Random forest can predict survival of cardiovascular heart disease patients from serum creatinine and ejection fraction alone

Davide Chicco
Peter Munk Cardiac Centre
davidechicco@davidechicco.it

1st March, 2019

Abstract

Abstract

1 Introduction

Ten quick tips paper [1]

Mesothelioma paper [2]

2 Dataset

We analyzed a dataset containing the electronic health records (EHR) of 299 patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015 [3, 4]. The patients were 105 women and 194 men, and their ages range between 40 and 95 years old (Table 1). All the 299 patients have left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) Classification of the stages of the heart failures [5].

The dataset contains 12 features, which report clinical, body, and lifestyle information (Table 1), that we briefly describe here. Some features are binary (age, anaemia, high blood pressure, diabetes, sex, and smoking Table 1), and need no additional explanation.

The hospital physician consider a patient having anaemia if haematocrit level was lower than 36% [3, 6].

The creatinine phosphokinase (CPK) states the level of the CPK enzyme in blood. When a muscle tissue gets damaged, CPK flows into the blood. Therefore, high levels of CPK in the blood of a patient might indicate a heart failure or injury [7]. The ejection fraction states percentage, of how much blood the left ventricle pumps out with each contraction. The serum creatinine is a waste product geneated by creatine, when a muscle breaks down. Especially, doctors focus on serum creatinine in blood to check kidney function. If a patient has high levels of serum creatinine, it might indicate a renal dysfunction going on [8]. Sodium is a mineral that serves for the correct functioning of muscles and nerves. The serum sodium test is a routine blood exam that tells if a patient has a normal level of sodium in the blood. An abnormally low level of sodium in the blood might be caused by a heart failure [9]. The death event feature, that we use as the target in our binary classification study, states if the patient died or survived at the end of the follow-up period, that was 130 days on average [3].

Regarding the dataset imbalance, the survived patients (death event = 0) are 203, while the dead patients (death event = 1) are 96. In statistical terms, there are 32.11% positives and 67.89% negatives.

As done by the original data curators [3], we represented this dataset as a table having 299 rows (patients) and 12 columns (features). For clarification purposes, we slightly changed the names of some features of the original dataset (Supplementary Information).

3 Methods

3.1 Survival prediction classifiers

feature	explanation	measurement	range
age	age of the patient	years	[40, ..., 95]
anaemia	decrease of red blood cells or hemoglobin	boolean	0, 1
high blood pressure	if a patient has hypertension	boolean	0, 1
creatinine phosphokinase (CPK)	level of the CPK enzyme in the blood	mcg/L	[23, ..., 7861]
diabetes	if the patient has diabetes	boolean	0, 1
ejection fraction	percentage of blood leaving the heart at each contraction	percentage	[14, ..., 80]
sex	woman or man	binary	0, 1
platelets	platelets in the blood	kiloplatelets/mL	[25.01, ..., 850.00]
serum creatinine	level of creatinine in the blood	mg/dL	[0.50, ..., 9.40]
serum sodium	level of sodium in the blood	mEq/L	[114, ..., 148]
smoking	if the patient smokes	boolean	0, 1
(target) death event	if the patient died during the follow-up period	boolean	0, 1

Table 1: ***Meanings, measurement units, and intervals of each feature of the dataset.*** *mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per litre.*

Random forest, one rule, linear regression, naïve Bayes, decision tree, neural network, support vector machine, and k-nearest neighbors

3.2 Random forest feature selection

Random forest feature selection

4 Results

4.1 Survival machine learning prediction

We employed several methods to predict the survival of the patients. We applied each method 10 times and reported the median result score (Table 2).

For methods that needed hyper-parameter optimization (neural network, support vector machine, and k-nearest neighbors), we split the dataset into 60% (179 randomly selected patients) for the training set, 20% (60 randomly selected patients) for the validation set, and 20% (the remaining 60 patients) for the test set. To choose the top hyper-parameters, we used a grid search and selected the models that generated the highest Matthews correlation coefficient.

