Advances in Artificial Intelligence Research (AAIR) **Vol. 1 (No.1), pp.13-18, 2021**

www.dergipark.com/aair/ **Published online: Jan 15, 2021**

Classification of Death Related to Heart Failure by Machine Learning Algorithms

Remzi Gürfidan[1,*], Mevlüt Ersoy[2]

[1] Isparta University of Applied Science, Yalvac Technical Sciences Vocational School; [0000-0002-4899-2219]
[2] Süleyman Demirel University, Computer Engineering; [0000-0003-2963-7729]

Abstract

The increase in the number of individuals with heart diseases and deaths associated with these diseases tops the list of causes of death. Early detection and treatment can reduce the risk of death of candidates with heart disease and people with heart disease. With the expansion of artificial intelligence technology in the field of health in recent years, artificial intelligence models with prediction and classification capability that will contribute positively to patients and health workers are being developed.

In this study, the heart disease mortality status was classified according to the clinical data and life information of the patients included in the heart failure data set. The aim of this study is to evaluate the mortality associated with heart disease based on the clinical data and life information of the patients and to guide patients and doctors to early diagnosis or early treatment methods. Classification processes were performed with different machine learning algorithms and success rates were shown. Different algorithms have been tested to achieve success rates between 73% and 83%. Among the tried algorithms, the most successful classification process is provided by the Support Vector Machine (SVM) algorithm.

Keywords: *Machine learning; healthcare; hearth failure; support vector machine.*

1. Introduction

Heart disease is common throughout the world and is at the top of diseases that pose a high risk to human life. According to the Health Statistics report of the Ministry of health of the Republic of Turkey, the cause of 46.2 percent of deaths in the world other than infectious diseases in 2012 was stated as cardiovascular diseases. The report shows that this rate is increasing. Deaths due to cardiovascular diseases are estimated to be 22.2 million in 2030 [1]. In studies conducted on ready data groups, prediction and classification processes are commonly performed by machine learning algorithms. The results obtained may vary depending on the selected algorithm and the characteristics of the data set being studied.

In this study, experimental studies were conducted on the classification of death conditions due to heart disease using machine learning algorithms according to the measurement values and life information obtained from individuals. The data set used in this study was taken from the University of California Irvine Machine Learning Repository [2], "Heart failure clinical records Data Set 2020" is used by various researchers [3,4] and can be accessed from online data mining repository of the University of California. This dataset was used in this research stud for designing machine-learning-based system for heart failure classification. In the data set, 12 arguments are defined as input data. The expected result as output from the system is the classification of death conditions due to the individual's heart condition as a result of the algorithms subjected to the input values. 33% of the 300 data lines in the data set were allocated as test data and trainings were conducted and 83% success rate was achieved as the highest value.

2. Literature Review

Machine learning algorithms are often used in academic studies in recent years due to the successful prediction and classification results they show on ready-made data sets. Machine learning occurs in different fields of study such as health [5,6], cryptology [7], time series [8,9]. There are many studies in the literature on the diagnosis and classification of heart diseases. Most of these studies were carried out using machine learning methods [10-14]. Mohan and etc. used a hybrid machine learning method where random forest and linear regression algorithms were combined to predict heart disease. There are 303 data in the data set they used in their study and the success rate of the model they trained is 88 % [15]. Haq and etc. also used a hybrid machine learning method for the classification of heart diseases. They used classifiers such as Support Vector Machine algorithm (SVM), K-Nearest Neighbor algorithm, Logistic Regression algorithm within the hybrid model. As a model success, they reached 89% [16]. Kukar and etc. used machine learning algorithms for the classification of the diagnosis of ischemic heart diseases. In their study, they used ECG (electrocardiogram),

*Corresponding author
E-mail address: remzigurfidan@isparta.edu.tr

Received: 10/August/2020; Received in revised form: 2/September/2020; Accepted: 23/October/2020.

sequential ECG test during controlled exercise, myocardial scintigraphy, and finally coronary angiography images as data sets. The model they developed in their study provides 84,448% accuracy [17].

3. Proposed Model

The data set contains 13 clinical features. Of these properties, the first twelve properties are considered as independent variables, and the last property as dependent variables. The proposed model is subjected to training to independent variables and predicts the dependent variable part. Since the predicted value is 0 or 1, the operation performed is classification. Table 1 shows the properties within the data set.

