American Journal of Clinical and Experimental Medicine
2023; 11(2): 33-38
http://www.sciencepublishinggroup.com/j/ajcem
doi: 10.11648/j.ajcem.20231102.12
ISSN: 2330-8125 (Print); ISSN: 2330-8133 (Online)

Performance Evaluation of Machine Learning Methods for Heart Failure Prediction

Jing Xia*, Xiaoying Wang

Medical School, Jiangnan University, Wuxi, China

Email address:
1282210120@stu.jiangnan.edu.cn (Jing Xia), xiaoyingwang@jiangnan.edu.cn (Xiaoying Wang)
*Corresponding author

To cite this article:
Jing Xia, Xiaoying Wang. Performance Evaluation of Machine Learning Methods for Heart Failure Prediction. *American Journal of Clinical and Experimental Medicine*. Vol. 11, No. 2, 2023, pp. 33-38. doi: 10.11648/j.ajcem.20231102.12

Received: February 21, 2023; **Accepted**: March 22, 2023; **Published**: March 31, 2023

Abstract: Heart failure is a syndrome of cardiac circulation disorder. Due to the dysfunction of the systolic function or diastolic function of the heart, the venous blood volume cannot be fully discharged from the heart, resulting in blood stasis in the venous system and insufficient perfusion in the arterial system. The symptoms of this disorder are concentrated in pulmonary congestion and vena cava congestion. The correlation between the inducement of heart failure and the incidence of heart failure is a subject that needs to be studied in the medical field. In recent years, with the development of data mining technology, more and more analytical models and algorithms have been applied in the medical field, which greatly improve the efficiency of medical data analysis and enable medical workers to cure diseases better. In this study, an ensemble learning model is applied to analyze the data of heart failure. First, the data is preprocessed and normalized, and features that are not associated with death rate of heart failure are removed. Secondly, multiple base classifiers are trained and compared. Finally, the competent base classifiers are selected and integrated with the Stacking-based ensemble learning algorithm for final classification. Comparative analysis showed that the prediction results of ensemble model are better than that of base classifiers in evaluation indexes such as accuracy, precision, AUC, Balanced accuracy and F1-score for the heart failure data.

Keywords: Data Mining, Prediction, Classification Models, Heart Failure, Clinical Medicine

1. Introduction

In clinical diagnosis, etiology is very important for the cure and prevention of disease. Due to the trend of population aging, heart failure has gradually become an important cause of global mortality increase. It was found that fifty to seventy-five percent of heart failure patients die within five years of diagnosis [4].

Therefore, it is beneficial for understanding the risk factors of heart failure, which helps people to better reduce their risk of developing heart failure. This study will use data mining and machine learning to help predict mortality of patients with heart failure. Ledley & Lusted introduced mathematical models into clinical medicine for the first time and proposed mathematical models of computer-aided diagnosis [9]. Since the 1990s, rapid advances in computing technology, in particular, machine learning methods have made the computer-aided diagnosis more available in clinical medicine, and more and more powerful predictive models have helped medical workers understand and cure diseases.

This study aims to compare the effects among various individual machine learning methods and ensemble learning methods in the prediction of heart failure, so that more appropriate classifiers can be used to predict the probability of heart failure and improve the accuracy of clinical diagnosis of heart failure.

In this study, firstly, the features that are not related to the prediction target are deleted. Then the data is normalized into the appropriate ranges of features. Next, the data is divided into training set and test set with a ratio of 7 to 3. Further, the multiple base classifiers including K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes Model (NBM), Gradient Boosting Decision Tree (GBDT), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and Extreme Gradient Boosting (EGB) are trained and their performances are compared. Finally, the Stacking-based ensemble learning algorithm is employed to select the competent base classifiers

and integrate them into a final ensemble model. Accuracy, precision, AUC, Balanced accuracy and F1-score are used to evaluate the model performance.

