Explainable machine learning approach with augmentation for mortality prediction

Firas Ketata
FEMTO-ST institute
Univ. Bourgogne Franche-Comté
ENSMM, CNRS
Besançon, France
firas.ketata@femto-st.fr

Zeina Al Masry
FEMTO-ST institute
Univ. Bourgogne Franche-Comté
ENSMM, CNRS
Besançon, France
zeina.almasry@femto-st.fr

Noureddine Zerhouni
FEMTO-ST institute
Univ. Bourgogne Franche-Comté
ENSMM, CNRS
Besançon, France
noureddine.zerhouni@femto-st.fr

Slim Yacoub
Laboratoire de Télédétection et
Systèmes d'Information à Référence Spatiale
Nat. Inst. of Applied Science and Technology
University of Carthage
Tunis, Tunisia
Slim.yacoub@insat.rnu.tn

***Abstract*—Cardiovascular diseases kill approximately 17.7 million people worldwide each year. They mainly occur in the form of myocardial infarction and heart failure. In this context, electronic medical records of patients with their physical characteristics and clinical laboratory test values are available. Biostatistical methods and machine learning (ML) techniques have already been used to find associations between patient characteristics and to predict the mortality in heart failure patients. However, ML models still not applicable in clinics and critical medical conditions. This may be due to the lack of explainability and clarity of ML prediction tools among physicians. Thus, the objective of this study is to propose an explainable approach to support physicians in their decision-making. This approach is based on several ML techniques combined with Shapley values. The goal is to increase the risk coefficients applied by Shapley with the k-fold technique in order to maximize the reliability of the explainability even for small datasets. The proposed approach is validated using the heart failure prediction public dataset. The explainability showed that the ejection fraction and serum creatinine variables are the most important and decisive for the prediction of mortality for patients with heart disease. Finally, the application of the k-fold technique with Shapley values allowed to improve the ranking of feature importance for mortality prediction and to provide meaningful visualization graphs.**

I. Introduction

Cardiovascular disease (CVD) is a disease of the heart and blood vessels, including cerebrovascular disease (stroke), heart failure (HF), and other types of conditions. Cardiovascular diseases account for approximately 17.7 million deaths, or 31% of global mortality [1]. In particular, according to the World Health Organization (WHO), heart failure is considered one of the leading causes of death worldwide. Heart failure occurs when the heart cannot pump enough blood to meet the body's needs, usually caused by diabetes, high blood pressure, or other illnesses [2]. It is responsible for 55,000 deaths each year. In the United States, 230,000 additional deaths were due to its indirect contribution. About 90% of patients with advanced heart disease die within a year [3]. Many electronic health records, also known as medical records, can be considered as useful information resources for the diagnosis or prediction of disease or mortality using machine learning. Many studies have proposed approaches using ML tools to predict the mortality risk fro heart failure. However, this still not applicable to critical medical conditions. This may be due to the lack of explainability and clarity of ML predictive tools among physicians. Additionally, it is very important for physicians to understand the causes associated with patient mortality and what are the most important risk variables when making predictions through ML. The aim of this study is to propose an approach based on explainable ML, Shapley values and k-fold technique to predict mortality in patients with heart failure by ensuring a maximum explainability. This explainability is largely based on the extraction of Shapley coefficients or risk factors for each input. A single prediction iteration to extract these coefficients may not be sufficient to guarantee the reliability of explainability, especially for small datasets. For that purpose, Shapley values allowed to increase the risk coefficients and to provide clearer decision support to physicians. The approach is applied on a public dataset [4] for heart failure prediction. This paper is organized as follows. Section 2 presents the related work. Section 3 provides the description of the dataset as well as the proposed methodology. The results and discussion are given in Section 4. Finally, conclusion and future work are described in Section 5.