For the other methods (random forest, one rule, linear regression, naïve Bayes, and decision tree), instead, we split the dataset into 80% (239 randomly selected patients) for the training set, and 20% (the remaining 60 patients) for the test set.

The prediction results show that the random forest classifier applied to the dataset containing only the most important features (serum creatinine and ejection fraction) outperformed all the other methods (Table 2). Random forest on the two top features, in fact, obtained the highest MCC (+0.42), $\mathrm{F}_1 score$ (0.59), true positive rate (0.86), precision-recall AUC (0.74), and receiver oprating characteristic AUC (0.80), and the second highest accuracy (0.75) among all the algorithms. The true negative rate obtained by random forest applied to the two top features obtained the lowest score (tied with the neural network) among all the methods employed, even if still high (0.86). Random forest on the two features is the only method which obtained true positive rate above the average (0.54), and a precision-recall AUC above two-thirds (0.74). Also, random forest on the two top features and random forest on all the features are the only methods which achieved a ROC AUC equal to 80%.

It is worth noticing that random forest on the two top features and random forest on all the features obtained almost the same true negative rate (0.86 and 0.87, respectively), but different true positive rate: this ensemble learning method on the two features, in fact, obtained sensitivity = 0.54, while the method on all the eleven features achieved sensitivity = 0.46 instead. Random forest on the two features outpeformed random forest on all the features also on the precision-recall AUC (0.74 and 0.62, respectively).

These results show that the removal of the non-top nine features helped random forest to improve its score on the prediction of the positive elements, while did not significantly changed its performance on the negative

method	**MCC**	F_1 score	accuracy	TP rate	TN rate	PR AUC	ROC AUC
Random forest (serum creatinine & ejection fraction)	**+0.42**	0.59	0.75	0.54	0.86	0.74	0.80
One rule	**+0.40**	0.39	0.77	0.25	0.98	0.50	0.63
Linear regression	**+0.37**	0.54	0.73	0.41	0.92	0.57	0.67
Naïve Bayes	**+0.35**	0.53	0.73	0.44	0.87	0.51	0.66
Random forest (all features)	**+0.35**	0.50	0.75	0.46	0.87	0.62	0.80
Decision tree	**+0.34**	0.55	0.72	0.50	0.87	0.49	0.66
Neural network	**+0.29**	0.48	0.70	0.41	0.86	0.47	0.76
Support vector machine	**+0.28**	0.34	0.74	0.22	0.96	0.59	0.78
k-nearest neighbors	**+0.05**	0.16	0.66	0.11	0.89	0.32	0.52

Table 2: ***Survival prediction results.*** *Random forest (serum creatinine & ejection fraction): random forest classifier applied to the dataset made of all the 299 patients but only of these 2 features. Random forest (all features): random forest classifier applied to the complete dataset (299 patients and 11 features). MCC: Matthewxs correlation coefficient. TP rate: true positive rate (sensitivity, recall). TN rate: true negative rate (specificify). Confusion matrix threshold for MCC, F_1 score, accuracy, TP rate, TN rate: $\tau = 0.5$. PR AUC: precision-recall area under the curve. ROC AUC: receiver operating characteristic 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. MCC, F_1 score, accuracy, TP rate, TN rate, PR AUC, ROC AUC formulas: Supplementary Information (subsection 5.2). Our hyper-parameter grid search optimization for k-nearest neighbors selected $k = 5$ on 3 runs out of 10. Our hyper-parameter grid search optimization for the support vector machine selected $C = 0.1$ on 6 runs out of 10. Our hyper-parameter grid search optimization for the neural network selected 1 hidden layer and 300 hidden unites on 5 runs out of 10.*

elements.

Because of the imbalance of the dataset (67.89% negative elements and 32.11% positive elements), all the methods obtained better prediction scores on the true negative rate, rather than on the true positive rate (Table 2). This results happens because the algorithms can see more negative elements during training, and therefore they are more trained to recognize them during testing. Especially, one rule and support vector machines achieved almost perfect prediction scores on the negative elements (specificity = 0.98 and 0.96, respectively), which outperformed random forest on the two top features, too.

