An Enhanced Random Forests Approach to Predict Heart Failure From Small Imbalanced Gene Expression Data

Davide Chicco and Luca Oneto

Abstract—Myocardial infarctions and heart failure are the cause of more than 17 million deaths annually worldwide. ST-segment elevation myocardial infarctions (STEMI) require timely treatment, because delays of minutes have serious clinical impacts. Machine learning can provide alternative ways to predict heart failure and identify genes involved in heart failure. For these scopes, we applied a Random Forests classifier enhanced with feature elimination to microarray gene expression of 111 patients diagnosed with STEMI, and measured the classification performance through standard metrics such as the Matthews correlation coefficient (MCC) and area under the receiver operating characteristic curve (ROC AUC). Afterwards, we used the same approach to rank all genes by importance, and to detect the genes more strongly associated with heart failure. We validated this ranking by literature review and gene set enrichment analysis. Our classifier employed to predict heart failure achieved MCC = +0.87 and ROC AUC = 0.918, and our analysis identified KLHL22, WDR11, OR4Q3, GPATCH3, and FAH as top five protein-coding genes related to heart failure. Our results confirm the effectiveness of machine learning feature elimination in predicting heart failure from gene expression, and the top genes found by our approach will be able to help biologists and cardiologists further our understanding of heart failure.

Index Terms—Heart failure, gene ranking, random forests, STEMI, infarction, gene expression, feature selection, machine learning, genetics, feature elimination

1 Introduction

CARDIOVASCULAR heart diseases are a global major health issue and are the cause of more than seventeen million deaths globally on an annual basis [1]. Myocardial infarction (MI), colloquially referred to as a heart attack, compromises a significant proportion of cardiovascular mortality. Myocardial infarction occurs when oxygen supply to a region of the heart is insufficient to oxygen demand; this results in damage to the myocardium, that is the cardiac muscle.

Cardiologists divide myocardial infarctions into two forms clinically: ST elevation myocardial infarction (STEMI) and non-ST elevation myocardial infarction (named NSTEMI or non-STEMI). This division refers to features seen on an electrocardiogram (ECG), a common and cheap clinical test available in almost every acute healthcare setting. STEMI is believed to occur when one of their heart's arteries is completely occluded, while NSTEMI is believed to be the result of an incomplete occlusion; although these rules generally hold true, they are not absolute. The STEMI-NSTEMI distinction is however critically important as the standard of care for STEMI almost always necessitates emergency percutaneous coronary intervention (PCI). Congestive heart failure is a complex disorder which often occurs in the aftermath of myocardial infarctions, and affects approximately 26 millions of people worldwide every year [2].

Several studies have shown that patients who experience STEMI are at a high risk of develop congestive heart failure, especially when they get elderly [3]. Identifying which patients are likely to develop congestive heart failure is an important task as it ensures patients can be connected with evidence based therapies which reduce their mortality and morbidity. In addition to the direct clinical impact that early identification might have, there is significant value in understanding what biological processes drive the development of heart failure as therapeutic interventions can be then be designed to target these processes. Machine learning can be a valuable tool in identifying which biological processes and genes are associated with the development of heart failure in patients post myocardial infarction.

Other studies employed computational intelligence algorithms to analyze gene expression data of patients with heart failure in the past. Pérez-Belmonte and colleagues [4], for example, employed Logistic Regression and statistical tests to investigate the thermogenic genes which might be involved in the pathogenesis of heart failure, from their gene expression. Huang and coauthors [5] and Kittleson and colleagues [6] took advantage of several computational intelligence algorithms to classify ischemic and non–ischemic heart failure from gene expression. Wang and colleagues [7] employed machine learning to propose a genetic signature for dilated cardiomyopathy.

Regarding infarctions and strokes, Fang *et al*. [8] took advantage of a Support Vector Machine model applied to gene expression to detect the genes that cause the risk of myocardial infarction. Neelima and colleagues [9] employed machine learning to perform a meta-heuristic analysis of gene expression related to myocardial infarction. O'Connell and coauthors [10] applied a genetic algorithm k-nearest neighbours method to gene expression data of peripheral blood to identify ischemic strokes. In a subsequent study, the same team [11] employed computational intelligence to verify the effectiveness of a stroke-associated gene signature.

To the best of our knowledge, no study involving applications of machine learning methods to gene expression data of patients diagnosed with STEMI exists in the scientific literature at the moment.

In this paper, we analyzed a publicly available gene expression dataset of patients who experienced STEMI and of healthy control individuals, released by Maciejak and colleagues [12]. The dataset includes genomic data from patients who experienced STEMI and follows them in the month afterwards in order to identify which patients ultimately developed congestive heart failure (Section 2). After data preprocessing, we retrieved the genes associated to these gene expressions, and applied a few machine learning classifiers to predict heart failure among the STEMI patients. We then took advantage of machine learning again to detect the most important genes related to heart failure, and validated the obtained gene ranking through four techniques: literature review and gene set enrichment analysis. Our prediction results show the effectiveness of our methods, and confirms the predictive power of machine learning applied to genomics for health and cardiology goals.

