.no/atc/structure_and_principles/.

[24] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, 2002, vol. 16, pp. 321-357.

[25] Miron B. Kursa, Aleksander Jankowski, Witold R. Rudnicki, "Boruta a system for feature selection", Fundamenta Informaticae, 2010, vol. 101, pp. 271-285.

[26] Adele Cutler, D. Richard Cutler and John R. Stevens, "Random Forests", Machine Learning, 2011.

[27] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society. Series B (Methodological), 1996, vol. 58, no.1, pp 267-288.

[28] RF vs LASSO, https://healthcare.ai/visual-tour-lasso-random-forest/

[29] Why accuracy alone is a bad measure for classification tasks, https://tryolabs.com/blog/2013/03/25/why-accuracy-alone-bad-measure-classification-tasks-and-what-we-can-do-about-it/.

[30] F1-Score, https://clusteval.sdu.dk/1/clustering_quality_measures/18.

[31] F2-Score, https://clusteval.sdu.dk/1/clustering_quality_measures/5.

[32] Paresh A. Metha and Martin R. Cowie, "Gender and heart failure: a population perspective", Heart, 2006, vol. 92, pp. 14-18.

[33] Matilde Razzanelli, Irene Bellini e Paolo Francesconi, "La banca dati MaCro delle malattie croniche. Aggiornamento 2018.", https://www.ars.toscana.it/images/pubblicazioni/Collana_ARS/2018/documento_99/Doc_Ars_99_2018_MaCro.pdf.

[34] Leo Breiman, "Bagging predictors", *Machine Learning*, vol. 24, 1996.

[35] Leo Breiman, Jerome Friedman, Charles J. Stone and R. A. Olshen, "Classification and Regression Trees", Wadsworth International Group, 1984.