

Classification of individuals at risk of heart disease using machine learning

Makine öğrenmesi kullanarak kalp hastalığı riski olan bireylerin sınıflandırılması

Betül Akalın¹, Ülkü Veranyurt¹, Ozan Veranyurt²

¹ Sağlık Bilimleri Üniversitesi, İstanbul, Türkiye

² Bahçeşehir Üniversitesi, İstanbul, Türkiye

Corresponding author: Ülkü Veranyurt, PhD, Sağlık Bilimleri Üniversitesi, İstanbul, Türkiye

E-mail: ulkuveranyurt@gmail.com

Received/Accepted: May 24, 2020 / August 17, 2020

Conflict of interest: There is not a conflict of interest.

SUMMARY




Objective: The aim of this study is to determine whether people have heart disease by using different machine learning algorithms with the data provided by the University of Cleveland.

Method: 303 patient data provided by the University of Cleveland were classified using Gaussian Bayes, K-Nearest Neighbor and Random Forest Algorithms with and without feature scaling. With each algorithm, the data is divided into random training and test sets. This process was repeated 50 times for each algorithm. The test results were subjected to the T-test to check statistical independence.

Results: In this study, 80.52% accuracy with K-Nearest Neighbor algorithm, 80.52% with Gaussian Bayes and 82.50% with Random Forest were observed with data scaling. The results of the three algorithms produced similar values and did not show statistical independence ($p > 0.05$). Without data scaling, 65.28% accuracy with the K-Nearest Neighbor algorithm, 80.52% with Gaussian Bayes and 82.19% with Random Forest were observed. The test results obtained with three algorithms showed statistical independence.

Conclusions: Although there were data from 303 patients in the study, over 80% accurate prediction was obtained. The presence of endpoints that distort the distribution in the data used results in differences in the methods used. It has been confirmed that much closer estimates can be obtained on a scaled patient data. This study is an example of the use of artificial intelligence in detecting cardiac diseases that pose a risk all over the world. With a more detailed patient data, much higher accuracy rates can be obtained and included in health management processes in the pre-diagnosis of heart disease in the future.

Keywords: Illness classification, machine learning in health, heart disease, heart failure

 Betül Akalın
 Ülkü Veranyurt
 Ozan Veranyurt

ORCID IDs of the authors:
B.A. 0000-0003-0402-2461
Ü.V. 0000-0003-4838-3373
O.V. 0000-0003-3652-2356

ÖZET

Amaç: Bu çalışmanın amacı Cleveland Üniversitesi tarafından sağlanan veriler ile farklı makine öğrenmesi algoritmaları kullanarak kişilerin kalp hastalığının olup olmadığını tespit etmektir.

Yöntem: Cleveland Üniversitesi tarafından sağlanan 303 kişilik hasta verisi özellik ölçekleme ile ve ölçekleme olmaksızın Gaussian Bayes, K-Nearest Neighbour ve Random Forest Algoritmaları kullanılarak sınıflandırılmıştır. Her

bir algoritma ile veri rastgele eğitim ve test kümelerine bölünmüştür. Bu işlem her bir algoritma için 50 kez tekrar edilmiştir. Test sonuçları istatistiksel bağımsızlığı kontrol etmek için T-testine tabi tutulmuştur.

Bulgular: Yapılan çalışmada veri ölçeklendirilmesi ile K-Nearest Neighbour algoritması ile %80.52, Gaussian Bayes ile %80.52 ve Random Forest ile %82.50 doğruluk gözlemlenmiştir. Kullanılan üç algoritmanın sonuçları birbirine benzer değerler üretmiş ve istatistiksel olarak bağımsızlık göstermemiştir ($p > 0.05$). Veri ölçeklendirmesi olmadan ise K-Nearest Neighbour algoritması ile %65.28, Gaussian Bayes ile %80.52 ve Random Forest ile %82.19 doğruluk gözlemlenmiştir. Üç algoritma ile elde edilen test sonuçları istatistiksel olarak bağımsızlık göstermiştir.