2 Dataset

We analyzed a dataset collected by Maciejak and colleagues [12] which they made publicly available on Gene Expression Omnibus (GEO) in May 2015. The dataset consists of gene expressions of the messenger RNA (mRNA) levels in pheriperal mononuclear cells of blood samples of 111 patients with ST-segment elevation myocardial infarction (STEMI) and of 46 healthy controls, that are individuals with a stable coronary artery disease and without a history of myocardial infarction [12]. The 111 patients with STEMI were admitted to the First Chair and Department of Cardiology of the Medical University of Warsaw (Warszawski Uniwersytet

- *Davide Chicco is with the Krembil Research Institute, Toronto, ON M5T 0S8, Canada. E-mail: davidechicco@davidechicco.it.*
- *Luca Oneto is with the Università di Genova, 16126 Genoa, Italy. E-mail: luca.oneto@gmail.com.*

Manuscript received 14 Apr. 2020; revised 16 Nov. 2020; accepted 27 Nov. 2020. Date of publication 1 Dec. 2020; date of current version 8 Dec. 2021.
(Corresponding author: Davide Chicco.)
Digital Object Identifier no. 10.1109/TCBB.2020.3041527

Medyczny, Warsaw, Poland, EU) between 2010 and 2013 [12]. We did not include any other clinical features, such as demographics variables.

Their original dataset contains gene expressions for three timepoints for the STEMI patients, but we included only the first timepoint (first day of admission) in our analysis. Maciejak and colleagues [12] also separated the 111 STEMI patients into a heart failure group (9 individuals) and a non-heart failure group (102 individuals), by analyzing their plasma N-terminal prohormone of brain natriuretic peptide (NT-proBNP) and left ventricular ejection fraction (LVEF). Through the NT-proBNP and LVEF analysis, the dataset curators split the STEMI patients into four equal groups. They then selected the individuals who had a high level of NT-proBNP and low LVEF in the first and fourth quartiles as the 9 heart failure patients [12].

For the heart failure classification task, we assigned the label *true* to the 9 patients with heart failure, and the label *false* to the 102 patients without heart failure. This dataset configuration had therefore 91.89 percent negative data instances (false labels), and 8.11 percent positive data instances (true labels).

Our data preprocessing associated 23,699 genes to this gene expressions. Thus, we applied the machine learning classifiers to a table made of 111 rows and 23,700 columns in the heart failure prediction task. We will describe the data processing more in detail in the next Section (Section 3.1).

3 Methods

In this section, we describe the preprocessing we applied to the original dataset to remove noise and to correctly associate genes to the gene expressions (Section 3.1), and then the statistical operations we applied to make the dataset ready for an effective machine learning binary classification (Section 3.2). We then described the details of the supervised machine learning approach we used for our heart failure binary classification and the methods we employed to detect the most relevant genes for heart failure.

3.1 Data Preprocessing

As previously mentioned, we analyzed a dataset of microarray gene expression data generated through the Affymetrix Human Gene (HuGene) 1.0 ST [transcript (gene) version] GPL6244 platform [12]. Even when gene expression data are generated from the same series of experiments, their results can be influenced by many different aspects, such as different reagents, different chemicals, different technicians, or different experiment conditions [13]. We call *batch* microarrays elaborated in the same laboratory, in the same short time, using the same biotechnology platform, and we call *batch effects* the sum of these experimental variations [13].

The removal of the batch effects in a gene expression dataset is pivotal to have all the gene expression samples consistent, and to generate a coherent analysis where statistical correlations have real biological meanings and are not caused by the batch effects instead.

We based our batch correction approach on the sample date, which is the date of the microarray scanner used to generate the microarray data, and employed the batch correction ComBat method [14].

3.2 Data Analytics

The problem described above can be easily mapped in a classical binary classification framework [15]. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space, consisting of f features, and let $\mathcal{Y} = \{0, 1\}$ be the output space. Conventionally we will indicate with 1 a positive outcome and with 0 a negative outcome. Let $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$ $\forall i \in \{1, \dots, n\}$, be a sequence of $n \in \mathbb{N}^*$ samples drawn independently from an unknown probability distribution μ over $\mathcal{X} \times \mathcal{Y}$. Let us consider a model (function) $f : \mathcal{X} \to \mathcal{Y}$ chosen from set $\mathcal{F}$ of possible hypotheses. An algorithm $\mathscr{A}_{\mathcal{H}} : \mathcal{D}_n \times \mathcal{F} \to f$ characterized by its hyperparameters $\mathcal{H}$ selects a model inside a set of possible ones based on the available dataset. Note that many algorithms for solving binary classification problems exists in literature [16] but surely Random Forests (RF) [17] (Supplementary Information, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.3041527) has shown to be one of the most powerful one [18], [19] especially in biomedical applications [20], [21], [22], [23].

Once the model is built one can investigate how and how much the model is affected by the different features that have been exploited to build the model itself during the feature ranking procedure (FR). This phase allows to understand the most important features and to remove the uninformative one, with the feature elimination (FE) phase, which may improve the quality of the final model due to the curse of dimensionality [24]. We describe the approach exploited in this paper for FR in Section 3.2.1.

The error of f in approximating $\mathbb{P}\{Y \mid X\}$ is measured by a prescribed metric $M : \mathcal{F} \to \mathbb{R}$. Note that, many different metrics are available in literature for binary classification which may provide insights on the performance of the model [25] (Supplementary Information, available online). Note also that, for binary classification problems, $\mathcal{D}_n$ may be imbalanced (namely the $|\{(X, Y) \in \mathcal{D}_n : Y = 0\}|$ may be $\gg$ or $\ll$ than the $|\{(X, Y) \in \mathcal{D}_n : Y = 1\}|$) and this may result in classifiers which produce unsatisfactory results on one of the two classes resulting in unsatisfactory metrics performance [26]; for this reason we discuss the problem and show how we tackle it (Section 3.2.2).

To tune the performance of the $\mathscr{A}_{\mathcal{H}}$, namely to select the best set of hyperparameters, and to estimate the performance of the final model according to the desired metrics, a Model Selection (MS) and Error Estimation (EE) phase needs to be performed [27]. We report the approach exploited in this paper for MS and EE purposes in Section 3.2.3.

