Interpretable Knowledge Mining for Heart Failure Prognosis Risk Evaluation

Shaobo Wang[1,2⋆], Guangliang Liu[2], Wenyan Zhu[2], Zengtao Jiao[2], Haichen Lv[3], Jun Yan[2], and Yunlong Xia[3]

[1]Beijing University of Technology, Beijing, China
[2]Yidu Cloud (Beijing) Technology Co Ltd., Beijing, China
{shaobo.wang,guangliang.liu,wenyan.zhu,zengtao.jiao,jun.yan}@yiducloud.cn
[3] Department of Cardiology, The First Affiliated Hospital of Dalian Medical University, Dalian, China
lvhaichen@dmu.edu.cn,yunlongxia01@163.com

Abstract. In this article, we propose a pipeline to mine interpretable knowledge from electronic health records (EHR) for the Heart Failure (HF) prognosis risk evaluation task. Mortality risk after first-diagnosis HF highly impacts patients' life quality, and is helpful for physicians to efficiently monitor patients' disease progress. How to mine medically reasonable and interpretable knowledge to assist physicians in evaluating mortality risk is a non-trivial task. The proposed pipeline leverages a gradient-boosting-based predictive model to estimate the risk of HF prognosis, and discovers variables and decision rules from the predictive model. The mined knowledge is confirmed as interpretable and inspirable by physicians.

Keywords: Heart Failure · Knowledge Mining · Interpretability.

1 Introduction

HF is a clinical syndrome characterized by blood congestion in pulmonary circulation or systemic circulation, and/or by insufficient blood perfusion in organs and tissues. HF is the late stage of various heart diseases and it can lead to serious manifestation. With high mortality and readmission rate, the prognosis of HF is often not satisfactory, resulting in a certain medical burden.

The incidence rate of HF, in modern society, has been increasing due to aging of population, change of disease spectrum and improved survival rate of various cardiovascular diseases [27]. The prevalence of HF in developed countries is 1.5%-2.0%, and, among people over 70 years old, the prevalence rate is higher than 10% [26]. To take China as an example, the average life expectancy grows, more people, especially the elderly, are suffering from chronic diseases. This causes many complications in HF patients. Technically, poor prognosis conditions of HF cannot be avoided. The 1-year all-cause mortality rate and 1-year

[⋆] The author is an intern at Yidu Cloud.

readmission rate were 7.2% and 31.9% in patients with chronic stable HF, and 17.4% and 43.9% in patients with acute HF, respectively [24] .

Some biomarkers have been individually used to predict outcomes of HF, for example BNP, age, cystatin C, serum uric acid, D-Dimer, etc. [10, 28]. Traditional biomarkers closely related to cardiovascular mortality in the general population, such as body mass index (BMI), serum cholesterol, and blood pressure (BP), are found useful to predict outcomes of patients with Chronic Heart Failure (CHF) [5]. Because of the complexity of HF prognosis, analysis towards multi-biomarkers might be worthy. Multi-biomarker strategies are gaining interest in tasks like clinical assessment and risk stratification of HF patients. Previous studies have shown that multi-biomarkers contribute to higher prognostic accuracy than an individual one [8].

How multi-biomarkers can affect the mortality rate of HF patients deserves further investigation. Predictive models (PM) based on machine learning have advantages in portraying interactions between biomarkers. Another significant issue of medical applications is interpretability. Therefore we employ an interpretable predictive model(IPM) to mine medical knowledge. When it comes to HF, widely recognized prognostic guidelines are not available, and current research of medical knowledge mining(MKM) in this field is not sufficient either. The motivation behind MKM for HF prognosis is to discover knowledge that can be a good supplement to medical guidelines. To the best of our knowledge, this is the first work regarding knowledge mining for 1-year in-hospital HF mortality risk evaluation. Our contributions are:

- We apply an interpretable model to understand the decisions made by predictive models for the task of 1-year in-hospital mortality risk evaluation of HF patients. The extracted knowledge is human-understandable.
- We set up a pipeline, incorporating medical expertise, to verify extracted knowledge, thereby the extracted knowledge is medically meaningful.
- We design a knowledge filtering method to extract knowledge that is applicable in both angles of medical logic and statistical analysis.