Sonuç: Çalışmada 303 hastanın verisi olmasına rağmen %80 üzerinde doğru tahminleme elde edilmiştir. Kullanılan veride dağılımı bozan uç noktaların olması kullanılan yöntemlerde sonuç farklarına sebep olmaktadır. Ölçeklendirilmiş bir hasta verisi üzerinde çok daha yakın tahminler elde edilebildiği doğrulanmıştır. Bu çalışma tüm dünyada risk teşkil eden kalp hastalıklarının tespit yapay zekâ kullanımında bir örnek teşkil etmektedir. Daha detaylı bir hasta verisi ile çok daha yüksek doğruluk oranları elde edilebilir ve gelecekte kalp hastalığının ön teşhisinde sağlık yönetimi süreçlerine dâhil edilebilir.

Anahtar sözcükler: Hastalık sınıflandırması, sağlıkta makine öğrenmesi, kalp hastalığı, kalp yetmezliği

INTRODUCTION

Death in heart diseases is one of the most common causes of death in developing countries. The points that make the diagnosis of heart diseases the most difficult are myocardial perfusion single-photon emission computed tomography (SPECT), and electrocardiogram (ECG) can be diagnosed by interpreting¹. The experience of the specialist who examines the medical diagnosis plays an important role. At this point, machine learning, which is a sub-branch of artificial intelligence, can learn similarities in the images, or the determining features in the diagnosis on patient information².

Unlike other areas of technology, applications of artificial intelligence and machine learning on health are still developing³. Its applications in the field of cardiology are very limited⁴. The first uses of artificial intelligence were basic estimation and image analysis. Developing hardware technology has started to increase in health as it allows working on big data⁵.

Machine learning⁶ works with iterations, it tries to learn common patterns on data without any assumptions. Machine learning types are summarized in Table 1.

Table 1: Types of machine learning

No	Method	Definition
1	Supervised	It is the most common form of learning. The data is divided into training and test sets. The data is marked in advance according to the procedure to be performed. It is used in operations such as regression, estimation, classification. Algorithms such as artificial neural networks, Random Forest are examples.
2	Unsupervised	In this form of learning, the model is not given any class or numerical information for education. The model tries to produce results through common points in the data. K-Nearest Neighbor (KNN), hierarchical clustering are examples.
3	Semi-supervised	It is divided into sections with or without result information marked in the data. Data that is not fully classified is used. Sound perception can be given as an example.
4	Reinforcement	It is based on the reward principle in behavioral psychology. The decision making mechanism learns the option that gives the highest reward based on the result obtained. Today, it is used in areas such as robotics, personalization in artificial intelligence, medical image processing.

In this study, Random Forest, Gaussian Bayes and unsupervised algorithm were chosen as supervised.

Random Forest: It is a model that uses decision trees. The data set is divided into decision breaks according to the introphy of the features. Decision

trees are created as much as the number of hyper parameters given. It is based on a certain number of features in each wood data set. The obtained trees are combined according to their accuracy and differences from each other and the results are obtained⁷. The biggest problem in models using decision trees is that if the data is low, it creates an over fitting⁸.

Gaussian Bayes: It is a supervised model. Probability theory calculates probabilities for each column of data. For the numerical data, the gaussian distribution is checked. Probability results are calculated for the class information to which each row of data belongs. The high probability class is considered to be the result⁹. This algorithm accepts each column of data independently. Therefore, it does not take into account the correlation in the data¹⁰.

K-Nearest Neighbor (KNN): It is an unsupervised algorithm. Common properties are determined without making a marking on the data¹¹. The basic parameter for this algorithm is the neighbor number of the selected data. The number of neighbors is given as the hyper parameter. Accordingly, the nearest n neighbors are

determined. The class of the data selected according to the class of the neighbors is determined. In the measurement of distance between data, distance criteria such as euclid and Minskovski are used¹⁰.