4.2 Biostatistics descriptive analysis

	full sample		dead patients		survived patients	
category feature	#	%	#	%	#	%
anaemia (0: false)	170	56.86	50	52.08	120	59.11
anaemia (1: true)	129	43.14	46	47.92	83	40.89
high blood pressure (0: false)	194	64.88	57	59.38	137	67.49
high blood pressure (1: true)	105	35.12	39	40.62	66	32.51
diabetes (0: false)	174	58.19	56	58.33	118	58.13
diabetes (1: true)	125	41.81	40	41.67	85	41.87
sex (0: woman)	105	35.12	34	35.42	71	34.98
sex (1: man)	194	64.88	62	64.58	132	65.02
smoking (0: false)	203	67.89	66	68.75	137	67.49
smoking (1: true)	96	32.11	30	31.25	66	32.51

Table 3: ***Statistical quantitative description of the category features.*** *#: number of patients. %: percentage of patients. Full sample: 299 individuals. Dead patients: 96 individuals. Survived patients: 203 individuals.*

	full sample		dead patients		survived patients	
numeric feature	median	mean	median	mean	median	mean
age	60.00	60.83	65.00	65.22	60.00	58.76
creatinine phosphokinase	250.00	581.80	259.00	670.20	245.00	540.10
ejection fraction	38.00	38.08	30.00	33.47	38.00	40.27
platelets	262.00	263.36	258.50	256.38	263.00	266.66
serum creatinine	1.10	1.39	1.30	1.84	1.00	1.19
serum sodium	137.00	136.60	135.50	135.40	137.00	137.20

Table 4: ***Statistical quantitative description of the numeric features.*** *Full sample: 299 individuals. Dead patients: 96 individuals. Survived patients: 203 individuals.*

4.3 Biostatistics feature selection results

feature	abs(t)	abs(t) rank	p-value	p-value rank	PCC	PCC rank	95% conf. int. min, max
age	87.90	2	1.28×10^{-215}	2	0.25	2	59.16, 61.87
anaemia	2.80	8	5.29×10^{-003}	8	0.07	4	0.03, 0.19
creatinine phosphokinase	10.36	6	1.08×10^{-021}	6	0.06	5	471.09, 691.95
diabetes	2.47	9	1.40×10^{-002}	9	0.00	7	0.02, 0.17
ejection fraction	55.13	3	1.12×10^{-158}	3	–0.27	11	36.41, 39.11
high blood pressure	0.78	10	4.37×10^{-001}	10	0.08	3	–0.05, 0.11
platelets	46.56	4	8.81×10^{-139}	4	–0.05	9	252.23, 274.49
serum creatinine	16.34	5	1.08×10^{-046}	5	0.29	1	0.94, 1.20
serum sodium	531.17	1	0.00	1	–0.20	10	135.80, 136.81
sex	8.47	7	1.87×10^{-016}	7	0.00	6	0.25, 0.4
smoking	0.00	11	$1.00 \times 10^{+000}$	11	–0.01	8	–0.08, 0.08

Table 5: ***Biostatistics feature selection.*** *Rates computed between each feature and the target feature death event. abs(t): absolute value of the t outcome of Student's t-test [10]. abs(t) range:* $[0, 1]$*, where 0 means minimum correlation, and 1 means maximum correlation. PCC: Pearson correlation coefficient between the specified feature (row) and the survival target. PCC range:* $[-1, +1]$*, where +1 means complete linear correlation, 0 is no linear correlation, and* -1 *means opposite linear correlation. abs(t) rank: ranking of the features based upon the abs(t) value. p-value rank: ranking of the features based upon the p-value value. PCC rank: ranking of the features based upon the PCC value.*

rank	feature	Shapiro–Wilk test
1	creatinine phosphokinase	7.05×10^{-28}
2	serum creatinine	5.39×10^{-27}
3	smoking	4.58×10^{-26}
4	death event	4.58×10^{-26}
5	sex	1.17×10^{-25}
6	high blood pressure	1.17×10^{-25}
7	diabetes	5.12×10^{-25}
8	anaemia	6.21×10^{-25}
9	platelets	2.89×10^{-12}
10	serum sodium	9.21×10^{-10}
11	ejection fraction	7.22×10^{-09}
12	age	5.34×10^{-05}