Table 1. *Dataset Features and Descriptions*

Features	Description	Feature Type
Age	Age of The Patient	Years
Anaemia	Decrease of Red Blood Cells or Hemoglobin	Boolean
High Blood Pressure	If the Patient Has Hypertension	Boolean
Creatinine Phosphokinase	Level of The Cpk Enzyme in The Blood	Mcg/L
Diabetes	Percentage of Blood Leaving the Heart at Each Contraction	Percentage
Ejection Fraction	Platelets in The Blood	Kiloplatelets/Ml
Platelets	Woman or Man	Binary
Sex	Level of Serum Creatinine in The Blood	Mg/Dl
Serum Creatinine	Serum Sodium: Level of Serum Sodium in The Blood	Meq/L
Serum Sodium	Smoking: If the Patient Smokes or Not	Boolean
Smoking	Follow-Up Period	Days
Time	Follow-Up Period	Days
[Target] Death Event	If the Patient Deceased During the Follow-Up Period	Boolean

There is a total of 300 data lines in the data set. 99 of these data sets were randomly allocated as test data and 201 as training data. Six different classification algorithms have been tried on the data set in order to obtain the best result of the training. The mathematical models of each algorithm, the classification success confusion matrix values obtained from the training are shown in tables.

3.1. Support Vector Machine (SVM)

The hypothesis function called "h". X and y parameters are classification dimensions. The point above or on the hyperplane will be classified as class +1, and the point below the hyperplane will be classified as class -1. This math model shown in Equation 1.

$$h(x_i) = \begin{cases} -1, if\ w.x + b < 0 \\ +1, if\ w.x + bx \geq 0 \end{cases} \quad (1)$$

n parameter is being the number of features had and w is a point on the hyperplane. use on the soft-margin classifier since choosing a sufficiently small value for lambda yields the hard-margin classifier for linearly-classifiable input data. The mathematical model that enables these operations to be executed is shown in Equation 2.

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0,1-\ y_i(w.x_i-b)\right)\right]+\ \gamma\|w\|^2 \quad (2)$$

SVM algorithm was applied to the studied data and an accuracy value of 0.83 was obtained. The confusion matrix of the classification process performed at the end of the model training is shown in Table 2.

Table 2. *Confusion Matrix for SVM Algorithm*

Confusion Matrix	False (0)	True (1)	Total
False (0)	63	7	70
True (1)	11	18	29
Total	74	25	99

The data shown in the columns on the confusion matrix show the actual data and the data shown in the rows show the classification results of the test data. Of the 99 randomly selected test data, 74 are "0" data, 25 are "1" data. The trained model classified 63 of its false - false classifications as correct and 7 incorrectly, and 18 of its true - true classifications as correct and 11 incorrectly.

3.2. Logistic Regression Algorithm

Given a row of data (x, y) in the data set, x is a matrix of values with m instances and n properties, and Y is a vector with m instances. The purpose of the algorithm is to train the model to predict which class the values to be given to it belong to in the future. Primarily, have been created a weight matrix with random initialization. Then have been multiply it by features. The mathematical model of the mentioned operations is shown in Eq. 3.

$$a = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \tag{3}$$

This is followed by calculating Eq. 4 the cost for that iteration.

$$cost(w) = -\frac{1}{m}\sum_{i=1}^{i=m} y_i \log(y_i) + (1 - y_i)\log(1 - y_i) \tag{4}$$

P function is defined probability and math model shown in Eq. 5.

$$P(y = 1|x; w) \ \& P(y = 0\ |x; w) \tag{5}$$

Logistic Regression algorithm was applied to the studied data and an accuracy value of 0.82 was obtained. The confusion matrix of the classification process performed at the end of the model training is shown in Table 3.

Table 3. *Confusion Matrix for Logistic Regression Algorithm*

Confusion Matrix	False (0)	True (1)	Total
False (0)	63	7	70
True (1)	10	19	29
Total	73	26	99

The data shown in the columns on the confusion matrix show the actual data and the data shown in the rows show the classification results of the test data. Of the 99 randomly selected test data, 73 are "0" data, 26 are "1" data. The trained model classified 63 of its false - false classifications as correct and 7 incorrectly, and 19 of its true - true classifications as correct and 10 incorrectly.

3.3. Decision Tree Classifier Algorithm

Firstly, Compute the entropy for the data set. Entropy is calculated by the H method shown in Equation 6. Entropy H(S) is measure of the amount of uncertain in the dataset. S is the current set for which entropy is being calculated. C = {True, False} is set of classes in S. P function is the proportion of the number of elements in class c to the number of elements in set S.

$$H(S) = \sum_{c \in C} -p(c)log_2p(c) \tag{6}$$

Decision Tree Classifier algorithm was applied to the studied data and an accuracy value of 0.73 was obtained. The confusion matrix of the classification process performed at the end of the model training is shown in Table 4.

Table 4. *Confusion Matrix for Decision Tree Classifier Algorithm*

Confusion Matrix	False (0)	True (1)	Total
False (0)	58	12	70
True (1)	15	14	29
Total	73	26	99

The data shown in the columns on the confusion matrix show the actual data and the data shown in the rows show the classification results of the test data. Of the 99 randomly selected test data, 73 are "0" data, 26 are "1" data. The trained model classified 58 of its false - false classifications as correct and 12 incorrectly, and 14 of its true - true classifications as correct and 15 incorrectly.