MATERIAL AND METHODS

This study was carried out on a database of 303 people with and without risk of heart disease provided by the University of Cleveland. The dataset consists of 13 different columns. In this dataset, respectively; age, gender, chest pain type, resting blood pressure, serum cholesterol level, fasting blood glucose, blood pressure when applying to the hospital, maximum heart rate limiting electrocardiographic results, exercise-induced angina, oldpeakST according to rest, individuals with or without disease. For coding, whether the person is sick or not, the number of individuals with the number of main vessels (0-3) colored with fluoroscopy were included in the study (Table 2). The dataset was prepared by a team which included cardiologists.

Table 2: Column specifications

Column No	Column Name	Column Specification	Values
1	age	age	in years
2	sex	sex	1= male; 0= female
3	cp	chest pain type	0-3
4	trestbps	resting blood pressure	in mm Hg on admission to the hospital
5	chol	serum cholestorol	in mg/dl
6	fb	fasting blood sugar > 120 mg/dl	1= true; 0= false
7	restecg	resting electrocardiographic results	0-2
8	thalach	maximum heart rate achieved	71-202 bps
9	exang	exercise induced angina	1 = yes; 0 = no
10	oldpeak	ST depression induced by exercise relative to rest	0-6.2
11	slope	the slope of the peak exercise ST segment	0-2
12	ca	number of major vessels (0-3) colored by flourosopy	0-4
13	thal	previous heart defect type	3= normal; 6= fixed defect; 7= reversable defect

Correlation relationship between data columns is presented in Figure 1. According to the color scale, the correlation relationship between the columns that are close to yellow is high, while the correlation relationship is low in those with dark color. Looking at Figure 1 slope, cp, thalac

columns show the highest correlation with the target column. There is no correlation between the target column and any other column above ± 0.5 . This shows that no column can be used independently in prediction.