Table 6: ***Shapiro Wilk test****. Shapiro Wilk test results.*

rank	feature	Wilcoxon signed-rank test	Kruskal-Wallis test
1	serum creatinine	0	0
2	ejection fraction	0.000001	0.000001
3	age	0.000167	0.000166
4	serum sodium	0.000293	0.000292
5	high blood pressure	0.171016	0.170746
6	anaemia	0.252970	0.252624
7	platelets	0.425559	0.425142
8	creatinine phosphokinase	0.684040	0.683513
9	smoking	0.828190	0.827500
10	sex	0.941292	0.940603
11	diabetes	0.973913	0.973244

Table 7: ***Wilcoxon signed-rank test (Mann-Whitney U test) and Kruskal-Wallis test****. Results.*

rank	feature	chi squared test
1	ejection fraction	0.000500
2	serum creatinine	0.000500
3	serum sodium	0.003998
4	age	0.005997
5	high blood pressure	0.181909
6	anaemia	0.260370
7	creatinine phosphokinase	0.377811
8	platelets	0.637681
9	smoking	0.889555
10	sex	1
11	diabetes	1

Table 8: ***Chi squared test****. Results.*

4.4 Machine learning feature selection results

Similarly to what [2] authors did for a dataset of patients having mesothelioma symptoms, we decided then to investigate the most important features of the cardiovascular heart disease patients dataset. To this aim, we again employed random forest [11], to take advantage of its twofold feature selection outcome: statistical outcome (accuracy decrease), and content-informative outcome (Gini impurity [12]).

Both the accuracy reduction and the Gini impurity rankings detected serum creatinine, age fraction, and age as the top three most important features of the dataset (Figure 1). The two rankings show high similarity: the Kendall τ rank correlation coefficient between them is $+0.56$ and the Spearman ρ rank correlation coefficient is $+0.73$. Both these coefficients range between -1 (when the ranking of a list is the opposite of the other one) and $+1$ (when the two rankings are similar) [13].

To have a unique final classification to evaluate, we then merged the two rankings into an aggregate ranking by using Borda's method [14]. For every feature f, we added its position in the accuracy decrease ranking $p_1(f)$ to its position in the Gini impurity raking $p_2(f)$, and saved this aggregate value in the ranking variable $score_f$. Finally, we sorted all the features increasingly based upon $score_f$ (Table 9).

In the aggregated ranking (Table 9), creatinine phosphokinase resulted the fourth most important features tied, while anaemia and diabetes resulted the least important features among all.

Once we obtained the ranking of the features based upon their importance, we wanted to understand what is the minimum number of features (and which features should be used) to still be able to perform an accurate prediction of the survival of patients. In fact, we want to provide a method that can be used by medical doctors in the hospital, in the scenario where just few features of the electronic health record (EHR) of a patient are available. We first wondered if the any singular feature was sufficient to make a survival prediction, that is if any top feature had a strong correlation with the target feature (survival). We calculated the Pearson correlation coefficient [15] between each feature and the survival, and found no feature having value close to $+1$ (PCC column in Table 9).

After having checked the impossibility to predict the patients survival by using one singular feature, we