II. Related work

In [7], the authors used time-based Cox regression and traditional Kaplan-Meier statistical estimates to identify significant predictors of heart failure (HF) mortality in 299 Pakistani

patients [4]. The authors concluded that age, serum creatinine, arterial hypertension were the most important features responsible for the high mortality in AD patients suffering from cardiovascular failure. Later, [8] developed a sex-based survival prediction model using the same dataset. The authors found that survival prediction patterns were significantly different for men than for women. For men, smoking, diabetes, and anemia were important features, while ejection fraction, sodium, and platelet count were important risk factors for women. Both of the aforementioned studies presented interesting results using statistical methods. Subsequently, [5] were used to apply data mining techniques and ML methods. These models were developed to predict patient survival and then ranked the most important features contained in medical records. However, only two features – ejection fraction and serum creatinine- were used in their ML analysis shown by the Gini approach of the random forest. A new version of the same dataset is provided to the UCI repository [5] for machine learning methods. After, [9] analyzed the new version of the dataset using the Synthetic Minority Oversampling (SMOTE) technique used in nine classification models to uncover important features and improve the machine learning models. The authors demonstrated that the ETC model achieved 92.6% accuracy in predicting patient survival. There are other studies that exploit the same dataset by incorporating the concept of machine learning interpretability. In [10], the authors compared different ML models and then used random forests to extract importance coefficients from each INPUT during prediction. The results show that the decision tree and random forest algorithms achieve the highest accuracy of 95% among the classifiers. An interpretable method called "DERExx" was developed in [11]. This method relies on scalable algorithms and provides users with an easy-to-understand set of IF-THEN rules that include data set parameters. Finally, [12] implements a ML classification algorithm to predict mortality in heart failure patients using patient-specific age risk factors. The problem of object class imbalance is handled by oversampling techniques. The results showed that the LGBM achieved the best accuracy of 96.8% in predicting survival of heart failure patients.

III. Methodology

Among the models used in the literature, random forest models and Gini methods have been used to add some machine learning interpretability to mortality prediction by exploiting the public dataset [4]. This interpretability is used to display the importance coefficients of each variable on the prediction. But in the context of medical decision support, it is still not enough for physicians to use machine learning models in critical situations. Generally, correlation studies added with importance coefficients to give confidence to these coefficients. Moreover, for small datasets, the interpretability of Gini and random forest lack of reliability in the extraction of importance coefficients. Thus, we propose to use Shapley method with the best model adapted to the dataset to predict mortality in patients with heart failure using a public dataset published on Kaggle [4]. Applying Shapley method will provide more explainability for ML tools, which combines the importance of each feature for prediction with the relationship between those features and the output, given the distribution of input values. Except that the application of an explainability approach on a database with only 299 cases decreases the reliability of the explainability. This may be an insufficient amount of data to determine the true significance of each INPUT in predicting mortality. In other words, each time we change the training and test data, the importance of each variable on the output changes. This results in a ranking of the importance of the inputs on the prediction which is not reliable. Therefore, we propose in this paper to combine the k-fold approach with Shapley to increase the importance coefficients of predictive features. This augmentation will ensure an importance ranking of each input on the output with respect to the fixed and unchangeable prediction. This combination will increase the reliability of importance coefficient extraction and thus confidence in the explanation, especially for small datasets. This will be beneficial to implement the ML-based decision support system in critical medical conditions for mortality prediction.

A. Methodology Description

The proposed approach starts with data analysis using visualization tools to understand the data and estimate the effective pre-processing before the learning phase. Next, machine learning models (logistic regression (LR), support vetor machine (SVM), XG-boost (XGB), random forest (RF)) are applied to predict the risk of patient death. These models are then compared by cross-validation, more precisely by measuring accuracy, precision and Recall. Finally, Shapley's values will be applied to the most effective model to ensure explainability to medical experts. This explainability is mainly based on the extraction of coefficients risk or on the importance of features in the forecasting process. Therefore, the increase in risk factor based on the k-fold technique will be applied with Shapley to obtain reliable risk coefficients and explainable visualizations Figure 1.

B. Description and analysis of the dataset

1) Statistical analysis: In this study, the heart failure clinical records dataset [4] was downloaded from Kaggle. This dataset provides the medical records of 299 patients with heart failure. The dataset includes 194 men and 105 women with 13 features presented in Table I .