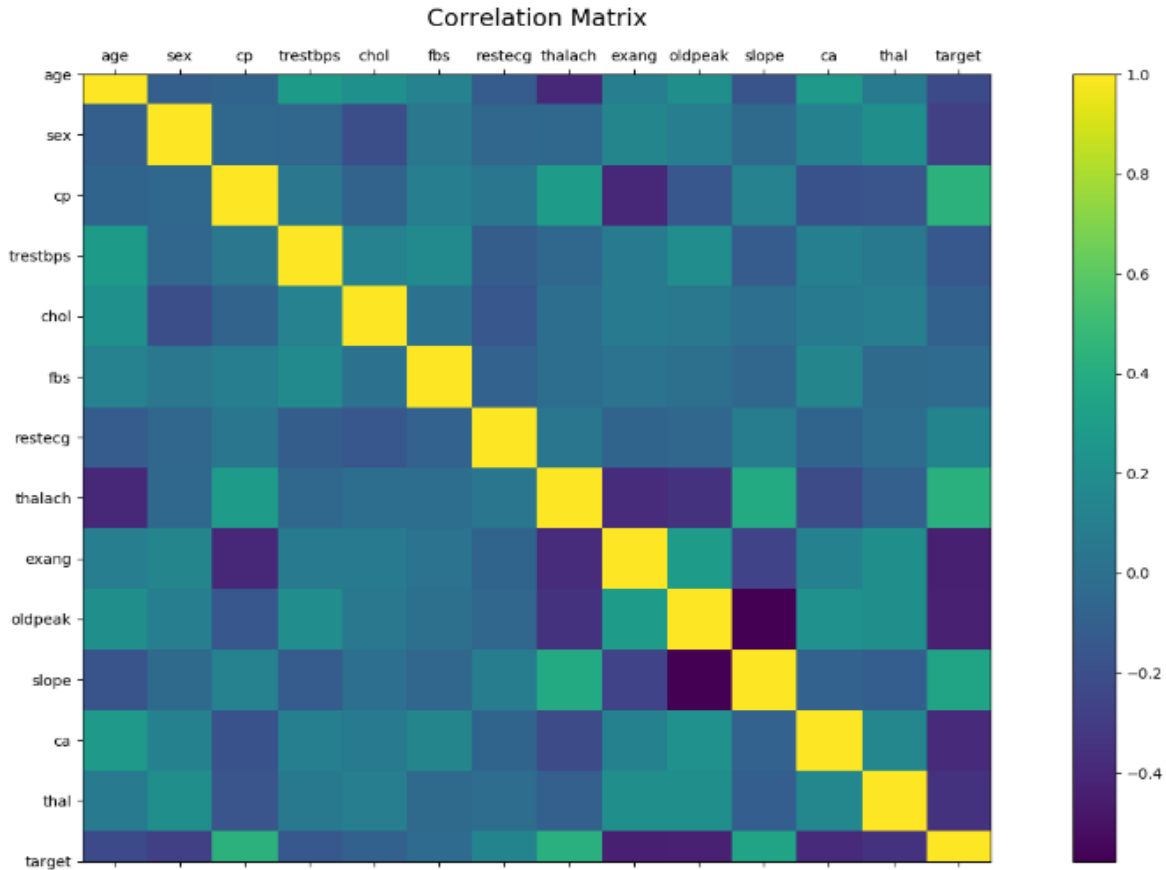


Figure 1: Correlation Matrix

The relationship between chest pain and heart disease is shown in Figure 2. As shown in the intersection graphs, the data does not show complete partitioning and there is no clear correlation.

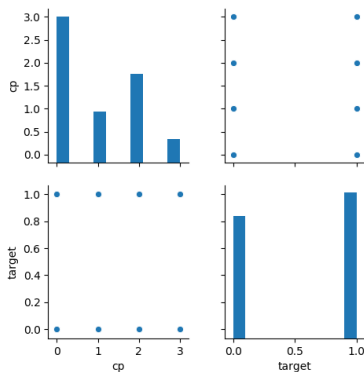


Figure 2 : Cp (chest pain) Corelation Graphic

After the data review; we applied Gaussian Bayes, K-Nearest Neighbor and Random Forest algorithms on the dataset. Supervised learning was performed for each algorithm by dividing the data set into training and test subsets and evaluation was performed for 50 times for each algorithm, and average accuracy values were taken into consideration. For each training and test data was randomly split. When this process was applied on the data without scaling, the results showed independence ($p < 0.05$). In the experiment with data scaling, similar results were obtained due to the small data set and no extreme data ($p > 0.05$). T tests were applied on the result sets of each algorithm.

RESULTS

In the study, 80.52% accuracy with K-Nearest Neighbor algorithm (KNN), 80.52% with Gaussian Bayes and 82.50% with Random Forest were observed with data scaling. The results of the three algorithms produced similar values and did not show statistical independence ($p > 0.05$). Without data scaling, 65.28% accuracy with the K-Nearest Neighbor algorithm, 80.52% with Gaussian Bayes

and 82.19% with Random Forest were observed. The test results obtained with three algorithms showed statistical independence. As a result of the data segmentation process performed 50 times in Table 3, percentage accuracy of each algorithm on the test data is given. In scaled data, 3 algorithms produced 81-83% results that were not statistically independent. Parametric (Random Forest,

Gaussian Bayes) and nonparametric (KNN) algorithms produced similar results in patient data without extreme data. The similarity of the results may be due to the number of data.

The Gaussian Bayes algorithm, which is a working principle based on the independence of data features, may produce similar results in scaled data.