3.2.1 Feature Ranking and Elimination

Once the models are built, it is possible to investigate how these models are affected by the different features used in the model identification phase, to understand if the models have also a foundation which relies on the underline phenomena or if the model just captures spurious correlations [28]. This procedure is called feature ranking (FR) and allows to detect if the importance of those features, that are known to be relevant from a physical perspective, are appropriately taken into account by the learned models. The failure of the computational model to properly account for the relevant features might indicate poor quality in the measurements or spurious correlations. FR therefore represents an important step of model verification, since it should generate consistent results with the available knowledge of the phenomena under exam. Moreover, FR allows to remove the uninformative features, with the feature elimination (FE) phase [29]. Retraining the models with just the informative variables usually improves the quality of the final model, because of the curse of dimensionality [24].

FR methods based on RF are among the most effective machine learning techniques [30], [31], particularly in the context of bioinformatics [20], [21] and health informatics [32]. Several measures are available for feature importance in RF. One approach is based on the Gini Importance or Mean Decrease in Impurity (MDI) which calculates each feature importance as the sum over the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits. Another powerful approach is the one based on the Permutation Importance or Mean Decrease in Accuracy (MDA), where the importance is assessed for each feature by removing the association between that feature and the target. This is achieved by randomly permuting [33] the values of the variable and measuring the resulting increase in error. The influence of the correlated features is also removed. In details, for every tree, two quantities are computed: the first one is the error on the out-of-bag samples as they are used during prediction, while the second one is the error on the out-of-bag samples after a random permutation of the values of a variable. These two values are then subtracted, and the average of the result over all the trees in the ensemble is the raw importance score for the variable under exam. Both MDI and MDA can be adopted since they can be easily carried out during the main prediction process inexpensively.

Despite most of the actual research studies in bioinformatics agrees on the effectiveness of MDI and MDA, they also agree that, when the number of samples is small, these methods may be unstable [22], [23], [34]. For this reason, in this work, instead of running the FR, we sub-sample $\mathcal{D}_n$ such that $\mathcal{S}_m \subset \mathcal{D}_n$ with $m = |\mathcal{S}_m| = p_{\mathrm{FR}} n$ (not overlapped with the test set), namely we randomly sample without replacement $100 \cdot p_{\mathrm{FR}}\%$ of the data in $\mathcal{D}_n$, we perform the FR using $\mathcal{S}_m$ and we repeat the procedure n_{FR} times. The final rank of a feature will be the mode of its ranking position in the different repetitions of the ranking position, and the MDI and MDA are the median value over the different repetitions.

Having the ranking of the features we now have a criteria for FE. Basically we will remove all the variables for which the median MDI or MDA is three order or magnitude smaller than the most important features. Then we will retrain the model with just the subset of the variables selected with this strategy, and we will perform again the FR procedure to have a more accurate method for ranking the most important features.

3.2.2 Handling Imbalanced Classes

Data available in bioinformatics for binary classification are often unbalanced [35], [36], [37]. However, most learning algorithms do not work well with imbalanced datasets and tend to poorly perform on the minority class; for these reasons, several techniques have been developed to address this issue [26].

The first step toward the solution of this problem is to apply the appropriate evaluation metrics for model generated using imbalanced data [38]. For example, overall accuracy is a very dangerous metric in this context since the more unbalanced is the dataset the more this metric tends to promote models which poorly perform on the minority class. For this reason, in this study we also included other metrics like Precision/Specificity, Recall/Sensitivity, F$_1$ score, MCC, and AUC which are more suited for the case of imbalanced data (Supplementary Information, available online).

The second step toward the mitigation of the effects of having an unbalanced dataset is to modify the algorithm or the data, but currently the most practical and effective method involves the resampling of the data in order to synthesize a balanced dataset [26]. For this purpose we can under- or over-sample the dataset. Undersampling balances the dataset by reducing the size of the abundant class. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modelling. Note that this method wastes a lot of information (many samples may be lost). For this reason, the over-sampling strategy is more often exploited. It tries to balance the dataset by increasing the size of rare samples. Rather than removing abundant samples, new rare samples are generated (for example by repetition, by bootstrapping, or by synthetic minority).

In this study, we initially tried both the under-sampling and the over-sampling techniques. Since under-sampling generated worse results than over-sampling, we decided to focus solely on over-sampling.

Often these two steps are still not enough to solve the problem and, with Random Forests (FR), we can perform another step toward the mitigation of the effects of having an unbalanced dataset. In particular RF is an ensemble classifier which means that the decision is taken as combination of many weak classifiers. A typical approach is to use the majority voting strategy to reach the final decision but, for unbalanced dataset, we can mitigate the tendency of the RF to predict the majority class by substituting this strategy with a thresholded majority voting strategy.

Basically, instead of performing majority voting (which, in binary classification, means that we take the class voted more than 50 percent of the times) we use a tunable threshold $T \in (0, 100)$ and we take the class voted more than $T\%$ of the times for one class (for example, the majority class) and the class voted more than $(100 - T)\%$ of the times for the other one (e.g., the minority class).

In this case, T becomes another hyperparameter of the RF to be tuned.

3.2.3 Model Selection and Error Estimation

MS and EE deal with the problem of tuning and assessing the performance of a learning algorithm [27]. Resampling techniques like k-fold cross validation and non-parametric bootstrap are often used by practitioners because they work well in many scientific domains [39]. Other alternatives exist, which represent bases in the Statistical Learning Theory and give more insight into the learning process. Examples of methods in this last category are: the seminal work of the Vapnik-Chervonenkis Dimension, its improvement with the Rademacher Complexity, the theory of compression, the Algorithmic Stability breakthrough, the PAC-Bayes theory, and more recently the Differential Privacy theory [27].