The patients were between 40 and 95 years old. Some features (for example, anemia, diabetes, high blood pressure, sex, and smoking) are binary, while others are numeric Table II. There are differences in proportions between the variables. For example, between the variables platelets, creatinine phosphokinase, and others. This difference is shown in Table II where statistical summary is provided. Therefore, scaling is a technique that is needed later in the data pre-processing process to ensure that all features are scaled equally. This scaling can be beneficial when training the models.

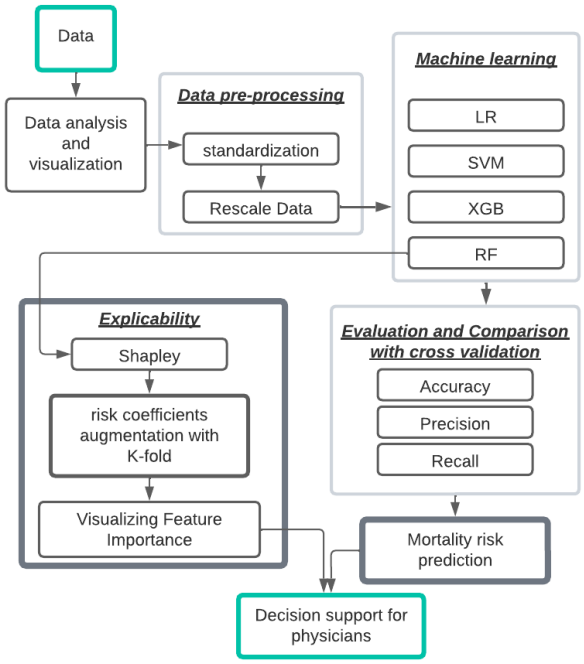

Fig. 1: Process of the methodology

Features	Description
age	age of patient
anaemia	decrease of red blood cells or hemoglobin
creatinine_phosphokinase	level of the CPK in the blood
diabetes	if the patient has diabetes
ejection_fraction	% of blood at each contraction
high_blood_pressure	if the patient has hypertension
platelets	platelets in the blood
serum_creatinine	level of serum creatinine in the blood
serum_sodium	level of serum sodium in the blood
sex	woman or man
smoking	if the patient smokes or not
time	follow-up period
DEATH_EVENT	if the patient deceased

TABLE I: Features description

Feature	mean	std	min	max
age	60.83	11.89	40.0	95.0
anaemia	0.43	0.49	0.0	1.0
creatinine_phosphokinase	581.83	970.28	23.0	7861.0
diabetes	0.41	0.49	0.0	1.0
ejection_fraction	38.08	11.83	14.0	80.0
high_blood_pressure	0.35	0.47	0.0	1.0
platelets	263358	97804.2	25100	850000
serum_creatinine	1.39	1.03	0.5	9.4
serum_sodium	136.62	4.41	113.0	148.0
sex	0.64	0.47	0.0	1.0
smoking	0.32	0.46	0.0	1.0
time	130.26	77.61	4.0	285.0
DEATH_EVENT	0.32	0.46	0.0	1.0

TABLE II: Statistical summary of the dataset

2) Data distribution analysis: When we visualize distributions of numerical data, we notice that most of the distributions have a Gaussian tendency (Figure 2). Hence the idea of applying standardization to the data at a later stage of the data preprocessing step. It is also important to examine the data volume distribution of the output columns. The number of positive and null values is fairly distributed compared to other medical datasets. 200 patients are alive and 99 patients have died. Therefore, the amount and distribution of data may be sufficient to train and test the model.