Table 3 : Algorithm test accuracies

Algorithm	Accuracy/Data Scaled	Accuracy/Data Unscaled
KNN	80.52%	65%
Random Forest	82.52%	82.19%
Gaussian Bayes	80.52%	80.52%

If we look at the one-time studies of the tested algorithms on unscaled data; The KNN algorithm correctly predicted 61 of 91 test data as shown in figure 2. However, there are 13 people who are not predicted to have heart disease, and 17 people who are considered to be not heart patients. For this test, the number of neighbors is determined as 5 as the hyper parameter. A decrease in data accuracy was observed at lower values (Figure 3).

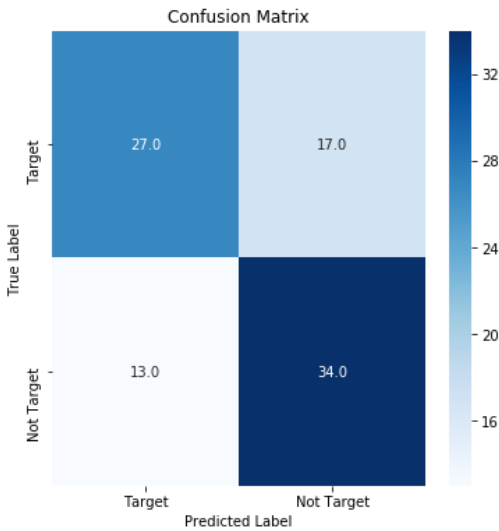


Figure 3: KNN Confusion Matrix

When Random Forest algorithm is applied on unscaled data, accurate estimation was made for 76 people in 91 test data. Four people were estimated to be patients when they were not heart patients, and 11 were classified as healthy, although they were heart patients. For the Random Forest algorithm, 100 sub-decision trees were used as hyper parameters (Figure 4).

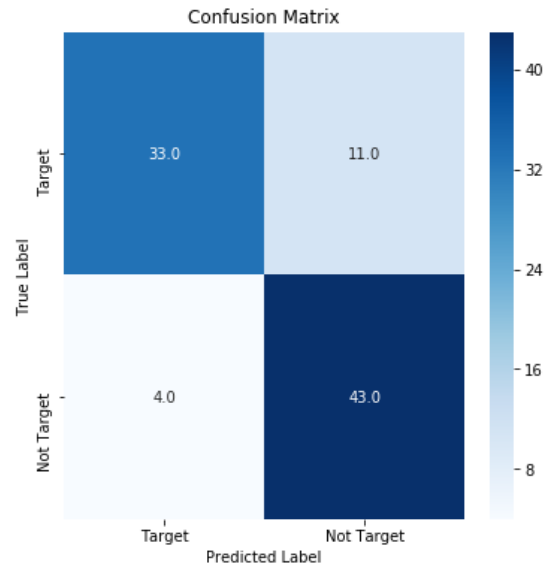


Figure 4: Random Forest Confusion Matrix

In the application made with the Gaussian Bayes algorithm, 73 people were correctly classified in the test set of 91 people. Although 6 people did not have heart disease, they were classified as patients. Although 12 people have heart disease, they are classified as healthy (Figure 5).

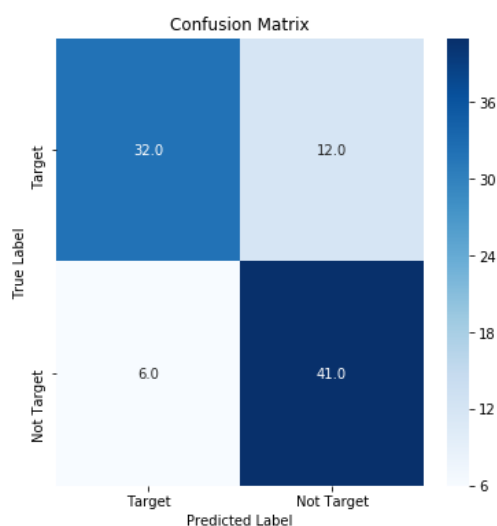


Figure 5: Gaussian Bayes Confusion matrix

In the one-time test application, the highest accuracy rate was provided by the Random Forest algorithm. In terms of false diagnosis rate, the lowest error rate was again provided by the Random Forest algorithm. Having non-standard samples in the patient data decreased the performance of the KNN algorithm. This shows that the KNN algorithm should not be applied on health data without scaling.