The remainder of this study is organized as follows. Section 2 introduces the related work of data mining technology and clinical heart failure. Section 3 explores the method of data preprocessing and modeling. In Section 4, the experimental results are analyzed. Section 5 presents the conclusion of this study and the future work.

2. Related Work

With the wide application of information technology, data mining technology has been paid great attention to in various fields. In the medical field, data mining can analyze and compare the physical signs and biological data of the human body, dig out the correlation, and analyze the precursor characteristics of diseases, to achieve the purpose of prevention or timely treatment.

Data mining has made great contributions in the areas of breast cancer, diabetes, kidney disease, lung cancer, gene association and so on [10]. However, most of data mining efforts in the medical field use a single data mining model and do not consider the different priorities required for different data sets. At the same time, the accuracy, adaptability, and robustness of the models have not been taken into consideration in a comprehensive manner.

In the field of heart failure, Ahmad et al. studied the heavy correlation between various factors and the incidence of heart failure [1]. Chicco & Jurman predicted the survival rate of patients with heart failure only by serum creatinine and ejection fraction [5].

Sohrabi et al. deployed classification algorithms (i.e., DT, Artificial Neural Networks (ANN), SVM and LR), and used AUC and accuracy as evaluation indicators to reduce costs and improve the quality of treatment in the hospital system [14]. To predict high frequency risk, decision support system based on ANN has also been used. Lafta et al. trained the neural network classifier with the global weight of attribute contribution to predict heart failure risk of patients [8]. The results showed that the method could accurately predict the clinical risk of heart failure.

Poolsawad et al. claimed that despite the large size of the clinical data set, missing value filling methods did not affect the data mining performance [11]. It is critical that the data set is an accurate representation of the clinical problem. Those methods that fill in the missing values do not affect the development of classifiers and prognostic/diagnostic models significantly. Supervised learning has been shown to be more suitable for mining clinical data than unsupervised methods. Rammal and Emam proposed that non-parametric classifiers such as decision trees give better results than parametric classifiers such as radial basis function networks (RBFNs) [12]. Rammal and Emam also explored the current analytical techniques that support prediction of heart failure, and then used the WEKA analytical tool to build an integrated data mining model based on big data technologies [12].

Hehde et al. established population-specific hematology reference intervals via data mining [6]. Shironoet al. used decision tree analysis to identify profiles associated with disease control rate (DCR) and the prognosis of patients with unresectable intrahepatic cholangiocarcinoma [13].

Augustine et al. employed LR to analyze the factors associated with high blood pressure and heart issues [3]. Animut & Berhanu explored the determinants of anemia status among pregnant women in ethiopia with LR [2]. Hu et al. employed LR to predict the mortality risk of acute respiratory distress syndrome [7].

However, at present, there is still not a mature classification model to analyze the heart failure data, and the predictive accuracy of the model is still a topic worth studying. Therefore, the purpose of this study is to explore the effectiveness of prediction model for heart failure data and suggest relevant personnel to use data mining technology for predictive analytics.

3. Methodology

3.1. Data Understanding and Preprocessing

The dataset was originally collected by Ahmad et al [1]. The data set including 299 patients (105 women and 194 men) with heart failure were published in July 2017, and all patients are over 40 years old. As shown in Table 1, there are 13 clinical features, which are age, anaemia, creatinine phosphokinase (CPK), diabetes, ejection fraction (EF), high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event. Features are independent against each other, and the last feature, death event, is the classification target of the dataset, which is represented by 0 and 1. Here, 1 indicates that the patient died during follow-up period, and 0 indicates that the patient did not die during follow-up period. Follow-up period ranges from 4 to 285 days, and is 130 days in average. Age, serum sodium, and CPK are continuous variables, while EF, serum creatinine, and platelets are categorical variables. EF is divided into three levels ($EF \leq 30$, $30 < EF \leq 45$ and $EF > 45$) and platelets are also classified into three levels based on quartiles. Serum creatinine higher than normal (1.5) is an indicator of renal dysfunction. Patients are evaluated for anemia based on their blood pressure levels [1].