In this work we will exploit the resampling techniques which rely on a simple idea: the original dataset $\mathcal{D}_n$ is resampled once or many (n_r) times, with or without replacement, to build three independent datasets called learning, validation and test sets, respectively $\mathcal{L}_l^r$, $\mathcal{V}_v^r$, and $\mathcal{T}_t^r$, with $r \in \{1, \ldots, n_r\}$. Note that $\mathcal{L}_l^r \cap \mathcal{V}_v^r = \emptyset$, $\mathcal{L}_l^r \cap \mathcal{T}_t^r = \emptyset$, $\mathcal{V}_v^r \cap \mathcal{T}_t^r = \emptyset$, and $\mathcal{L}_l^r \cup \mathcal{V}_v^r \cup \mathcal{T}_t^r = \mathcal{D}_n$ for all $r \in \{1, \ldots, n_r\}$.

Then, to select the best combination of the hyperparameters $\mathcal{H}$ in a set of possible ones $\mathfrak{H} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}$ for the algorithm $\mathscr{A}_{\mathcal{H}}$ or, in other words, to perform the MS phase, the following procedure has to be applied

$$\mathcal{H}^* : \quad \arg\min_{\mathcal{H} \in \mathfrak{H}} \sum_{r=1}^{n_r} M(\mathscr{A}_{\mathcal{H}}(\mathcal{L}_l^r), \mathcal{V}_v^r), \tag{1}$$

where $\mathscr{A}_{\mathcal{H}}(\mathcal{L}_l^r)$ is a model built with the algorithm $\mathscr{A}$ with its set of hyperparameters $\mathcal{H}$ and with the data $\mathcal{L}_l^r$ and where $M(f, \mathcal{V}_v^r)$ is a desired metric. Since the data in $\mathcal{L}_l^r$ are independent from the ones in $\mathcal{V}_v^r$, the idea is that $\mathcal{H}^*$ should be the set of hyperparameters which allows to achieve a small error on a data set that is independent from the training set.

Then, to evaluate the performance of the optimal model which is $f^*_{\mathscr{A}} = \mathscr{A}_{\mathcal{H}^*}(\mathcal{D}_n)$ or, in other words, to perform the EE phase, the following procedure has to be applied:

$$M(f^*_{\mathscr{A}}) = \frac{1}{n_r} \sum_{r=1}^{n_r} M(\mathscr{A}_{\mathcal{H}^*}(\mathcal{L}_l^r \cup \mathcal{V}_v^r), \mathcal{T}_t^r). \tag{2}$$

Since the data in $\mathcal{L}_l^r \cup \mathcal{V}_v^r$ are independent from the ones in $\mathcal{T}_t^r$, $M(f^*_{\mathscr{A}})$ is an unbiased estimator of the true performance, measured with the metric M, of the final model [27].

If $n_r = 1$, if l, v, and t are aprioristically set such that $n = l + v + t$, and if the resample procedure is performed without replacement, we obtain the hold out method [27]. For implementing the complete nested k-fold cross validation, instead, we need to set $n_r \leq \binom{n}{k}\binom{n-\frac{n}{k}}{k}$, $l = (k-2)\frac{n}{k}$, $v = \frac{n}{k}$, and $t = \frac{n}{k}$ and the resampling must be done without replacement [39]. Finally, for implementing the nested non-parametric bootstrap, $l = n$ and $\mathcal{L}_l^r$ must be sampled with replacement from $\mathcal{D}_n$, while $\mathcal{V}_v^r$ and $\mathcal{T}_t^r$ are sampled

TABLE 1
Heart Failure Prediction Results on Patients Having STEMI

method	**MCC**	F_1 score	accuracy	TP rate	TN rate	PR AUC	ROC AUC
Random Forests FE	0.870 ± 0.054	0.878 ± 0.075	0.982 ± 0.067	0.843 ± 0.071	0.994 ± 0.072	0.802 ± 0.055	0.918 ± 0.067
XGBoost FE	0.829 ± 0.061	0.838 ± 0.059	0.971 ± 0.077	0.819 ± 0.060	0.981 ± 0.057	0.787 ± 0.040	0.898 ± 0.057
AdaBoost FE	0.796 ± 0.054	0.811 ± 0.045	0.951 ± 0.059	0.807 ± 0.058	0.966 ± 0.057	0.781 ± 0.051	0.888 ± 0.067
Logistic Regression FE (Lasso)	0.728 ± 0.059	0.719 ± 0.045	0.918 ± 0.051	0.826 ± 0.071	0.935 ± 0.066	0.779 ± 0.041	0.880 ± 0.068
Random Forests all	0.623 ± 0.037	0.643 ± 0.038	0.782 ± 0.044	0.624 ± 0.039	0.913 ± 0.071	0.689 ± 0.054	0.782 ± 0.068
XGBoost all	0.605 ± 0.043	0.616 ± 0.058	0.779 ± 0.058	0.611 ± 0.047	0.901 ± 0.086	0.678 ± 0.041	0.771 ± 0.052
AdaBoost all	0.583 ± 0.044	0.594 ± 0.038	0.761 ± 0.052	0.598 ± 0.047	0.898 ± 0.057	0.669 ± 0.047	0.765 ± 0.037
Logistic Regression all	0.524 ± 0.042	0.543 ± 0.042	0.732 ± 0.050	0.554 ± 0.037	0.892 ± 0.064	0.529 ± 0.034	0.632 ± 0.039