2 Previous works

2.1 HF Risk Prediction

In recent years, predictive modelling, a powerful risk prediction tool, has been gaining increasing interests in the study of cardiovascular diseases. Early and effective intervention according to risk evaluation is of great significance for HF patients [6, 31]. There are many studies on predicting the outcome of HF based on statistical or machine learning methods. [20, 21] proposed the Seattle HF Model (SHFM) to predict the mortality rate of HF patients. [19] made a prediction model to predict HF mortality called Enhanced Feedback for Effective Cardiac Treatment (EFFECT), and [10] used a Classification And Regression

Tree (CART) model to predict the in-hospital mortality of acutely decompensated HF and made a risk stratification. Logistic Regression (LR) is often used in the research of HF prognosis [12], yet it fails to model the non-linear relations among features. [13] concluded that it was more reasonable to construct a non-linear model than LR. In our study, we consider a method of applying gradient-boosting-based algorithms to establish the one-year mortality prediction model of HF, and we choose Rulefit [11] in the subsequent medical knowledge mining task.

2.2 Medical Knowledge Mining

MKM aims to extract meaningful patterns from medical datasets, and these patterns are expected to support physicians and patients in the process of screening, diagnosis, treatment, prognosis, health monitoring and management. A popular data source of MKM is EHR which records a patient's routine in a hospital, for example demographic data, diagnosis, laboratory test results, nursing records and prescriptions. Compared to the general applications of knowledge mining, there are some specific difficulties in the study of MKM: data availability and data standardization [18].

Cancer, heart diseases and diabetes are the top 3 most common diseases that are considered in previous works, most of which focus on the diagnosis and prognosis stage [9]. [2] compared the performance of various machine learning models for the task of heart disease prediction. [3] applied social network analysis, text mining, temporal analysis and higher order feature construction to healthcare data analysis. [7] used 13 attributes, such as gender and blood pressure, to estimate the likelihood of a patient being diagnosed with heart disease. [30] evaluated the performance of machine learning algorithms in four benchmark prediction tasks and suggested that recurrent neural networks achieved the most promising results in mortality prediction. According to [15], 64% of cardiology studies are devoted to classification techniques, and predictive modeling is the second most popular technique.

Previous research mainly focus on diagnosis-related tasks or onset risk evaluation tasks. Refined-Clinical Knowledge Model (R-CKM) which is a tree-based PM could produce medical knowledge from EHR [14]. [4] mined some knowledge through the R-CKM and made it possible to enrich and optimize the medical guidelines of HF diagnosis. [22] developed a medical knowledge mining pipeline based on temporal pattern mining for early detection of Congestive HF, and the mined patterns can make more accurate predictions than PM [22]. In contrast to previous studies, we interpret extracted patterns through incorporating physicians. The mined knowledge in our work not only conforms to medical common sense, but also gives supportive evidence to unverified hypotheses.

2.3 Interpretable Predictive Model

Interpretable machine learning(IML) has been a hot topic in current machine learning communities, especially due to the popularity of deep learning models. There is no clear mathematical definition of interpretability, though a natural language definition by Miller is *'Interpretability is the degree to which a human can understand the cause of a decision'* [25]. Some methods, such as neural network, have very high ability of feature abstraction and nonlinear fitting, yet the intermediate process is a black box. Medical experts would view these algorithms with suspicion.

IML can be categorized into three groups: interpretable models, model-agnostic methods and example-based explanation methods. Interpretable models include algorithms that are interpretable themselves, for instance linear regression, logistic regression and the decision tree. The RuleFit model employed in this work is one of interpretable models as well [11]. Model-agnostic methods enjoy high flexibility because they can be applied to any models, but they might influence models' performance adversely, like SHAP [38]. LIME [34], Anchor [35] and LORE [39] are representative model-agnostic methods that could generate rule-based explanations, but they can only achieve local interpretability. In the medical domain, data is represented in a structured format, thereby example-based explanation methods can not work [36].