CONCLUSION

Although 303 patients have data in the study, over 80% accurate estimation was obtained. The presence of endpoints that distort the distribution in the data used results in differences in the methods used. It has been confirmed that much closer estimates can be obtained on a scaled patient data. Heart diseases and their derivatives are known as the most complex diseases in the medical literature¹². 3.6 million People are heart patients in north-western Europe¹³ and it is projected to increase to 5 million by 2025. While advanced heart diseases create an annual cost of about 10,000 euros per patient, the lifetime cost is over 100,000 euros¹⁴. In its current form, it is not possible to manage the diagnosis and diagnosis in the health sector only with human power¹⁵. The rapid development of modern technology, especially the use of artificial intelligence, will change the health sector¹⁵. The aging population, the more complex of the diseases seen and the increase in the population in the rural areas require artificial intelligence-supported health processes that are not manpowered¹⁶

The use of artificial intelligence in clinical management, management of heart attacks and other heart diseases is increasing day by day¹⁷. As a result of the recent studies, it has been seen that

the reliability of artificial intelligence applications is related to the quality of the data used. Since this study has been studied with a limited number of data, an approximately 80% success rate in diagnostic accuracy has been achieved. With a richer clinical data, the accuracy rate in this study can be increased, and artificial intelligence applications can be used to detect heart attacks and different heart diseases¹⁸. Algorithmic solutions can be used for diagnosis in the field of health, and uses can be increased in areas such as diagnosis, treatment, evaluation of visual test results, and radiology¹⁷. This study is an example of an algorithmic solution on clinical data. According to a study conducted in 2019, studies on artificial intelligence applications in the field of health have increased 3 times compared to the last 3 years¹⁸. If we give a few examples of artificial intelligence applications in the field of health; detection of risk of cardiac arrest due to eco-cardiography¹⁹, detection of heart disease using tomographic images²⁰. Detection of coronary atherosclerosis using a biomarker¹⁵.

There are difficulties such as medical prolongation faced by the use of artificial intelligence in health¹⁷. Some of these difficulties are; the accuracy of the clinical data used, the reliability of the model, the management of the data, the legal processes related to the data used, testing and verification of the model²¹. In order for artificial intelligence models to be functional, patient databases containing high-size and independent data are needed¹⁷. While keeping the confidentiality of the patient data used constitutes a serious problem, ethical processes differ in each country¹⁷. The reliability of the model and tests can be achieved by using multiple validation or different data sets²².

In this study, multiple validation was applied on the same data set. Increasing the quality of the data used or applying to machine learning while selecting the data set are alternative methods that can be used to increase the reliability of the results²². This study is an example of the use of artificial intelligence in detecting cardiac diseases that pose a risk all over the world. With a more detailed patient data, much higher accuracy rates can be obtained and included in health management processes in the pre-diagnosis of heart disease in the future.

REFERENCES

1. Box LC, Angiolillo DJ, Suzuki N, Box LA, Jian J, Guzman L, et al. Heterogeneity of atherosclerotic plaque characteristics in human