FE: method applied after feature elimination. all: method applied on all the features. MCC: Matthews correlation coefficient. TP rate: true positive rate (sensitivity, recall). TN rate: true negative rate (specificity). Confusion matrix threshold for MCC, F_1 score, accuracy, TP rate, TN rate: $\tau = 0.5$. PR: precision-recall curve. ROC: receiver operating characteristic curve. AUC: area under the curve. MCC: worst value = –1 and best value = +1. F_1 score, accuracy, TP rate, TN rate, PR AUC, ROC AUC: worst value = 0 and best value = 1. We report the formulas of MCC F_1 score, accuracy, TP rate, TN rate, PR AUC, ROC AUC in the Supplementary Information, available online.

without replacement from the sample of $\mathcal{D}_n$ that have not been sampled in $\mathcal{L}_l^r$ [39].

Note that, for the bootstrap procedure: $n_r \leq \binom{2n-1}{n}$. In this paper, we exploit the complete nested k-fold cross validation because it represents the state-of-the-art approach [27], [39].

Hyper-Parameter Optimization. For the selection of the hyperparameters of the method used, we followed these steps:

1) We built a model using the MS strategy where we set $k = n/2$ and $n_r = 1000$. It means that the size of the training set is 107 patients, the size of the validation set is 2 patients, and the size of the test set is 2 patients;
2) During the MS we searched the hyperparameters using the following ranges: $n_b = n$, $n_v \in d^{\frac{1}{16},\frac{1}{8},\frac{1}{4},\frac{1}{2}}$, $n_d = \infty$, $n_t = 1000$, and $T \in \{5, 10, 15, \ldots, 95\}$;
3) We reported the results using the EE strategy and previously introduced the metrics together with the standard deviation;
4) We performed the FR and FE where we set $p_{\mathrm{FR}} = 0.9$ and $n_{\mathrm{FR}} = 1000$.

Since we used k-fold cross-validation, each data instance of the whole dataset is included in the test set, in turn.

In our results, we give more importance to the scores of MCC because it is the only binary evaluation rate that takes into account the ratio of the negative set and positive set [40], [41], rather than using other metrics such as ROC AUC [42].

3.2.4 Other Methods

To highlight the impact of feature elimination (Section 3.2.1), we also performed the binary classification through a traditional Random Forests classifier without this phase.

For a general comparison, we then performed the binary classification through other methods and their feature elimination variants: Logistic Regression [43], [44], Extreme Gradient Tree Boosting (XGBoost) [45], and Adaptive Boosting (AdaBoost) [46].

3.3 Validation of the Data Analytics Results

After generating the ranking of the most impactful genes for heart failure through Random Forests, we decided to employ several techniques to validate this ranking.

Literature Review . To validate our gene ranking, we manually looked for scientific publications associating the top genes to heart failure or cardiology in general [47]. We manually queried PubMed [48] and Google Scholar [49] by inserting the symbol of the gene and the keywords “heart failure” or “heart”. If we could not find any scientific publication related to heart and the specific gene, we looked for articles involving genes of its same gene family.

Gene Set Enrichment Analysis (GSEA) . Gene set enrichment analysis (GSEA) techniques associate genes to functional annotations, and then select the ones with highest statistical significance as the most *relevant* pathways for those genes [50]. GSEA can also be used to validate the predicted association between a tissue and a set of genes, by checking if the tissue of the functional annotations found is the same of the predicted tissue [51]. In our GSEA analysis, we used g:Profiler [52], an online tool that reads in a list of genes and reports a list of Gene Ontology annotations that are statistically associated to the genes.

4 Results

In this section, we first report and describe the results we obtained for the heart failure classification and for the gene ranking.

We report the results of our heart failure prediction in Table 1. Our enhanced Random Forests classifier was able correctly predict the majority of positive data instances and negative data instances, by reporting MCC = +0.870, with a sensitivity of 0.843 and a sensitivity of 0.994.

The Random Forests approach was able to outperform all the other classifiers, both as the baseline model and after feature elimination, and to outperform the traditional Random Forests classifier without FE (Table 1). XGBoost resulted being the second top performing method both in the “after feature elimination” mode and in the traditional mode, followed by AdaBoost.

The binary classification results show also that the feature elimination enhanced all the methods, leading to a MCC increase of 39.64 percent for Random Forests, of 36.54 percent for XGBoost, of 26.76 percent for AdaBoost, and of 39.93 percent for Logistic Regression. These results confirm the effectiveness of feature elimination.

Since Random Forests with feature elimination achieved the top results in the binary classification, we reported the ranking of the top 23 genes, that fall in the top 0.1 percent percentile, obtained through this technique in Table 2. We validated the appropriateness and the relation of our gene ranking to heart failure through literature review and gene set enrichment analysis (Section 3.3).

Validation Through Literature Review . We performed a systematic literature search to find scientific publications confirming the association of heart failure and the top 23 genes ranked by our machine learning approach (Table 2). First, we discarded the pseudogenes (such as NPM1P8 and RN7SL612P), that are genes which lost functionality in gene expression or in the capability to code proteins [53], genes related to uncharacterized proteins (such as HSP90AA4P), and RNA genes that are actually non-coding RNA molecules (such as AL035696.1, MIR124-3, MIR3620). Among the first positions related to protein-coding genes, we found KLHL22, which has a role in congenital heart disease [54], WDR11 which is known to be related to cardiac anomaly [55], and OR4Q3, that is associated to congenital heart disease [56].