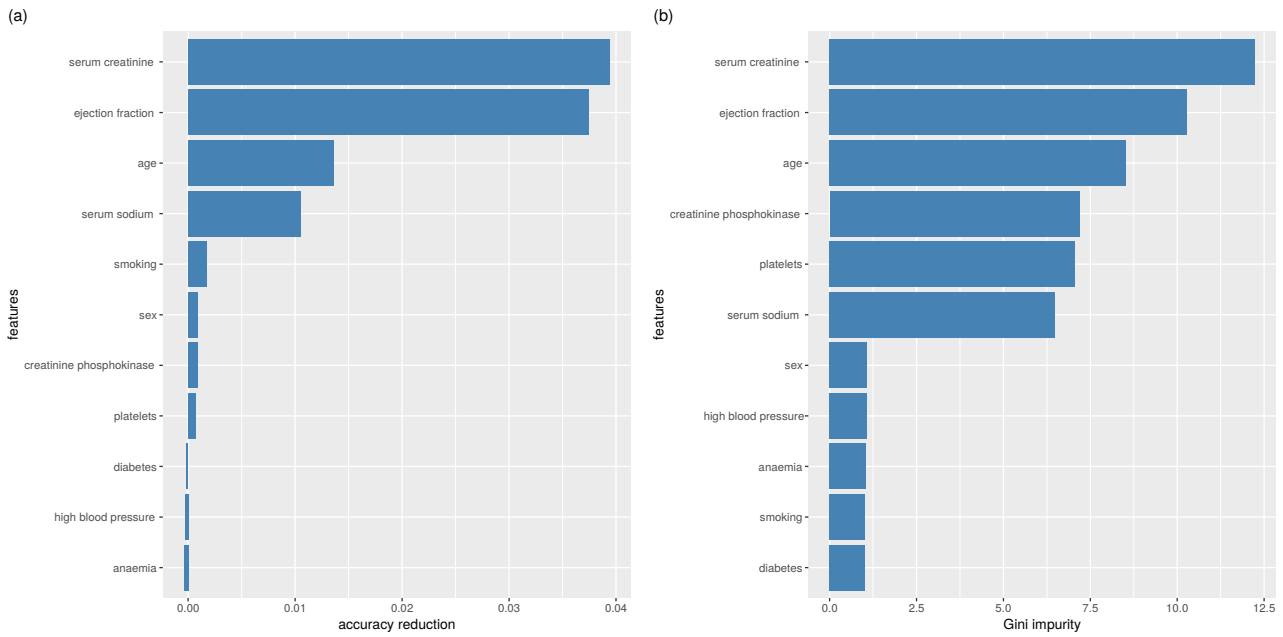

Figure 1: ***Random forest feature selection.*** *Random forest feature selection through accuracy reduction (a). Random forest feature selection through Gini impurity (b).*

final rank	feature	accuracy decrease	accuracy decrease rank	Gini impurity	Gini impurity rank
1	serum creatinine	3.78×10^{-2}	1	11.84	1
2	ejection fraction	3.43×10^{-2}	2	10.71	2
3	age	1.53×10^{-2}	3	8.58	3
4	creatinine phosphokinase	7.27×10^{-4}	6	7.26	4
4	serum sodium	7.20×10^{-3}	4	6.49	6
6	sex	1.64×10^{-3}	5	1.12	8
6	platelets	2.47×10^{-4}	8	6.80	5
8	high blood pressure	-1.68×10^{-3}	11	1.13	7
8	smoking	3.68×10^{-4}	7	0.95	11
10	anaemia	-5.91×10^{-4}	10	1.06	9
10	diabetes	-1.41×10^{-4}	9	1.02	10

Table 9: ***Random forest feature selection aggregate ranking.*** *We merged the two ranking through their position.*

decided to use the two most important features for this scope: serum creatinine, and ejection fraction. As we described earlier (Survival prediction results), a random forest classifier applied to the dataset containing only these two features and all the 299 patients outperformed all the other methods, including random forest itself applied to all the features (Table 2).

To verify further the predictive power of serum creatinine and ejection fraction, we depicted a scatterplot with the serum creatinine values on the x axis and the ejection fraction values on the y axis, and we colored every patient-point based on its status (alive or dead, Figure 2). This plot shows a clear distinction between alive patients and dead patients, that we highlighed by manually inserting a black straight line.

5 Discussion and conclusions

– Random forest applied to serum creatinine and ejection fraction alone can predict survival better than the other methods applied to all the features.