Tree-based machine learning has been widely used in medical research as a reason of its self-interpretability, and the decision making process is humanly understandable. On the other hand, physicians are interested in figuring out the role of key variables at the population level. For instance, patients whose BNP is above 35 ng/L and NYHA is less than 40% will be diagnosed as HFrEF, while another patient whose NYHA level is larger than 49% might be diagnosed with HFpEF even though his/her NYHA level is around 90%. Given the aforementioned reasons, we employed RuleFit in this work. RuleFits consists of two components: a tree model and a linear model. The tree model implements classification or regression tasks, and associated decision rules are extracted from the learned model. Then the extracted decision rules, together with original features, would be fitted into the linear model.

3 Method

3.1 Data

In this study, we retrospectively collected EHR data of hospitalized patients diagnosed with HF between December 2010 and August 2018. Included patients are over 18 years old and were diagnosed with HF according to diagnosis guidelines. We used off-the-shelf natural language processing tools to structrize and standardrize collected raw data. Finally, we enrolled 13,602 patients with HF, and 537 (3.95%) died within 1 year.

3.2 Outcomes

Firstly, we build a predictive model to predict the mortality risk of HF patients. Any patients with a clear hospital death record within one year are labelled with high risk, and the others are considered low risk. Then, we interpret knowledge learned by the predictive model through calculating feature importance.

3.3 Feature engineering

Variables with filling rate greater than 80% were selected, and, finally, a total of 73 features were extracted, including demographics (age or sex), living habits (smoking or drinking), previous medical history (comorbidities or surgery), etiology, vital signs, routine laboratory examinations, interventions and admission medications.

Given the normal range of each laboratory test item from hospital, we discretize continuous features through labelling them into three tags(lower , normal and higher). For example, the normal range of white blood cell (WBC) is [3.5-9.5]10^9/L, and it will be transferred to 'lower' if WBC<3.5. This is because qualitative values are more meaningful than quantitative values in the process of making clinical assessment. Generally speaking, physicians focus far more on what items are abnormal.

Some continuous features will be labelled with only two categories, such as Basophils (BASO), the normal range of BASO is (0,0.06]10^9/L, and it falls to two tags (normal and higher). In order to achieve human-understandable interpretability from RuleFit, we adopt one-hot feature representations, examples of one-hot features are shown in Table 1. After normalization, missing values are fixed through calculating mean values (for continuous features like age) or mode numbers (for one-hot features).

Feature	One-hot Representation	Normal Range	Raw Data
WBC low	(1,0,0)	[3.5,9.5]	WBC<3.5
WBC normal	(0,1,0)	[3.5,9.5]	3.5≤WBC≤9.5
WBC high	(0,0,1)	[3.5,9.5]	WBC>9.5
BASO normal	(1,0)	(0,0.06]	BASO≤0.06
BASO high	(0,1)	(0,0.06]	BASO>0.06

Table 1. One-hot feature representation examples

3.4 Predictive Modelling

We compare the performance of several predictive models: Logistic Regression(LR), Support Vector Machine(SVM), Gradient Boosting Decision Tree(GBDT) and RuleFit. LR is generally considered as a baseline model for classification tasks, and SVM shows good performance in binary classification tasks. RuleFit applies a GBDT in the decision rules generation, so a pure GBDT model is trained to make comparisons.

The prediction outcome of RuleFit is defined as

$$F(\mathbf{x}) = 1/\left(1 + e^{-g(\mathbf{x})}\right)$$

$$g(\mathbf{x}) = \hat{a_0} + \sum_{k=1}^{K} \hat{a_k} r_k(\mathbf{x}) + \sum_{j=1}^{n} \hat{b_j} l_j(x_j)$$

$r_k(\mathbf{x})$ is the feature representation of k_{th} rule, and x_j is the original feature. Considering normalization of input feature x_j, it is transformed into $l_j(x_j)$: $min(\delta_j^+, max(\delta_j^-, x_j))$. δ_j^+ and δ_j^- are respectively the upper and lower quantiles of feature x_j, then linear parameters can be available with

$$\hat{a_k}, \hat{b_j} = \arg\min_{a_k, b_j} \sum_{i=1}^{N} Loss(y_i, F(\mathbf{x})) + \lambda(\sum_{k=1}^{K} |a_k| + \sum_{j=1}^{n} |\hat{b_j}|)$$