Regarding the following protein-coding genes in the ranking (GPATCH3, FAH, PKP3, DOK2, PASD1), we found only an article

TABLE 2
Heart Failure-Associated Gene Ranking Results Found by Random Forests

final position	gene	aggregated position	Gini position	Gini importance	accuracy position	accuracy importance
1	AL035696.1	5	1	6.530	4	0.456
2	KLHL22	7	4	4.460	3	0.504
3	NPM1P8	7	6	3.700	1	0.592
4	WDR11	10	3	4.960	7	0.407
5	OR4Q3	12	2	4.830	10	0.369
6	GPATCH3	16	5	3.390	11	0.385
7	FAH	17	11	3.270	6	0.444
8	PKP3	18	9	3.070	9	0.391
9	DOK2	19	14	2.430	5	0.451
10	PASD1	20	8	3.000	12	0.372
11	METTL7B	21	13	2.580	8	0.376
12	AMMECR1	24	22	0.613	2	0.506
13	SLFN5	24	7	3.660	17	0.340
14	HSP90AA4P	25	12	2.360	13	0.340
15	CRB3	31	10	2.980	21	0.293
16	MIR124-3	33	19	1.850	14	0.331
17	CUX1	34	16	1.570	18	0.323
18	MIR3620	35	20	0.955	15	0.344
19	CTU1	37	15	2.250	22	0.279
20	DEDD2	37	21	0.887	16	0.334
21	PRKAR1A	37	17	2.100	20	0.296
22	RN7SL612P	37	18	1.680	19	0.336
23	ESYT2	46	23	0.550	23	0.281

Results of the top 23 genes out of 23,699 obtained through machine learning approach, ranked through the Borda's method. Gini position: position of the gene in the Random Forests Gini impurity ranking. Accuracy position: position of the gene in the Random Forests accuracy decrease ranking. Aggregated position: sum of the Gini ranking position and of the accuracy ranking position.

TABLE 3
Results of the Gene Set Enrichment Through g:Profiler

term name	sub-ontology	term ID	URL	p-value
fumarylacetoacetase activity	molecular function	GO:0004334	[65]	4.147×10^{-2}

Gene Ontology term associated by g:Profiler [52] to the top 23 genes found through our machine learning approach.

associating DOK2 coronary artery disease [57]. The METTL7B is overexpressed in the heart atrial appendage [58], while AMMECR1 is known to be related to heart activation [59]. SLFN5 plays a role in the cardiovascular System [60], and CRB3 plays one in congestive heart failure after childhood cancer [61].

We found an article associating CUX1 to pediatric dilated cardiomyopathy [62], and no article related to cardiology for gene CTU1.

DEDD2 results being highly expressed in the heart [63], while PRKAR1A plays a role in heart development [64].

We found no article relating ESYT2 to cardiac biology.

To recap, we our literature review found relationships to the heart for most of the top protein-coding genes (KLHL22, WDR11, OR4Q3, METTL7B, AMMECR1, SLFN5, CRB2, CUX1, DEDD2, and PRKAR1A) but not for six of them (GPATCH3, FAH, PKP3, PASD1, CTU1, and ESYT2).

Validation Through Gene Set Enrichment Analysis . To further validate our gene ranking, we applied a gene set enrichment analysis (GSEA) on the top 23 genes of the ranking, equal to the top 0.1 percent percentile. We inputed these 236 genes to g:Profiler [52], and analyzed the Gene Ontology annotations that g:Profiler associated to the selected genes. We filtered the annotations having significance p-value lower than 0.005, and computed the significance threshold through the Bonferroni correction

g:Profiler associated a Gene Ontology annotation to the 23 genes: *fumarylacetoacetase activity* (Table 3). Fumarylacetoacetate is an enzyme encoded by the human FAH gene.

The fumarylacetoacetase enzyme correlates to cardiomyopathy: a deficiency of fumarylacetoacetase hydrolase (FAH) can generate type I tyrosinemia, a congenital metabolic condition that can cause reversible hypertrophic cardiomyopathy among children [66].

This gene set enrichment associated a cardiomyopathy related Gene Ontology annotations to the top 23 genes of the ranking generated by our random forests approach. These results therefore additionally confirm the robustness and the validity of our machine learning gene ranking.

5 Discussion and Conclusions

Heart failures and myocardial infarctions kill millions of people every year, especially in developing countries and among elderly populations. Predicting these two lethal conditions in patients can be time-consuming and require technologies that might be absent from many hospitals. In this context, machine learning can provide a more efficient, fast, effective way to detect them. In this project, we analyzed a publicly available dataset of gene expressions of patients having STEMI. Among the 111 STEMI patients, we separated the 9 patients having heart failure from the 102 patients without heart failure. After a data preprocessing phase and a data engineering phase (involving gene expression batch correction, gene names retrieval, dimensionality reduction, and undersampling class adjustment), we applied several machine learning classifiers to predict heart failure among the STEMI patients.

Our results show that machine learning enhanced with feature elimination, can predict heart failure in STEMI patients with high accuracy. Moreover, we showed that the Random Forests method can recognize genes particularly impactful to heart failure. We validated this list of genes through two different validation techniques, which confirmed the soundness of our approach.

Our Random Forests variant method and our results about heart failure genes can have a strong clinical significance. Clinical experts, in fact, can apply our proposed method to datasets of other disease to detect rankings of genes significant for those diseases, too. Moreover, medical doctors and biologists can treasure the list of genes related to heart failure that we present in this study (Table 2) and use them for diagnostic purposes or to plan wet-lab analyses about heart attack genetics.