Table 1. *Data description.*

Features	Description
Age	Age of the patient (years)
Anaemia	Decrease of red blood cells or hemoglobin (boolean)
Creatinine phosphokinase	Level of the CPK enzyme in the blood (mcg/L)
Diabetes	The patient has diabetes or not (boolean)
Ejection fraction	Percentage of blood leaving the heart at each contraction (percentage)

Features	Description
High blood pressure	The patient has hypertension or not (boolean)
Platelets	Platelets in the blood (kiloplatelets/mL)
Serum creatinine	Level of serum creatinine in the blood (mg/dL)
Serum sodium	Level of serum sodium in the blood (mEq/L)
Sex	Woman or man (binary)
Smoking	The patient smokes or not (boolean)
Time	Follow-up period (days)
Death event [target]	The patient deceases during the follow-up period or not (boolean)

Because the follow-up period has no causal relationship with the death of patients with heart failure, the Time feature is removed. The objective of the study is to evaluate the rate of death within 285 days.

Rest features like age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium are normalized by MinMaxScaler [15]. It scales the data into between 0 and 1 according to the below Equation (1).

$$X(\text{normalized})=\frac{x-\min(x)}{\max(x)-\min(x)} \quad (1)$$

By making the features have the same measurement scale and unifying their statistical probability distribution between 0 and 1, the construction of machine learning models can be simplified, the efficiency of machine learning can be improved, and the prediction accuracy can be improved.

3.2. Modeling Method

In this study, data is first randomly divided into training set and test set in a 7:3 ratio, namely 70% as training data and 30% as test data.