Procedures for choosing the regularization coefficient λ are discussed in [37]. In this work, a squared-error ramp loss is used to achieve better robustness against out-of-distribution cases. The loss function is defined as

$$Loss(\mathrm{y}_i, \mathrm{F}(\mathbf{x}_i)) = [\mathbf{y}_i - max(-1, min(1, \mathbf{F}(\mathbf{x}_i)))]^2$$

3.5 Knowledge Mining

Decision Rules Extraction We decompose the trained GBDT into decision rules: any path through the root nodes in trees can be converted into decision rules. The representation of the rule is as: *IF* $x_1 > 60$ *and* $x_2 = 1$ *and* $x_3 = 0$ *THEN 1 ELSE 0.* Rules r_m are defined

$$r_m(\boldsymbol{x}) = \prod_{j \in T_m} I(x_j \in s_{jm})$$

Where T_m is the set of features used in the m-th tree, $I(.)$ is the indicator function that is 1 if feature x_j is in the subset of the value s_{jm} and 0 otherwise. An example of the rule is like

$$r_{256}(x) = I(is\ digoxin)I(BIL\ is\ high)I(HGB\ is\ not\ low)$$

$r_{256}(x)$ is 1 if and only if all the three conditions above are met. The total number of rules derived from the GBDT model K is: $\sum_{m=1}^{M} 2(t_m - 1)$ where t_m is the number of terminal nodes within the m_{th} tree.

Feature Importance RuleFit calculates the importance of the rule feature as: k-th rule-feature $I_k = |\hat{\alpha_k}|\sqrt{s_k(1 - s_k)}$, where the first term $\hat{\alpha_k}$ is the coefficient value calculated above, measuring the estimated predictive relevance. And the second term $\sqrt{s_k(1 - s_k)}$ represents the standard deviation, in order to mitigate the impact of features' scales. $s_k = \frac{1}{N}\sum_{i=1}^{N} r_k(\mathbf{x}_i)$ is the support on the training data, which means the proportion of data point where the specific decision rule k can applies on training data. The importance of original feature is calculated as: $I_j = |\hat{b_j}|std(l_j(x_j))$. $std(l_j(x_j))$ is the standard deviation of $l_j(_j)$. I_k and I_j measure global feature importance. An advantage of RuleFit is no variables or rules are dropped before being fitted into a linear model, while inspirable decision rules or variables might be excluded if some of them have been removed for the purpose of dimensional reduction.

Knowledge Filtering The linear model of RuleFit takes both original features and features constructed through decision rules into account. The final feature space is large and is difficult to interpret. It is a straightforward way to rank decision rules and individual variables by feature importance. However, there are no guarantees that all extracted knowledge are consistent with medical logic. To maintain reasonable knowledge, we design a filtering method, incorporating medical experts, to revise extracted knowledge. As shown in table 2, there are three kinds of criteria. Firstly, decision rules associated with calculated mortality rate, from original data, lower than 7.2% are rejected, this is because we far more concern decision rules that are likely to cause death. Secondly, reserved rules would be ranked by feature importance, and two cardiologists review them and mark them by two labels: *is content with medical common sense* and *is inspirable*. The difference of their reviewing results would be re-checked by a more experienced cardiologist to reach a final decision. All rules marked as **NO** with the first label would be filtered out. The second label regrading inspiration aims to detect knowledge that might be supportive to unverified medical hypotheses.

Criteria	Priority	Primary
Mortality	1	Yes
Medical Expertise	1	Yes
Feature Importance	2	No

Table 2. Knowledge filtering criteria

3.6 Evaluation Metrics

We adopt sensitivity, specificity, accuracy, AUC and ROC to evaluate performance of predictive models. The first two metrics are popular in the medical domain, and they can mathematically describe the performance of predictive models on high risk patients and low risk patients.