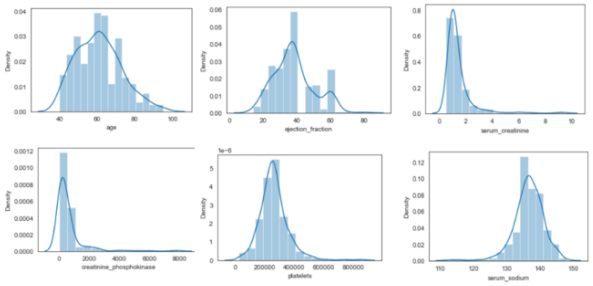

Fig. 2: Distribution of digital inputs

C. Data pre-processing

1) Rscale Data: The idea is to scale the properties in a range between 0 and 1 using the following (1) [13]:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

This scaling is useful for optimization algorithms that are central to machine learning algorithms like gradient descent. It is also useful for algorithms that weight inputs, such as regression and neural networks, and algorithms that use measures of distance, such as k-nearest neighbors.

2) Standardisation: Standardization is a useful technique to transform attributes with a Gaussian distribution and different means and standard deviations into a standard Gaussian distribution with mean 0 and standard deviation 1 using the following (2) [14].

$$x' = \frac{x - mean(x)}{std(x)} \quad (2)$$

It is best suited for techniques such as linear regression, logistic regression, and linear discriminant analysis which assume a Gaussian distribution for input variables.

D. The used models

We briefly present the used models in our study.

1) Logistic regression: Logistic regression is a statistical model that studies the relationship between a set of qualitative variables Xi and a qualitative variable Y. It is a generalized linear model using a logistic function as a link function [15].

2) Support Vector Machine: A Support Vector Machine (SVM) is a machine learning algorithm that performs supervised learning to classify or regress a dataset. Support Vector Machines are used to classify two sets of data based on similar classifications. The algorithm separates groups by drawing lines (hyperplanes) based on the model.[16].

3) Random Forest: The random forest technique consists of many decision trees that classify a data frame individually. Each decision tree sorts the same data according to their respective optimal distribution. Using random forests gives us minimal overfitting: the major risk of overfitting is reduced due to the use of multiple trees. Also, it provides high accuracy (the algorithm performs well on large datasets, and accuracy improves as the quality of the data used for training improves). [17].

4) XG-boost: The idea of eXtreme Gradient Boosting (Xgboost) is not to use a single model, but to use several models and then combine them to obtain a single result. It is above all a practical approach, making it possible to manage the problems of regression and classification. The algorithm works in sequence. Unlike random forests, for example. This way of doing will of course make it slower, but above all it improves the algorithm by capitalization compared to previous executions [18].

The hyperparameters used for all models have been presented in Table III :

Models	Hyperparameters
RL	Solver=lbfgs, C=1, max_iter=100
SVM	kernel=linear, C=1, cache_size=200, decision_function=avr
RF	n_estimators=50, random_state=42, criterion='Gini'
Xgb	Default configurations

TABLE III: Hyperparameters used for all models

E. Evaluation tools

Evaluation and comparison between models was performed by k-fold cross-validation with k = 5 iterations. The estimators used are accuracy, recall and precision, using the following equations [19]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

where TP refers to the true positive, FP is the false positive, TN refers to the true negative and FN to the false negative.

F. The shapley approach

Shapley explainability is based on the extraction of Shapley coefficients, which are then used to determine and visualize the importance of each variable for prediction. The Shapley value is defined as the marginal contribution of a variable value to predicting all possible "coalitions" or subsets of features. In other words, it's a way to redistribute overall benefits among traits, provided they are all cooperative[20].

In this study, the implementation of Shapley coefficients approach with visualizations for explainability were developed by the Python SHAP library.

G. k-fold with shapley (Coefficients increase)

The k-fold technique is used to divide the original sample into k samples or files and then select one of the k samples as the validation set, while the other k-1 samples are the training set for the model learning. Then some reminders that the data resulting from the validation will be the data used for the training, while another training file will be used for validation. Finally, we get a list containing 5 sub-lists of shapley's values, each sub-list is due to a prediction with different training and test data. This increases reliability when visualizing the interpretability of each variable's importance to prediction. In this study, we choose the number of files k equal to 5 as shown in Figure 3.