Regarding limitations, we have to report that the small size of the patients having heart failure (9 individuals) restricts the generalisability of our approach. Clearly, if we had a dataset with a higher number of patients, our computational approach would have been able to obtain more robust and reliable results, especially for the heart failure gene ranking.

In the future, we aim at applying our computational techniques to other heart failure datasets, to obtain further confirmation of the genes we found to be most related to these conditions. We also plan to apply these methods to gene expression datasets of other heart-related conditions. Additionally, we aim at investigating the semantic similarity among the genes detected in our study [67].

Data Availability

The dataset is publicly available on Gene Expression Omnibus (GEO) at the following URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59867

Competing Interests

The authors declare they have no competing interests.

Acknowledgments

The authors thank Pingzhao Hu for his suggestions regarding the batch correction phase.

References

[1] S. Mendis *et al.*, "World health organization definition of myocardial infarction: 2008–09 revision," *Int. J. Epidemiol.*, vol. 40, no. 1, pp. 139–146, 2010.

[2] G. Savarese and L. H. Lund, "Global public health burden of heart failure," *Card. Failure Rev.*, vol. 3, no. 1, 2017, Art. no. 7.

[3] J. M. I. H. Gho *et al.*, "Heart failure following STEMI: A contemporary cohort study of incidence and prognostic factors," *Open Heart*, vol. 4, no. 2, 2017, Art. no. e000551.

[4] L. M. Pérez-Belmonte *et al.* "Expression of epicardial adipose tissue thermogenic genes in patients with reduced and preserved ejection fraction heart failure," *Int. J. Med. Sci.*, vol. 14, no. 9, 2017, Art. no. 891.

[5] X. Huang *et al.*, "A comparative study of discriminating human heart failure etiology using gene expression profiles," *BMC Bioinf.*, vol. 6, no. 1, 2005, Art. no. 205.

[6] M. M. Kittleson *et al.*, "Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy," *Circulation*, vol. 110, no. 22, pp. 3444–3451, 2004.

[7] H. Wang and H. Zheng, "Signature genes in human heart failure based on gene expression analysis: Can we identify a unique set?" in *Proc. 8th IEEE Int. Conf. Bioinf. Bioeng.*, 2008, pp. 1–6.

[8] H.-Z. Fang, D.-L. Hu, Q. Li, and S. Tu, "Risk gene identification and support vector machine learning to construct an early diagnosis model of myocardial infarction," *Mol. Med. Rep.*, vol. 22, no. 3, pp. 1775–1782, 2020.

[9] E. Neelima and M. S. P. Babu, "MAGED: Metaheuristic approach on gene expression data: Predicting the coronary artery disease and the scope of unstable angina and myocardial infarction," *Global J. Comput. Sci. Technol.*, vol. 16, pp. 1–7, 2016.

[10] G. C. O'Connell *et al.*, "Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke," *NPJ Genomic Med.*, vol. 1, no. 1, pp. 1–9, 2016.

[11] G. C. O'Connell, P. D. Chantler, and T. L. Barr, "Stroke-associated pattern of gene expression previously identified by machine-learning is diagnostically robust in an independent patient population," *Genomics Data*, vol. 14, pp. 47–52, 2017.

[12] A. Maciejak *et al.*, "Gene expression profiling reveals potential prognostic biomarkers associated with the progression of heart failure," *Genome Med.*, vol. 7, no. 1, 2015, Art. no. 26.

[13] C. Chen *et al.*, "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods," *PLoS One*, vol. 6, no. 2, 2011, Art. no. e17238.

[14] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[15] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[16] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[18] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.

[19] M. Wainberg, B. Alipanahi, and B. J. Frey, "Are random forests truly the best classifiers?," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3837–3841, 2016.

[20] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinf.*, vol. 7, no. 1, 2006, Art. no. 3.

[21] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*. Berlin, Germany: Springer, 2012.

[22] H. Wang, F. Yang, and Z. Luo, "An experimental study of the intrinsic stability of random forest variable importance measures," *BMC Bioinf.*, vol. 17, no. 1, 2016, Art. no. 60.

[23] M. B. Kursa, "Robustness of random forest-based gene selection methods," *BMC Bioinf.*, vol. 15, no. 1, 2014, Art. no. 8.

[24] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning and Data Mining*. Berlin, Germany: Springer, 2017.

[25] C. C. Aggarwal, *Data Mining: The Textbook*. Berlin, Germany: Springer, 2015.

[26] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017.

[27] L. Oneto, *Model Selection and Error Estimation in a Nutshell*. Berlin, Germany: Springer, 2019.

[28] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. Mar., pp. 1157–1182, 2003.

[29] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Berlin, Germany: Springer, 2007.

[30] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2008, pp. 313–325.

[31] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, 2010.

[32] D. Chicco and C. Rovelli, "Computational prediction of diagnosis and feature selection on mesothelioma patient health records," *PLoS One*, vol. 14, no. 1, 2019, Art. no. e0208737.

[33] P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Berlin, Germany: Springer, 2013.

[34] M. L. Calle and V. Urrea, "Letter to the editor: Stability of random forest importance measures," *Brief. Bioinf.*, vol. 12, no. 1, pp. 86–89, 2010.

[35] K. F. Kerr, "Comments on the analysis of unbalanced microarray data," *Bioinformatics*, vol. 25, no. 16, pp. 2035–2041, 2009.

[36] R. Laza, R. Pavón, M. Reboiro-Jato, and F. Fdez-Riverola, "Evaluating the effect of unbalanced data in biomedical document classification," *J. Integrative Bioinf.*, vol. 8, no. 3, pp. 105–117, 2011.