3.4. K-Nearest Neighbor Algorithm (KNN)

The basic principle of K-nearest neighbor algorithm in classification problems is that a selected algorithm K can detect the closest neighbor of the hidden data point. Then, assigning the most obviously classified one to the hidden data point. The data numbers refer to all K neighboring classes. The Euclidean equation shown in Equation 7 is used to measure distances.

$$d(x,x') = \sqrt{(x-x_1')^2 + \cdots + (x-x_n')^2} \tag{7}$$

Finally, the input x gets assigned to the class with the largest probability. The variable x is defined to indicate a property and y to indicate the target. The K in KNN is a hyperparameter, must decide get the most suitable fit for the data set.

$$P(y=j|X=x) = \frac{1}{K}\sum_{i\in A} I\left(y^{(i)}=j\right) \tag{8}$$

KNN algorithm was applied to the studied data and an accuracy value of 0.73 was obtained. The confusion matrix of the classification process performed at the end of the model training is shown in Table 5.

Table 5. *Confusion Matrix for K-Nearest Neighbor Algorithm*

Confusion Matrix	False (0)	True (1)	Total
False (0)	63	7	70
True (1)	19	10	29
Total	82	17	99

The data shown in the columns on the confusion matrix show the actual data and the data shown in the rows show the classification results of the test data. Of the 99 randomly selected test data, 82 are “0” data, 17 are “1” data. The trained model classified 63 of its false - false classifications as correct and 7 incorrectly, and 10 of its true - true classifications as correct and 19 incorrectly.

3.5. Linear Discriminant Analysis Algorithm

Population π_i the probability density function of x is multivariate normal with mean vector μ_i and variance-covariance matrix Σ (same for all populations). normal probability density function is shown Eq. 9.

$$P(X|\pi_i) = \frac{1}{2\pi^{\frac{p}{2}}|\sum i|^{\frac{1}{2}}}\, exp\left[-\frac{1}{2}(X-\mu_i)'\frac{1}{\sum i}(X-\mu_i)\right] \tag{9}$$

According to the Naive Bayes classification algorithm, have been classified the population for which $P(\pi_i)$ $P(X|\pi_i)$ is the maximum. Linear Discriminant Analysis algorithm was applied to the studied data and an accuracy value of 0.82 was obtained. The confusion matrix of the classification process performed at the end of the model training is shown in Table 6.

Table 6. *Confusion Matrix for Linear Discriminant Analysis Algorithm*

Confusion Matrix	False (0)	True (1)	Total
False (0)	64	6	70
True (1)	12	17	29
Total	76	23	99

The data shown in the columns on the confusion matrix show the actual data and the data shown in the rows show the classification results of the test data. Of the 99 randomly selected test data, 76 are “0” data, 23 are “1” data. The trained model classified 64 of its false - false classifications as correct and 6incorrectly, and 17 of its true - true classifications as correct and 12 incorrectly.

3.6. Gaussian Naive Bayes Algorithm

In Bayes Theorem, the property vector X = (x_1, x_2, ..., x_n) is denoted by the given class variable C_k. The Naive Bayes classification problem have been formulated like Eq. 10.

$$C' = \text{argmax}\, P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \tag{10}$$

Gaussian Naive Bayes algorithm was applied to the studied data and an accuracy value of 0.78 was obtained. The confusion matrix of the classification process performed at the end of the model training is shown in Table 7.

Table 7. *Confusion Matrix for Gaussian Naive Bayes Algorithm*

Confusion Matrix	False (0)	True (1)	Total
False (0)	67	3	70
True (1)	19	10	29
Total	86	13	99

The data shown in the columns on the confusion matrix show the actual data and the data shown in the rows show the classification results of the test data. Of the 99 randomly selected test data, 86 are "0" data, 13 are "1" data. The trained model classified 67 of its false - false classifications as correct and 3 incorrectly, and 10 of its true - true classifications as correct and 19 incorrectly.

4. Discussion

The accuracy value of the model trained with machine learning takes a value between 0 and 1. The closer to 1, the higher the success of the model. The accuracy values obtained as a result of the training of the model trained according to different algorithms are shown graphically in Figure 1. When the obtained accuracy values were examined, the highest success was obtained from the model trained with SVM algorithm. LR and LDA follow this success.