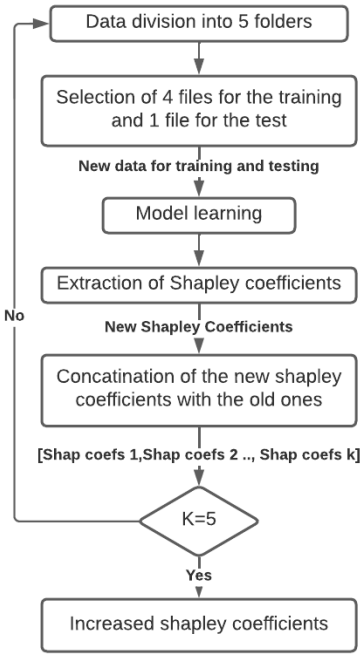

Fig. 3: The k-fold technique with Shapley

IV. Results and discussion

In this study, several existing libraries in python were used. The Sklearn library [21] is used to build the classification algorithms. The Pandas and NumPy libraries [22] are used to read and manipulate the data. The visualization of two-dimensional (2D) graphs is done with the Seaborn library [23]. Finally, the SHAP library [6] is used to manipulate and visualize the Shapley coefficients and the importance of each input to the prediction.

A. Prediction of mortality for patients with heart failure

The Random Forest admits an accuracy of 0.833, higher than that of the linear regression, XGB and SVM equal respectively to 0.823, 0.810, 0.806. On the other hand, the linear regression admits a better precision compared to the other models. Hence, according to the formula, the false positive values is very low for the linear regression, so it is

the best model to detect the presence of mortality risk. Then, for the Recall, the SVM is the best with a recall equal to 0.71. Therefore, the false negative values are very few. Therefore, it is the best model to detect the absence of mortality risk. All the evaluators for all the models are presented in IV.

Models	Accuracy	Precision	Recall
RL	0.823	0.72	0.62
SVM	0.806	0.69	0.71
RF	0.833	0.675	0.672
Xgb	0.81	0.59	0.62

TABLE IV: Models evaluation

B. Explainability with SHAP

We applied Shapley with the model that admits the best accuracy which is the random forest. The diagram presented in 4 is generated with Python's SHAP library [6]. The y-axis are the features (inputs) and the Shapley values are shown in the x-axis. The color degradation represents whether the values of each feature are large or small (red color: the largest values, blue color: the smallest values). Since Shapley is a local explainability approach, each point on the graph presents a Shapley value for each specific feature related to a specific patient. Thus, the number of points for each INPUT is equal to the number of patients = 299. Features are ranked from most important (top) to least important (bottom) in predicting mortality. When the smallest values of a feature admit negative Shapley values and the largest admit positive Shapley values, in this case, the greater this feature, the higher the risk of mortality (Example: serum creatinine). If the largest values of a feature admit negative Shapley values, and the smallest admit positive Shapley values, then the smaller this feature, the higher the risk of mortality (Example: ejection fraction). The explainability showed that the variable serum creatinine is the most important and the most decisive for the prediction of mortality. Then, ejection fraction and platelets are also decisive for the prediction but a little less decisive than serum creatinine. Then we find the other variables (smoking, age, serum sodium, high blood pressure, creatinine phosphokinase, anemia, diabetes, sex) ranked from most important to least important for the prediction. But the ranking of the importance of each input on the prediction of the risk of mortality is changed each time the training and test data are modified during the training of the associated model (random forest) with Shapley.

C. Increasing Shapley coefficients with k-fold

Figure 5 shows the impact of the features on the risque prediction after the application of the technique for multiplication of the Shapley coefficients. Hence, the length of Shapley values has been multiplied by 5. Thus, we see a bigger cloud of points in Figure 5 compared to that of Figure 4. We have 299 times 5 points instead of 299. There is an impressive change in the order of feature importance on the prediction of mortality presented in Figure 5 compared to that of Figure 4. After