[37] K. Han, K.-Z. Kim, and T. Park, "Unbalanced sample size effect on the Genome-wide population differentiation studies," in *Proc. IEEE Int. Conf. Bioinf. Biomed. Workshops*, 2010, pp. 347–352.

[38] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, 1995, pp. 1137–1143.

[40] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 35, pp. 1–17, 2017.

[41] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020, Art. no. 6.

[42] D. Chicco and M. Masseroli, "A discrete optimization approach for SVD best truncation choice based on ROC curves," in *Proc. 13th IEEE Int. Conf. Bioinf. Bioeng.*, 2013, pp. 1–4.

[43] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic Regression*. Berlin, Germany: Springer, 2002.

[44] L. E. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination for the Lasso and sparse supervised learning problems," pp. 1–31, 2010, *arXiv:1009.4219*.

[45] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[46] R. E. Schapire, "Explaining AdaBoost," in *Empirical Inference*. Berlin, Germany: Springer, 2013, pp. 37–52.

[47] D. Chicco and M. Masseroli, "Validation pipeline for computational prediction of Genomics annotations," in *Proc. 12th Int. Meeting Comput. Intell. Methods Bioinf. Biostatist.*, 2016, pp. 233–244.

[48] National Center for Biotechnology Information, U.S. National Library of Medicine, "PubMed," Jun. 6, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/

[49] Google Inc. "Google Scholar," Jun. 6, 2019. [Online]. Available: https://scholar.google.com

[50] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting Genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 43, pp. 15545–15550, 2005.

[51] D. Chicco, H. S. Bi, J. Reimand, and M. M. Hoffman, "BEHST: Genomic set enrichment analysis enhanced through integration of chromatin long-range interactions," pp. 1–29, 2020, *arXiv:168427*.

[52] J. Reimand *et al.*, "g:Profiler–A web server for functional interpretation of gene lists (2016 update)," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W83–W89, 2016.

[53] E. S. Balakirev and F. J. Ayala, "Pseudogenes: Are they "junk" or functional DNA?," *Annu. Rev. Genet.*, vol. 37, no. 1, pp. 123–151, 2003.

[54] S. E. Racedo *et al.*, "Mouse and human CRKL is dosage sensitive for cardiac outflow tract formation," *Amer. J. Hum. Genet.*, vol. 96, no. 2, pp. 235–244, 2015.

[55] A. Sutani *et al.*, "WDR11 is another causative gene for coloboma, cardiac anomaly and growth retardation in 10q26 deletion syndrome," *Eur. J. Med. Genet.*, vol. 63, no. 1, 2020, Art. no. 103626.

[56] S. Yu, L. Shao, H. Kilbride, and D. L. Zwick, "Haploinsufficiencies of FOXF1 and FOXC2 genes associated with lethal alveolar capillary dysplasia and congenital heart disease," *Amer. J. Med. Genet. Part A*, vol. 152, no. 5, pp. 1257–1262, 2010.

[57] P. van der Harst and N. Verweij, "Identification of 64 novel genetic Loci provides an expanded view on the genetic architecture of coronary artery disease," *Circ. Res.*, vol. 122, no. 3, pp. 433–443, 2018.

[58] Gene Cards, "METTL7B," Apr. 13, 2020. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=METTL7B&keywords=METTL7B

[59] M. Moysés-Oliveira *et al.*, "Inactivation of AMMECR1 is associated with growth, bone, and heart alterations," *Hum. Mutation*, vol. 39, no. 2, pp. 281–291, 2018.

[60] Gene Cards, "SLFN5," Apr. 13, 2020. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SLFN5&keywords=SLFN5

[61] J. G. Blanco *et al.*, "Genetic polymorphisms in the carbonyl reductase 3 gene CBR3 and the NAD (P) H: Quinone oxidoreductase 1 gene NQO1 in patients who developed anthracycline-related congestive heart failure after childhood cancer," *Cancer: Interdisciplinary Int. J. Amer. Cancer Soc.*, vol. 112, no. 12, pp. 2789–2795, 2008.

[62] P. D Tatman *et al.*, "Pediatric dilated cardiomyopathy hearts display a unique gene expression profile," *J. Clin. Invest. Insight*, vol. 2, no. 14, 2017, Art. no. e94249.

[63] Gene Cards, "DEDD2," Apr. 13, 2020. [Online]. Available: https://www.genecards.org/cgi-bin/carddisp.pl?gene=DEDD2&keywords=DEDD2

[64] Z. Yin *et al.*, "Heart-specific ablation of PRKAR1A causes failure of heart development and myxomagenesis," *Circulation*, vol. 117, no. 11, pp. 1414–1422, 2008.

[65] QuickGO, "GO:0004334 – Fumarylacetoacetase activity," Oct. 1, 2019. [Online]. Available: https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0004334

[66] S. Mohamed *et al.*, "Tyrosinemia type 1: A rare and forgotten cause of reversible hypertrophic cardiomyopathy in infancy," *BMC Res. Notes*, vol. 6, no. 1, 2013, Art. no. 362.

[67] D. Chicco and M. Masseroli, "Software suite for gene and protein annotation prediction and similarity search," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 4, pp. 837–843, Jul./Aug. 2015.

[68] L. Rokach and O. Z. Maimon, *Data Mining With Decision Trees: Theory and Applications*, vol. 69. Singapore: World Scientific, 2008.

[69] I. Orlandi, L. Oneto, and D. Anguita, "Random forests model selection," in *Proc. 24th Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn.*, 2016, pp. 1–6.