4 Experimental Results

4.1 Experimental setup

In our study, we divide the original dataset into training set and testing set with a ratio of 7:3 (training set with 9521 samples and testing set with 4081 samples). To address the data imbalance issue, we employ the Boarderline SMOTE algorithm [40] which is a over-sampling method generating samples for the minority class. We implement 10-fold cross validation on the training set. In terms of the Gradient Boosting Classifier, we train 100 classifiers, and the depth of each tree model is set to 3, and nodes will not split if there are less than 182 samples associated with them, the minimum number of samples required to be at a leaf node is set to 15. The decision rules are integrated into additional feature sets, combined with original feature sets, work as input to the logistic regression model.

4.2 Performance of the Predictive Models

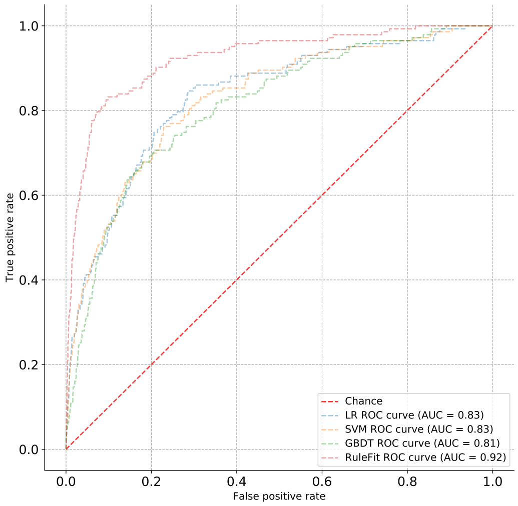

Fig. 1. ROC curves of PMs

Figure 1 shows ROC curves of four predictive models, the results of evaluation metrics are available in Table 3. According to Figure 1, RuleFit achieves a AUC of 0.92 and outperforms other models with a large margin, and SVM is better than LR and GBDT thanks to its outstanding performance in binary classification tasks. The tree model component of RuleFit implements feature interaction and selection, this proves the advantage of non-linear features. RuleFit enjoys the

Model	Accuracy	Sensitivity	Specificity	AUC
LR	0.83	0.66	0.84	0.83
SVM	0.85	0.63	0.85	0.83
GBDT	0.88	0.54	0.89	0.81
RuleFit	0.97	0.45	0.98	0.92

Table 3. Evaluation results

best performance in both accuracy(0.97) and specificity(0.98), but its sensitivity value of 0.45 is dramatically lower than that of other models. The sensitivity value of LR(0.66) is the best among four PMs, and all models show worse sensitivity than their specificity values. This means they are not good at detecting high risk HF patients, but are more likely to make correct decisions for low risk HF patients. Another reason is the unbalanced source data.

Rule	Mortality (%)	Importance	|Coef|	Death toll
Mono% low: no Urea high: yes CK-MB high: yes	19.74	0.047	0.23	92
gender: female hs-cTnl high: yes UA normal: no	19.39	0.090	0.37	145
temperature low: yes Ca high : no age>81.5	17.31	0.014	0.12	9
TP low: yes cardi-surgery history: yes GGT normal: no	15.00	0.016	0.13	12
PLT normal: yes age>81.5	12.96	0.142	0.39	245

Table 4. Examples of Mined Knowledge

4.3 Evaluation of Mined Knowledge

After filtering procedures, there are 110 (34.38% of initial rules) valid rules left. Table 4 reveals top-ranked decision rules and their statistical characteristics. The average mortality rate corresponded with extracted rules is 6.26%, and 36 rules (11.25%) exceed the average Chinese HF mortality rate(7.2%).

According to physicians' evaluation, there is no any knowledge against medical common sense. For instance, the second rule in table 4, *‘gender: female hs-cTnI high: yes UA normal: no’*, interprated as "a female patient with high hs-cTni and abnormal UA", is in line with previous findings. [17] confirms hs-cTnI as a useful biomarker for CHF patients and uric acid is an important prognostic marker for all-cause mortality of HF [29]. Also, some rules are of inspiration and are evident to unconfirmed medical hypotheses. For example, the rule *‘Mono% low: no Urea high: yes CK-MB high: yes’* demonstrates the significance of *Mono%* which has been proved by [32] relevant with the pathogenesis of cardiovascular diseases, but its impact on HF is unclear. The rule *’ALP high: yes ChE low: yes gender: male’* confirms that higher level of ALP and lower level of ChE are acceerative factors to the death of HF patients, whereas these two biomarkers have not been seriousely considered before and are worthy of further investigation.