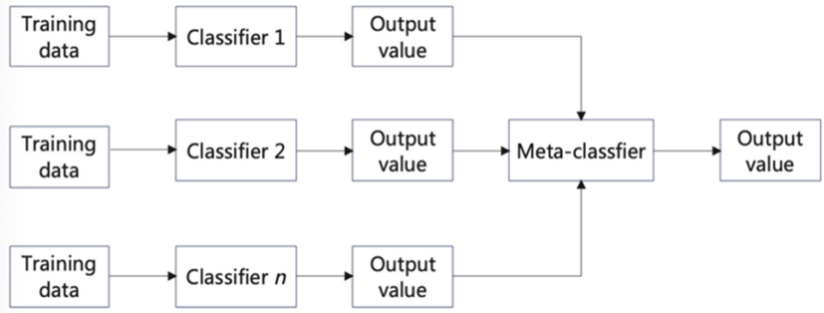

Figure 1. *Operation principle of Stacking.*

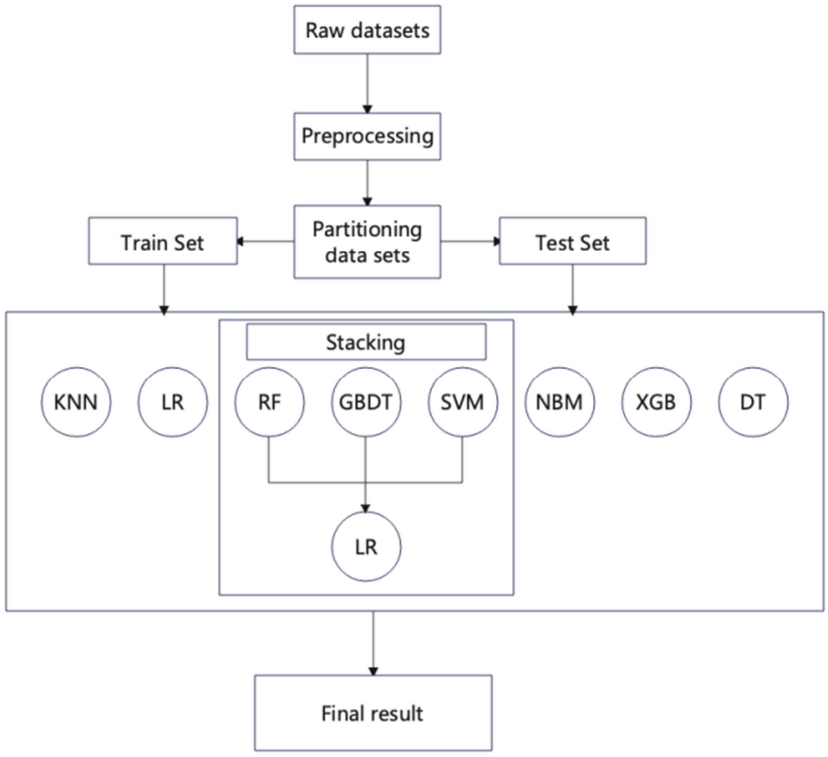

Figure 2. *Process of the proposed model.*

As shown in Figure 1, a Stacking-based ensemble learning model is employed to analyze the data through a multi-layer learning structure. The first layer (also called learning layer) uses *n* different classifiers or models with different parameters to combine the predicted results into a new feature set, which is used as the input of the next layer of classifier (i.e., meta-classifier).

As shown in Figure 2, KNN, LR, NBM, GBDT, SVM, RF DT and XGB are respectively used as base classifiers to predict the death of patients with heart failure, and then the three competent base classifiers (GBDT, RF and SVM) with the higher predictive accuracy are selected to construct the Stacking-based ensemble model, with LR used as the meta-classifier.

LR applies Maximum Likelihood Estimate (MLE) method to calculate the risk factors of disease, and predicts the probability of occurrence of disease according to the risk factors.

The training set is used to train the base classifiers to fit the model better. Then the evaluation index of eight base classifiers is obtained and compared through the test set. Similarly, the evaluation indicators of the Stacking-based integration model can be obtained.

4. Experimental Results

In this experiment, KNN, LR, NBM, GBDT, SVM, RF, DT and XGB are used as base classifiers and trained respectively. After comparison, GBDT, RF, and SVM were found to outperform others and then selected as competent base classifiers to form the Stacking integration model, with LR used as the meta-classifier. Evaluation indicators like accuracy, precision, AUC, Balanced accuracy and F1-score were used to evaluate the model performance. All classification models were implemented using Python programming language.

The confusion matrix is shown in Table 2. Calculation formulas of accuracy, precision, AUC, recall, Balanced accuracy and F1-score are shown in Equations (2) ~ (6).

Table 2. *Confusion Matrix.*

		True condition	
		positive	**negative**
Predicted condition	positive	True Positive (TP)	False Positive (FP)
	negative	False Negative (FN)	True Negative (TN)

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (4)$$

$$\text{F1-score}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (5)$$

$$\text{Balanced Accuracy}=\frac{TPR+TNR}{2} \quad (6)$$

Accuracy represents the proportion of samples which are correctly classified in the total number of samples; Precision represents the proportion of samples that are truly positive from the samples predicted positively; Recall represents the proportion of samples that are correctly predicted positively from the samples predicted positively. In practice, when Precision is high, Recall tends to be low, and vice versa. So, the harmonic average of Precision and Recall F1-score is also utilized. Accuracy sometimes does not perform better in imbalanced data sets, so Balanced accuracy is also utilized. AUC is defined as the area under the Receiver Operating Characteristic (ROC) curve and used as the evaluation standard, because the ROC curve cannot clearly indicate which classifier has the better effect in many cases. ROC is based on a series of different dichotomies (boundary value or determination threshold), with true positive rate as the Y-axis and false positive rate as X-axis.

To verify the performance of the classifiers fairly, the different random segmentation to training and test data are conducted. Then the statistics are tabulated, and ROC curves of various classification models are obtained, as shown in Figure 3 and Table 3.