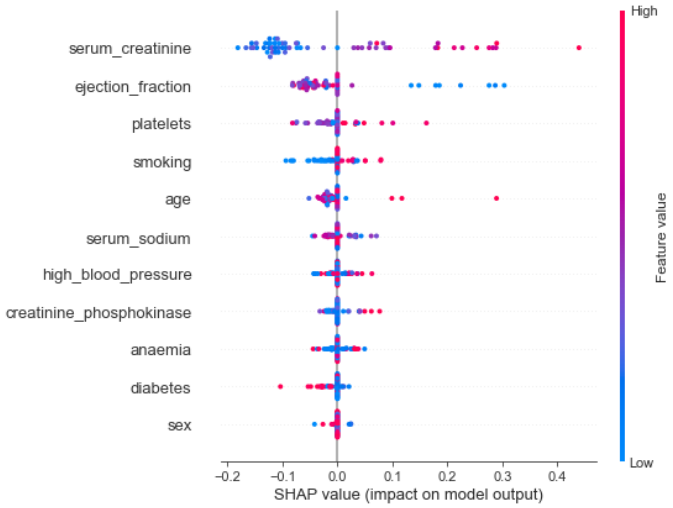

Fig. 4: Shapley Visualization (impact features on model output)

applying the Shapley coefficient multiplication technique, we notice that the ranking of the most important inputs on mortality prediction has been fixed. When testing with any test and training data when learning random forest, the ranking remains the same. Figure 5 presents the final ranking of the inputs on mortality prediction. The most important variable became ejection fraction then serum creatinine contrary to what is presented Figure 4. Age has become more important than platelets and tuxedos. Also, creatinine phosphokinase, it is presented as more important than high blood pressure. Also, the relationship between features and output exhibited by staining degradation was clearer and more significant after applying increased shapley coefficients. As we can see in Figure 5, more the features (serum creatine, age, creatine phosphokinase, hign blood pressure, anemia) are high, the more the risk of mortality is important. On the other hand, for the features (ejection fraction, serum soduim) more they are small, more the risk of mortality is important. So, the Figure 5 is presented in a clearer way to the physicians in order to understand the most important coefficients in the prediction and to see the trend of the influence of each variable in relation with the output. Hence, the application of the k-fold technique with the Shapley values offers a more reliable importance ranking with more meaningful visualization graphs of explainability.

V. Conclusion

In this paper, we have presented an approach based on explainable machine learning to predict mortality in heart failure patients, while comparing several ML models. The purpose of this approach is to ensure a maximum of explainability with visualization tools using the Shapley method and the SHAP python library. This explainability may support physicians with their decision-making. Next, we apply the k-fold method to multiply Shapley's coefficients in order

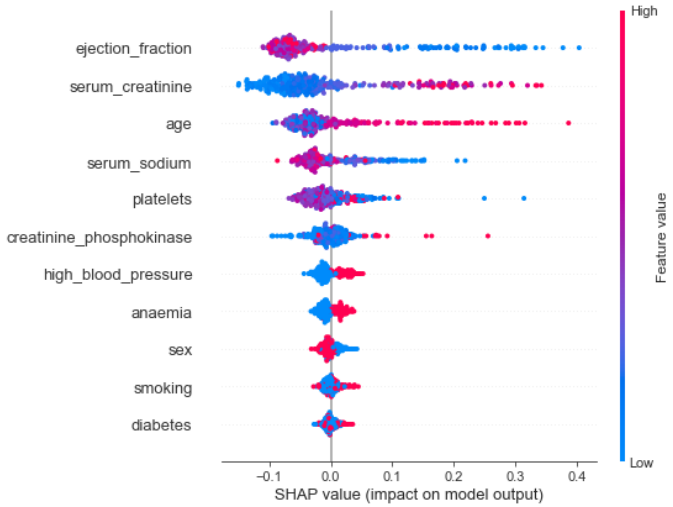

Fig. 5: Shapley Visualization with k-fold (K=5)

increase the reliability of the proposed approach. The results showed that the random forest admits a better accuracy of 0.833, which is higher than the one of the linear regression, XGB and SVM equal respectively to 0.823, 0.810, 0.806. On the other hand, the linear regression admits a better precision of 0.72 compared to the other models and a better recall of 0.71 for the support vector machine. We found out that the ejection fraction variable is the most important and the most decisive for the prediction of mortality. Serum creatinine and Age are also decisive variables for the prediction but a little less decisive than serum creatinine. The other variables (serum sodium, platelets, creatinine phosphokinase, high blood pressure, anemia, sex, smoking, diabetes) are ranked from most important to least important for the prediction. Finally, the application of the k-fold technique with Shapley's values offered a more reliable ranking of the importance of the features for the prediction of mortality, with more meaningful visualization graphs of explainability.