– The original dataset curators identified these top features: age, serum creatinine (renal dysfunction), high blood pressure, ejection fraction and anemia. In our feature selection instead, high blood pressure is 8th out of

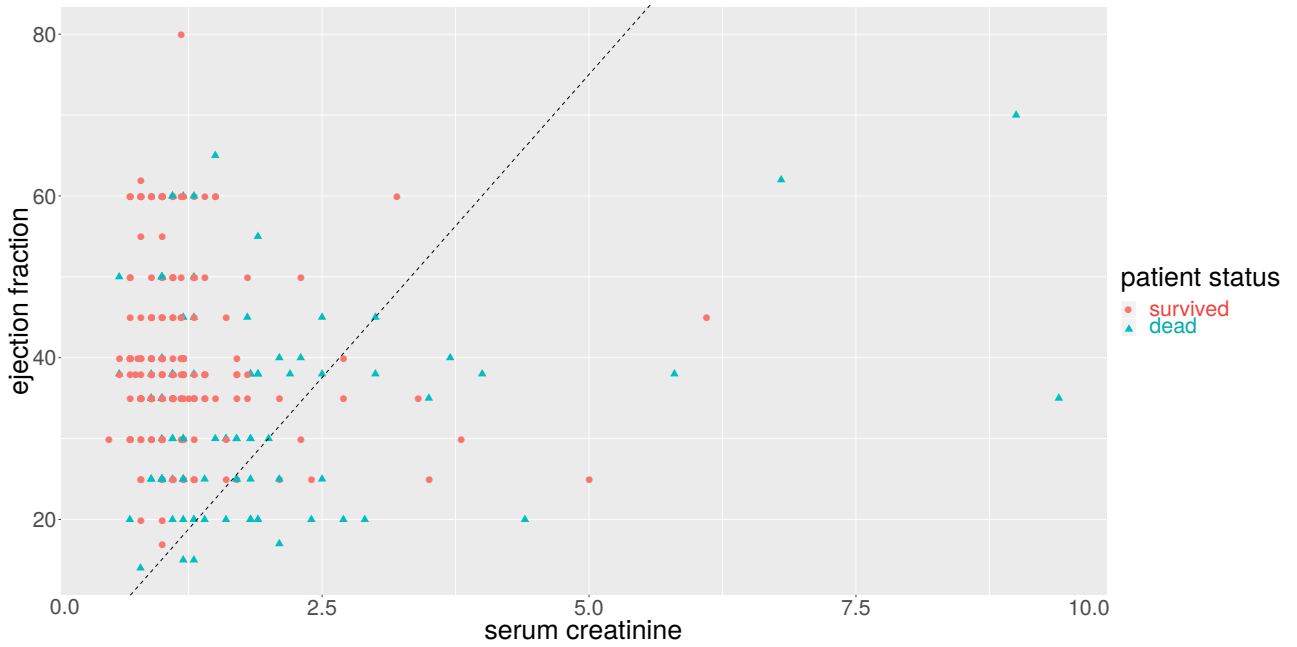

Figure 2: ***Scatterplot of serum creatinine versus ejection fraction*** *Serum creatinine (x axis) range: [0.50, 9.40] mg/dL. Ejection fraction (y axis) range: [14, 80]%. We manually drew a black straight line to highlight the discrimination between alive and dead patients.*

11, and anemia is the 10th out of 11 (last position tied with diabetes).

Acknowledgments

The author thanks Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza for having collected the dataset and have put it publically available online.

Data and software availability

The dataset file used in this project [4] is publically available on FigShare at: `https://plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1`

Our software code is publically available on GitHub at:
`https://github.com/davidechicco/cardiovascular_heart_disease`

Authors' details

Davide Chicco (ORCID: 0000-0001-9655-7142) is in the Peter Munk Cardiac Centre, Toronto, Ontario, Canada.

Correspondence should be addressed to Davide Chicco: davidechicco@davidechicco.it

References

[1] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(35):1–17, 2017.

[2] Davide Chicco and Cristina Rovelli. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PLOS One*, 14(1):e0208737, 2019.