Table 3. *The detailed information about evaluation indicators of various classification models.*

Classifier	**Accuracy**	**Precision**	**AUC**	**Recall**	**Balanced Accuracy**	**F1-score**
KNN	0.6000	0.5111	0.4551	0.6000	0.4551	0.5429
LG	0.7111	0.6849	0.5847	0.7111	0.5847	0.6698
NBM	0.7000	0.6667	0.5668	0.7000	0.5668	0.6527
DT	0.6778	0.6573	0.5899	0.6778	0.5899	0.6634
SVM	0.6889	0.4746	0.5000	0.6889	0.5000	0.5620
GBDT	0.7000	0.6971	0.6452	0.7000	0.6452	0.6985
RF	0.6778	0.6573	0.5899	0.6778	0.5899	0.6634
XGB	0.6556	0.6391	0.5737	0.6556	0.5737	0.6453
Stacking Classifier	0.7556	0.7450	0.6855	0.7556	0.6855	0.7465

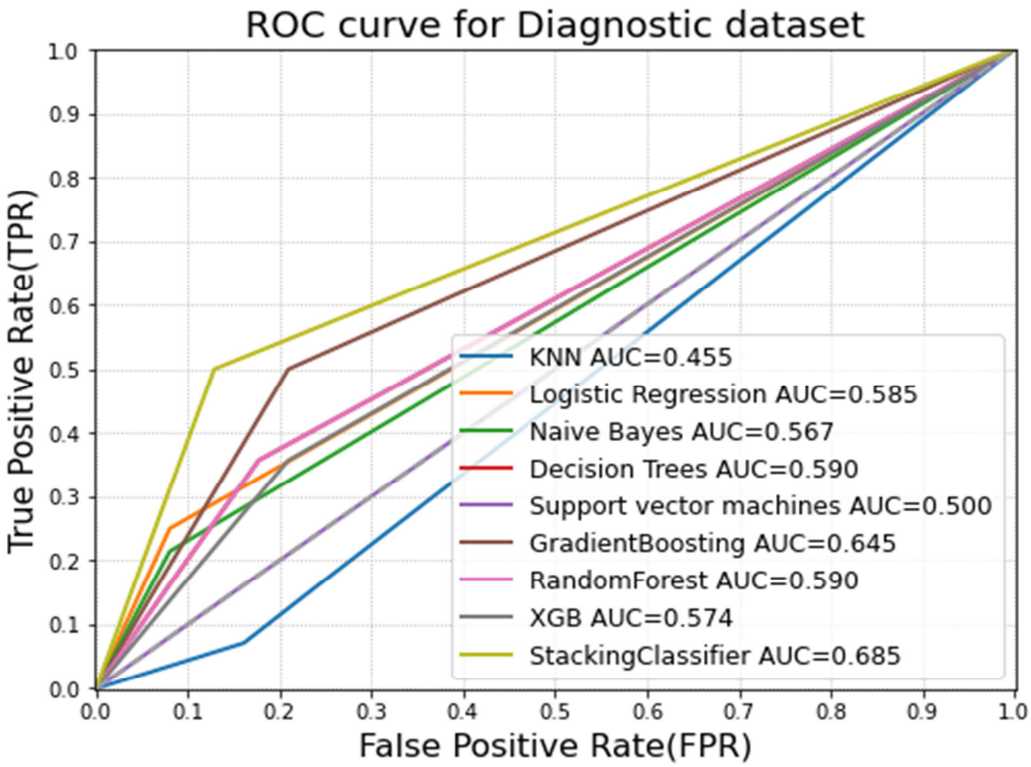

Figure 3. *ROC curves for diagnostic dataset.*

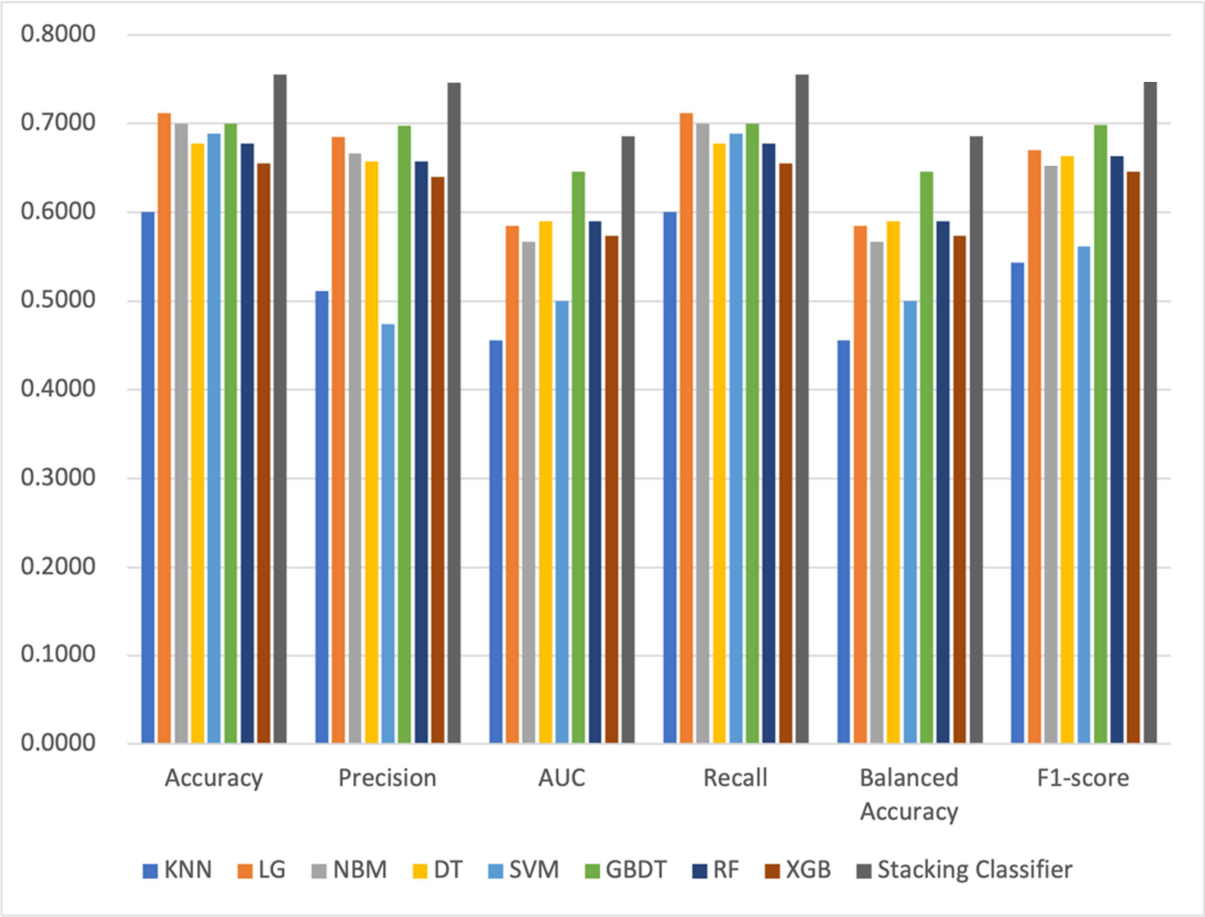

Figure 4. *Histogram of evaluation indicators of classification models.*

According to Figure 4, all the evaluation indicators, show that the Stacking-based ensemble model perform better than the base classifiers, therefore, the ensemble model should be selected as the best model for predicting this data set.

5. Conclusion

At present, with the change of lifestyle, the incidence of heart failure is increasing year by year. It is urgent to study the pathogenic factors and treatment. In this study, features which are not associated with heart failure are first removed, and then the data are preprocessed into the appropriate ranges through normalization to unify their statistical probability distribution. KNN, LR, NBM, GBDT, SVM, RF, DT and XGB are used as the base classifiers, and GBDT, RF, and SVM are used to construct the Stacking-based ensemble learning model. In addition, Accuracy, precision, AUC, Balanced accuracy and F1-score are used to evaluate the model performance. Finally, it is concluded that the Stacking integrated model performs better than the base classifiers. Therefore, the results of this study provide reference for clinical judgment of heart failure and prevention of disease occurrence.