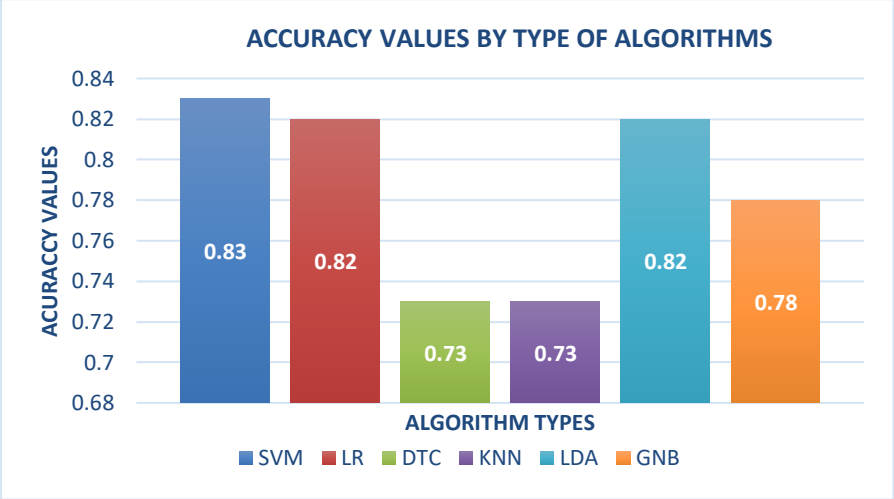

Figure 1. *Accuracy values by type of algorithms*

The 83 percent success rate measured in the developed model cannot be considered as bad, but considering the algorithms applied in this study, higher successes were expected from the developed models. The reason for the success rate to remain at this level is considered to be the number of data in the data set. 201 data were presented to the models as training data. As the number of data in the data set increases, the learning ability of the model will naturally improve and the accuracy value will increase accordingly.

5. Conclusion

In this study, experimental studies were conducted on the classification of heart failure related death conditions using different machine learning algorithms according to the measurement values and life

information obtained from individuals. The heart_failure_clinical_records data set was used in the study. In the data set, 12 arguments are defined as input data. The expected result as output from the system is to be able to classify the individual's death conditions due to heart disease as a result of algorithms subjected to input values. Of the 300 rows of data in the dataset, 33% of test data is divided into 67% of training data. The highest accuracy value obtained from different algorithms has been increased to 83%. In future studies, it is aimed to increase the amount of data in the data set and increase the final success rate to over 90%.

References

[1] Türkiye Cumhuriyeti Sağlık Bakanlığı Sağlık İstatistikleri Yıllığı, 2015

[2] Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16 (2020).

[3] 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.

[4] 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), 16.

[5] Zemmal, N., Azizi, N., Sellami, M., Cheriguene, S., Ziani, A., AlDwairi, M., & Dendani, N. (2020). Particle Swarm Optimization Based Swarm Intelligence for Active Learning Improvement: Application on Medical Data Classification. Cognitive Computation, 1-20.

[6] Hsu, C. N., Liu, C. L., Tain, Y. L., Kuo, C. Y., & Lin, Y. C. (2020). Machine Learning Model for Risk Prediction of Community-Acquired Acute Kidney Injury Hospitalization from Electronic Health Records: Development and Validation Study. Journal of *Medical Internet Research*, *22*(8), e16903.

[7] Al Shibli, M. (2020). Hybrid Artificially Intelligent Multi-Layer Blockchain and Bitcoin Cryptology (AI-MLBBC): Anti-Crime-Theft Smart Wall Defense. In Encyclopedia of Criminal Activities and the Deep Web (pp. 1089-1111). IGI Global.

[8] Jiang, Y., Bao, X., Hao, S., Zhao, H., Li, X., & Wu, X. (2020). Monthly Streamflow Forecasting Using ELM-IPSO Based on Phase Space Reconstruction. Water Resources Management, 1-17.

[9] He, L., Chen, S., Liang, Y., Hou, M., & Chen, J. (2020). Infilling the missing values of groundwater level using time and space series: case of Nantong City, east coast of China. Earth Science Informatics, 1-15.

[10] Yin, H. (2020). Smart Healthcare Via Efficient Machine Learning (Doctoral dissertation, Princeton University).

[11] Anand, H., Anand, A., Das, I., Rautaray, S. S., & Pandey, M. (2020, July). Hridaya Kalp: A Prototype for Second Generation Chronic Heart Disease Detection and Classification. In International Conference on Innovative Computing and Communications (pp. 321-329). Springer, Singapore.

[12] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. In 2017 IEEE Symposium on Computers and Communications (ISCC) (pp. 204-207). IEEE.

[13] Buettner, R., & Schunter, M. (2019, October). Efficient machine learning based detection of heart disease. In 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom) (pp. 1-6). IEEE.

[14] Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems (IJAIS), 3(7), 25-30.

[15] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.

[16] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018

[17] Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial intelligence in medicine*, *16*(1), 25-50.