- coronary artery disease: A three-dimensional intravascular ultrasound study. *Catheter Cardiovasc Interv* 2007;70(3):349-56. <https://doi.org/10.1002/ccd.21088>.
2. Alonso DH, Wernick MN, Yang Y, Germano G, Berman DS, Slmoka P. Prediction of cardiac death after adenosine myocardial perfusion SPECT based on machine learning. *J Nucl Cardiol* 2018. <https://doi.org/10.1007/s12350-017-0924-x>.
 3. Narula S, Shameer K, Salem Omar AM, Dudley JT, Sengupta PP. Reply: Deep learning with unsupervised feature in echocardiographic imaging. *J Am Coll Cardiol* 2017;69:2101-2.
 4. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2017 May 6.
 5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115-8.
 6. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: Are we there yet? *Heart* 2018. <https://doi.org/10.1136/heartjnl-2017-311198>.
 7. Weng SF, Reys J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;12:e0174944.
 8. C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Cambridge, Mass.: MIT Press, 2006.
 9. M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, pp. 111-147, 1974.
 10. T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, pp. 21-27, 1967.
 11. Savarese G, Lund LH. Global public health burden of heart failure. *Cardiac Fail Rev*. 2017;3(1):7-11. <https://doi.org/10.15420/cfr.2016:25:2>.
 12. Braunschweig F, Cowie MR, Auricchio A. What are the costs of heart failure? *Europace*. 2011;13(Suppl 2):ii13-7. <https://doi.org/10.1093/europace/eur081>.
 13. Sanders-van Wijk S, van Asselt AD, Rickli H, Estlinbaum W, Erne P, Rickenbacher P, et al. Cost-effectiveness of N-terminal pro-B-type natriuretic-guided therapy in elderly heart failure patients: results from TIME-CHF (Trial of Intensified versus Standard Medical Therapy in Elderly Patients with Congestive Heart Failure). *JACC Heart Fail*. 2013;1(1):64-71. <https://doi.org/10.1016/j.jchf.2012.08.002>.
 14. Conrad N, Judge A, Tran J, Mohseni H, Hedgecott D, Crespillo AP, et al. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet*. 2018;391(10120):572-80. [https://doi.org/10.1016/s0140-6736\(17\)32520-5](https://doi.org/10.1016/s0140-6736(17)32520-5).
 15. KrumH, Forbes A, Yallop J, Driscoll A, Croucher J, Chan B, et al. Telephone support to rural and remote patients with heart failure: the Chronic Heart Failure Assessment by Telephone (CHAT) study. *Cardiovasc Ther*. 2013;31(4):230-7. <https://doi.org/10.1111/17555922.12009>.
 16. Tran BX, Latkin C.A, Giang V.T, et al. The Current Research Landscape of the Application of Artificial Intelligence in Managing Cerebrovascula and Heart Diseases: A Bibliometric and Content Analysis. *Int. J. Environ. Res. Public Health* 2019;16:2699. <https://doi:10.3390/ijerph16152699>.
 17. Pakhomov, S.S.; Hemingway, H.; Weston, S.A.; Jacobsen, S.J.; Rodeheer, R.; Roger, V.L. Epidemiology of angina pectoris: Role of natural language processing of the medical record. *Am. Heart J*. 2007, 153, 666-673.
 18. Kwon, J.M. Kim, K.H. Jeon, K.H.; Park, J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography* 2019, 36, 213-218.
 19. Herweh, C.A, Ringleb, P, Rauch, G, Gerry, S, Behrens, L, Möhlenbruch, M, Gottorf, et al. Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients. *Int. J. Stroke* 2016, 11, 438-445.
 20. Char, D.S.; Shah, N.H.; Magnus, D. Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *N. Engl. J. Med*. 2018, 378, 981-983.
 21. Beck, A.H.; Sangoi, A.R.; Leung, S.; Marinelli, R.J.; Nielsen, T.O.; Van De Vijver, M.J.; West, R.B.; Van De Rijn, M.; Koller, D. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Sci. Transl. Med*. 2011, 3, 108-113.
 22. Hae, H.; Kang, S.J.; Kim, W.J.; Choi, S.Y.; Lee, J.G.; Bae, Y.; Cho, H.; Yang, D.H.; Kang, J.W.; Lim, T.H.; et al. Machine learning assessment of myocardial ischemia using angiography: Development and retrospective validation. *PLoS Med*. 2018, 15, e1002693.