[3] Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza. Survival analysis of heart failure patients: a case study. *PLOS One*, 12(7):e0181001, 2017.

[4] Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza. Survival analysis of heart failure patients: a case study. Dataset. `https://plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1`. URL visited on 25th January 2019.

[5] Charlene Bredy, Margherita Ministeri, Alexander Kempny, Rafael Alonso-Gonzalez, Lorna Swan, Anselm Uebing, Gerhard-Paul Diller, Michael A Gatzoulis, and Konstantinos Dimopoulos. New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome. *European Heart Journal-Quality of Care and Clinical Outcomes*, 4(1):51–58, 2017.

[6] William M McClellan, W Dana Flanders, Robert D Langston, Claudine Jurkovitz, and Rodney Presley. Anemia and renal insufficiency are independent risk factors for death among patients with congestive heart failure admitted to community hospitals: a population-based study. *Journal of the American Society of Nephrology*, 13(7):1928–1936, 2002.

[7] Johns Hopkins Rheumatology. Creatine Phosphokinase (CPK). `https://www.hopkinslupus.org/lupus-tests/clinical-tests/creatine-phosphokinase-cpk/`. URL visited on 25th January 2019.

[8] Carissa Stephens. What is a creatinine blood test? `https://www.healthline.com/health/creatinine-blood`. URL visited on 25th January 2019.

[9] Christine Case-Lo. What is a sodium blood test? `https://www.healthline.com/health/sodium-blood`. URL visited on 25th January 2019.

[10] DB Owen. The power of Student's t-test. *Journal of the American Statistical Association*, 60(309):320–333, 1965.

[11] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[12] Leo Breiman and Adele Cutler. Random forests - Gini importance. `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#giniimp`. URL visited on 31st October 2016, 2004.

[13] Davide Chicco, Eleonora Ciceri, and Marco Masseroli. Extended Spearman and Kendall coefficients for gene annotation list correlation. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 19–32. Springer, 2014.

[14] D Sculley. Rank aggregation for similar items. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 587–592. SIAM, 2007.

[15] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, pages 1–4. Springer, 2009.

Supplementary Information

5.1 Original feature renaming

For clarification and precision purposes, we slightly changed some names of the features from the original dataset [3, 4]:
We renamed "Gender" to "sex".
We renamed "BP" to "high blood pressure".
We renamed "Pletelets" to "platelets".
We renamed "Sodium" to "serum sodium".
We renamed "CPK" to "creatinine phosphokinase".
We renamed "Creatinine" to "serum creatinine".
We renamed "Event" to "death event", which is the survival target.
We discarded the "TIME" feature.

5.2 Binary statistical rates

List of statistical rates and their formulas:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{1}$$
(worst value $= -1$; best value $= +1$)

$$F_1\ score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2}$$
(worst value $= 0$; best value $= 1$)

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$
(worst value $= 0$; best value $= 1$)

$$true\ positive\ rate\ =\ recall\ =\ sensitivity\ = \frac{TP}{TP+FN} \tag{4}$$
(worst value $= 0$; best value $= 1$)

$$true\ negative\ rate\ = specificity\ =\ \frac{TN}{TN+FP} \tag{5}$$
(worst value $= 0$; best value $= 1$)

$$precision = \frac{TP}{TP+FP} \tag{6}$$
(worst value $= 0$; best value $= 1$)

$$false\ positive\ rate\ = fallout\ =\ \frac{FP}{FP+TN} \tag{7}$$
(worst value $= 1$; best value $= 0$)

$$Precision\text{-}recall\ (PR)\ curve = \begin{cases} true\ positive\ rate & on\ the\ x\ axis \\ precision & on\ the\ y\ axis \end{cases} \tag{8}$$
(worst value $= 0$; best value $= 1$)

$$ROC\ curve = \begin{cases} false\ positive\ rate & on\ the\ x\ axis \\ true\ positive\ rate & on\ the\ y\ axis \end{cases} \tag{9}$$
(worst value $= 0$; best value $= 1$)