There are still some shortcomings in the methods used in this study. As this data set is small and has no missing values, and the data is relatively balanced, the prediction is relatively simple, and the preprocessing is also convenient. However, the clinical medical data in practice tend to be large and contain more missing values and outliers. The practical data can also be imbalanced. Therefore, in the processing of practical clinical data, the missing values should be paid additional attention to, and the imbalanced data should also be handled in the preprocessing stage. There will be also much more features with high dimensionality, the importance of each feature should be checked (e.g., through correlation matrix analysis), and only salient features are chosen to predict the disease. Data mining for heart failure still needs more in-depth research.

References

[1] Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PloS One*, 12 (7), e0181001.

[2] Animut, K., & Berhanu, G. (2022). Determinants of anemia status among pregnant women in ethiopia: using 2016 ethiopian demographic and health survey data; application of ordinal logistic regression models. *BMC Pregnancy and Childbirth*, doi: 10.1186/S12884-022-04990-8.

[3] Augustine, K. A., Pascal, K., K., Faustina, A., et al. (2022). A binary logistic regression analysis on the factors associated with high blood pressure and its related heart issues. *Science Journal of Applied Mathematics and Statistics*, doi: 10.11648/J.SJAMS.20221003.12.

[4] Bleumink, G. S., Knetsch, A. M., Sturkenboom, M. C., et al. (2004). Quantifying the heart failure epidemic: prevalence, incidence rate, lifetime risk and prognosis of heart failure: the Rotterdam study. *European Heart Journal*, 25 (18), 1614-1619.

[5] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20 (1), 1-16.

[6] Hehde, J., Olschinsky-Szermer, M., Pahl, J., et al. (2022). Indirectly determined hematology reference intervals for pediatric patients in Berlin and Brandenburg. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 60 (3), 408-432.

[7] Hu, J., Fei, Y., & Li, W. Q. (2021). Predicting the mortality risk of acute respiratory distress syndrome: radial basis function artificial neural network model versus logistic regression model. *Journal of clinical monitoring and computing*, doi: 10.1007/S10877-021-00716-X.

[8] Lafta, R., Zhang, J., Tao, X., Li, Y., Tseng, V. S., Luo, Y., & Chen, F. (2016). An intelligent recommender system based on predictive analysis in telehealthcare environment. *Web Intelligences*, 4 (4), 325-336.

[9] Ledley, R. S. & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130 (3366), 9-21.

[10] McCullough, P. A., Jurkovitz, C. T., Pergola, P. E., et al. (2007). Independent components of chronic kidney disease as a cardiovascular risk state: results from the Kidney Early Evaluation Program (KEEP). *Archives of Internal Medicine*, 167 (11), 1122-1129.

[11] Poolsawad, N., Moore, L., Kambhampati, C., & Cleland, J. G. (2014). Issues in the mining of heart failure datasets. *International Journal of Automation and Computing*, 11, 162-179.

[12] Rammal, H. F., & Emam, A. Z. (2018). Heart failure prediction models using big data techniques. *International Journal of Advanced Computer Science and Applications*, 9 (5), doi: 10.14569/IJACSA.2018.090547.

[13] Shirono, T., Niizeki, T., Iwamoto, H., et al. (2021). Therapeutic outcomes and prognostic factors of unresectable intrahepatic cholangiocarcinoma: a data mining analysis. *Journal of Clinical Medicine,* 10 (5), 987.

[14] Sohrabi, B., Vanani, I. R., Gooyavar, A., & Naderi, N. (2019). Predicting the readmission of heart failure patients through data analytics. *Journal of Information & Knowledge Management*, 18 (01), 1950012.

[15] Tran, M. T., & Lee, G. S. (2019). Super-resolution in music score images by instance normalization. *Smart Media Journal*, 8 (4), 64-71.