A highly interesting discovery is the rule of *‘PLT normal: yes age >81.5’*. It is well known and has been verified in [19] that age is a high-risky factor to HF. However, in clinical scenarios, normal level of PLT is generally not an influential factor not to mention interprating it a more important factor than age. The inverse association is referred as ”reverse epidemiology” or the ”risk factor paradox”, and it deserves more research.

5 Conclusion

In this work, we test four predictive models on the task of HF prognosis risk evaluation, and incorporate RuleFit and medical expertise to mine knowledge in an interpratable manner. The extracted knowledge, screened by our knowledge filtering method, is reasonable statistically and medically.

We found that detecting highly risky HF patients is difficult but predicting outcomes of HF patients with low risks is easier. Our filtering method is helpful to reject unacceptable results, though it is not scalable. Some of extracted knowledge is valuable in providing statistical evidence to support physicians' hypotheses, and is also inspirable in discovering novel variables that have not been considered before.

6 Future Work

Despite the compelling results from our models, our work can be improved in several aspects. Firstly, rules embedded with intrinsic temporal dependencies are helpful to mine knowledge of different clinical stages. Secondly, RuleFit equally treats each decision rule derived from the tree model component, conversely their spatial dependencies should be taken into account in the linear model. Thirdly, more features can be considered, like the examinations of cardiac ultrasound, cardiac synchronization therapy. Last but not least, scalably automatic evaluation on the extracted knowledge is non-trivial and indispensable. The difficulty towards automatic evaluation of medical knowledge mining stems from the medical expertise behind it. We suggest a multi-task learning solution to mine knowledge from a wide range of heart diseases in order to decrease the usage of domain knowledge.

References

1. Aronson D., Mittleman M A., Burger A J.: Elevated blood urea nitrogen level as a predictor of mortality in patients admitted for decompensated heart failure. The American journal of medicine, **116**(7), 466–473 (2004)

2. Bhatla N., Jyoti K.: An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering **1**(8), 1–4 (2012)
3. Chandola V., Sukumar S R., Schryver J C.: Knowledge discovery from massive healthcare claims data.Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 1312–1320 (2013)
4. Choi D J., Park J J., Ali T., et al.: Artificial intelligence for the diagnosis of heart failure. NPJ digital medicine **3**(1), 1–6 (2020)
5. Curcio F., Sasso G., Liguori I., et al.: The reverse metabolic syndrome in the elderly: Is it a "catabolic" syndrome?. Aging clinical and experimental research **30**(6), 547–554 (2018)
6. Damen J A A G., Hooft L,, Schuit E,, et al.: Prediction models for cardiovascular disease risk in the general population: systematic review. bmj **353**, (2016)
7. Dangare C S., Apte S S.: Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications **47**(10), 44–48 (2012)
8. Demissei B G., Cotter G., Prescott M F., et al.: A multimarker multi-time point-based risk stratification strategy in acute heart failure: results from the RELAX-AHF trial. European journal of heart failure **19**(8),1001–1010 (2017)
9. Esfandiari N., Babavalian M R., Moghadam A M E., et al.: Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications **41**(9), 4434–4463 (2014)
10. Fonarow G C., Adams K F., Abraham W T., et al.: Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. Jama **293**(5), 572–580 (2005)
11. Friedman J H., Popescu B E.: Predictive learning via rule ensembles. Annals of Applied Statistics **2**(3), 916–954 (2008)
12. Frizzell J D., Liang L., Schulte P J., et al.: Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA cardiology **2**(2), 204–209 (2017)
13. Golas S B., Shibahara T., Agboola S., et al.: A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC medical informatics and decision making **18**(1), 1–17 (2018)
14. Hussain M., Afzal M., Ali T., et al.: Data-driven knowledge acquisition, validation, and transformation into HL7 Arden Syntax. Artificial intelligence in medicine **92**, 51–70 (2018)
15. Kadi I., Idri A., Fernandez-Aleman J L.: Knowledge discovery in cardiology: A systematic literature review. International journal of medical informatics **97**, 12–32 (2017)
16. Kalantar-Zadeh K., Block G., Horwich T., et al.: Reverse epidemiology of conventional cardiovascular risk factors in patients with chronic heart failure. Journal of the American College of Cardiology **43**(8), 1439–1444 (2004)
17. Kawahara C., Tsutamoto T., Sakai H., et al.: Prognostic value of serial measurements of highly sensitive cardiac troponin I in stable outpatients with nonischemic chronic heart failure. American heart journal **162**(4), 639–645 (2011)
18. Lavrač N., Zupan B.: Data mining in medicine. Data Mining and knowledge discovery handbook. Springer, Boston (2005)
19. Lee D S., Austin P C., Rouleau J L., et al.: Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. Jama **290**(19), 2581–2587 (2003)