REFERENCES

[1] World Health Organization, World Heart Day, *https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)*, Accessed 27 November 2022.

[2] National Heart Lung and Blood Institute. Heart failure, *https://www.nhlbi.nih.gov/health-topics/heart-failure*. Accessed 27 November 2022.

[3] A. N. Nowbar, M. Gitto, J. P. Howard, D. P. Francis and R. Al-Lamee, *Mortality from ischemic heart disease: Analysis of data from the world health organization and coronary artery disease risk factors from NCD risk factor collaboration*, Circulation Cardiovascular Quality and Outcomes. 2019.

[4] D. Chicco, *Heart failure clinical records Data Set*, https://archive.ics.uci.edu/ml/datasets/Heart+ failure+clinical+records . 2022.

[5] D. Chicco and G. Jurman, *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*, BMC Med. Inform. Decis. Mak. 2020.

[6] Poduska, Joshua, *SHAP and LIME Python Libraries: Part 1–Great Explainers, with Pros and Cons to Both* 2018.

[7] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, *Survival analysis of heart failure patients: A case study*, PLoS One. 2017.

[8] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, *Gender based survival prediction models for heart failure patients: A case study in Pakistan*, PLoS One. 2019.

[9] Ishaq, Abid, Saima Sadiq, Muhammad Umer, Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara, and Michele Nappi, *Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques.*, EEE access. 2021.

[10] Rahman, Polin, Ahmed Rifat, MD IftehadAmjad Chy, Mohammad Monirujjaman Khan, Mehedi Masud, and Sultan Aljahdali, *Machine Learning and Artificial Neural Network for Predicting Heart Failure Risk*. 2021.

[11] Sannino, Giovanna, Giuseppe De Pietro, and Ivanoe De Falco, *Automatic Extraction of Interpretable Knowledge to Predict the Survival of Patients with Heart Failure*, IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) IEEE. 2021.

[12] Lee and HeeJeong Jasmine, *Comparative Study on Prediction of Mortality in Heart Failure Patients using Nine Machine Learning Algorithms*, Journal of the Korean Society of Information Technology. 2022.

[13] Ilyas, Ihab F and and Xu Chu, *Data cleaning*, Morgan and Claypool. 2019.

[14] Chu, Xu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang, *Data cleaning: Overview and emerging challenges*, Proceedings of the 2016 international conference on management of data. 2016.

[15] Zhou, Zhi-Hua, *Machine learning*, Springer Nature. 2021.

[16] Harrington, Peter, *Machine learning in action*, Simon and Schuster. 2012.

[17] Alzubi, Jafar, Anand Nayyar and and Akshi Kumar, *Machine learning from theory to algorithms: an overview*, Journal of physics: conference series. 2018.

[18] Chen, T. and Guestrin, *XGBoost: A scalable tree boosting system*, . In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[19] Zuluaga-Gomez J, Al Masry Z, Benaggoune K, Meraghni S, Zerhouni N, *A CNNbased methodology for breast cancer diagnosis using thermal images*, Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization. 2021.

[20] Rozemberczki, Benedek, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar, *The Shapley Value in Machine Learning* , arXiv preprint. 2022.

[21] Pedregosa F, Varoquaux, Ga"el, Gramfort A, Michel V, Thirion B, Grisel O, and others, *Scikit-learn: Machine learning in Python*, Journal of machine learning research. 2011.

[22] McKinney W and others, *Data structures for statistical computing in python*, Proceedings of the 9th Python in Science Conference. 2010.

[23] Waskom, M. and others, *mwaskom/seaborn: v0.8.1*, Zenodo. 2017.