20. Levy W C., Mozaffarian D., Linker D T., et al.: The Seattle heart failure model. Circulation **113**(11), 1424–1433 (2006)
21. Levy W C., Aaronson K D., Dardas T F., et al.: Prognostic impact of the addition of peak oxygen consumption to the Seattle Heart Failure Model in a transplant referral population. The Journal of heart and lung transplantation **31**(8), 817–824 (2012)
22. Li J., Tan X., Xu X., et al.: Efficient mining template of predictive temporal clinical event patterns from Patient Electronic Medical Records. IEEE journal of biomedical and health informatics **23**(5), 2138–2147 (2018)
23. Lv H., Yang X., Wang B., et al.: Machine Learning–Driven Models to Predict Prognostic Outcomes in Patients Hospitalized With Heart Failure Using Electronic Health Records: Retrospective Study. Journal of medical Internet research **23**(4), e24996 (2021)
24. Maggioni A P., Dahlström U., Filippatos G., et al.: EURObservational Research Programme: regional differences and 1-year follow-up results of the Heart Failure Pilot Survey (ESC-HF Pilot). European journal of heart failure **15**(7), 808–817 (2013)
25. Miller T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
26. Mosterd A., Hoes A W.: Clinical epidemiology of heart failure. heart **93**(9), 1137–1146 (2007)
27. van Riet E E S., Hoes A W., Wagenaar K P., et al.: Epidemiology of heart failure: the prevalence of heart failure and ventricular dysfunction in older adults over time. A systematic review. European journal of heart failure **18**(3), 242–252 (2016)
28. Lauren S., Michael M G.: Acute heart failure. Trends in cardiovascular medicine **30**(2), 104—112 (2020)
29. Tamariz L,. Harzand A., Palacio A., et al.: Uric acid as a predictor of all-cause mortality in heart failure: a meta-analysis. Congestive heart failure **17**(1), 25–30 (2011)
30. Tang F., Xiao C., Wang F., et al.: Predictive modeling in urgent care: a comparative study of machine learning approaches. Jamia Open **1**(1), 87–98 (2018)
31. Taslimitehrani V., Dong G., Pereira N L., et al.: Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. Journal of biomedical informatics **60**, 260–269 (2016)
32. Wrigley B J., Lip G Y H., Shantsila E.: The role of monocytes and inflammation in the pathophysiology of heart failure. European journal of heart failure **13**(11), 1161–1171 (2011)
33. Molnar C. Interpretable machine learning. Lulu. com, (2020)
34. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ”Why should I trust you?: Explaining the predictions of any classifier.” Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).
35. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. ”Anchors: High-Precision Model-Agnostic Explanations.” AAAI Conference on Artificial Intelligence (AAAI), 2018
36. Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. ”Examples are not enough, learn to criticize! Criticism for interpretability.” Advances in Neural Information Processing Systems (2016).
37. Friedman J, Popescu B. Gradient directed regularization for linear regression and classification. In Tech rep Stanford University, Department of Statistics (2004)

38. Lundberg S, Lee S I. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874,(2017).
39. Guidotti R, Monreale A, Ruggieri S, et al. Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820, (2018)
40. Han, H., Wang, W.-Y., Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In ICIC'05 Proceedings of the 2005 international conference on Advances in Intelligent Computing - Volume Part I (pp. 878–887).