International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

The Prediction of Heart Disease using Machine Learning Technique

Jae Won Choi, Young Keun Choi

Abstract: ***Medical errors are generally costly and harmful. There are many deaths worldwide every year. Clinical decision support systems provide opportunities to reduce medical errors and improve patient safety. Certainly, one of the most important aspects of applying such a system is the diagnosis and treatment of heart disease. Machine learning technology is implemented to analyze different kinds of heart-based problems. For this, this study essentially had two primary goals. Firstly, this paper intends to understand the role of variables in heart disease modeling better. Secondly, the study seeks to evaluate the predictive performance of the decision trees. Based on these results, first, men seem to be more susceptible to heart disease than women. Age, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, and the number of major vessels also show increased odds of having heart disease. Second, for the full model, the accuracy rate is 0.873, which implies that the error rate is 0.127. Among the patients who predicted not to have heart disease, the accuracy that would not have heart disease was 85.43%, and the accuracy that had heart disease was 89.10% among the patients predicted to have heart disease.***

Keywords: Decision tree, Heart disease, Machine learning

I. INTRODUCTION

Machine running is one of the fastest-growing areas of artificial intelligence, and is used in most medical applications. Because it is an intelligent tool for analyzing data, it has great value in health care, and it is rich in data. In the past few years, a large amount of data has been collected and stored due to the digital revolution. Monitoring and other data collection devices are available and used daily in modern hospitals, and large amounts of data are being collected. It is very difficult or impossible for humans to derive useful information from this vast amount of data. This is why machine running is widely used today to analyze this data and diagnose health care issues. A brief description of what the machine running algorithm does is learned from a previously diagnosed patient case. To save lives, you need to diagnose heart problems quickly, efficiently and accurately. This has led researchers to be interested in predicting the risk of heart disease and to use a variety of machine learning techniques to create different cardiac risk prediction systems.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Jae Won Choi, Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, TX, USA.
Young Keun Choi, Division of Business Administration, School of Business and Economics, Seoul, The republic of Korea.

Missing and outlier data in training sets often degrade model performance and inaccurate predictions. Therefore, it is important to process missing values and specific values before making any predictions.

Machine learning is a technology that grows fast and works with the human mind. Represents a multi-level record and effectively addresses the selectivity dilemma. Machine learning techniques are often used for scientific prediction. Process large amounts of data and smoothly decode complex hassles.

The purpose of this study is to find and analyze the cause of heart disease so that doctors and scientists can use it to formulate possible solutions to problems. The proactive approach and modeling methodology used in this white paper can be viewed as a roadmap for readers to follow the steps taken in this study and apply procedures to identify the causes of many other medical problems.

II. RELATED STUDY

In medicine, classification is one of the most important, important and widely used decision-making tools. Many modern technologies have been introduced to accurately and accurately predict heart disease. Some tasks related to this area are briefly described as follows:

Guan et al. [1] proposed a support vector machine-based system (SVM) that could effectively predict heart disease. The proposed system is 76.5% accurate. The system compared various models, such as standard SVMs, L0 standard SVM strategies, to eliminate and approximate recursive properties. Shilaskar and Ghatol [2] predicted heart algorithms using various classification strategies and election algorithms. They used the SVM classifier to optimize, select, and select forward properties. Their experiments have shown that they reduce the number of input variables and improve accuracy. The accuracy of the system is approximately 85%. Shao et al. [3] proposed a system that uses machine learning strategies such as logistic regression to find accuracy in predicting coronary heart disease and to select the most important and important functions. The system uses an approximate set strategy and multivariate adaptive regression splines to reduce the size of the description function for diagnosing heart disease. The proposed system achieved an accuracy of 82.14%.

Rajati and Radhamani [4] proposed a model using K-Nearest Neighbor (KNN) combined with Ant Colony Optimization (ACO) technology to predict coronary heart disease. Accuracy (70.26%) was compared with four machine operating algorithms. Bashir et al.

[5] proposed a system using SVM, Naïve Bayes, and DT-GI to predict heart disease. In the preprocessing step, nodule values in the data set are removed. The voting strategy was then used to determine the accuracy of the heart disease prediction. In our test results we found 82% accuracy.

Amin et al. [6] proposed a hybrid model for classifying heart disease using key risk features. They used two widely used tools in the system, the neural network system and the genetic algorithm. With the help of genetic algorithms and global optimization techniques, each neuron was measured in a neural network. Their experiments showed that the model is faster than other models and is 89% accurate. Khatibi and Montaser [7] adopted a fuzzy-based system using the fuzzy set concept and the Dempster-Shaper theory. The proposed method follows two steps: First, the input is described through the fuzzy unit and the fuzzy setup is performed through the fuzzy inference system. Second, we created a confidence interval for the hybrid inference engine and combined the various information using combinatorial rules.

Temurtas and Tanrikulu [8] proposed a model for classifying heart disease data sets using neural network technology. Dividing the dataset by three times the cross-validation was found using neural network strategy. In the experiment they achieved 96.30% accuracy of the model. Yumusak and Temurtas [9] used multilayer neural networks to predict coronary artery disease. Two hidden layers with an average accuracy of 91.60% were used between the input and output layers. Liu et al. [10] used a regression and local linear insertion (LLE) strategy for cardiac disease classification to achieve approximately 80% accuracy.

Nashif et al. [11] developed an application to monitor cardiac patients using various machine learning algorithms.The study proposed a cloud-based system that allows patients to upload physiological data to verify cardiac data. Data from the Canadian Community Health Survey data set with 20 important attributes was validated. The SEM model is defined by the relationship between CCC 121 and 20 properties. Where CCC 121 is a variable that defines whether the patient has heart disease. Ghadge et al. [12] developed an intelligent heart attack prediction system using Big Data. The main contribution of this study was to find models for intelligent heart failure prediction systems using big data and data mining modeling techniques.

III. METHODOLOGY

3.1 Dataset

Variable information is provided by a third-party website, an international issue on the popular internet platform Kagle (www.kaggle.com), which provides data called a 'heart disease dataset' uploaded by David Rap. This dataset was launched in 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. This includes 76 properties, including predicted ones, but all published experiments use some of the 14 properties. Kagle asked participants to predict heart disease. To help develop the algorithm, the organizer provided data stream types for large individual factors. These variables are listed and defined in Table 1.

Table 1. The variables in each category

Variables	Definitions
age	age in years
sex	(1 = male; 0 = female)
cp	chest pain type (4 values)
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results (values 0, 1, 2)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by flourosopy
thal	1 = normal; 2 = fixed defect; 3 = reversable defect
target	The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

3.2 Decision Tree

Among various analytical techniques, Decision Tree (DT) is a powerful and widely used machine learning algorithm for predicting and classifying medical data to date. Used for both classification and regression problems. Now you may have a question about why you want to use a DT classifier instead of another classifier. I can tell you two reasons to answer that question. One is that decision trees often try to mimic the way the human brain thinks, so understanding data and making good conclusions or interpretations is very simple. The second reason is that you can see the logic the data interprets, not the black box algorithms like SVM and NN in the decision tree. It has simple and clear expertise and has become one of the favorite among programmers of this generation. Now we've looked at why decision trees can take a closer look at what a decision tree classifier is. The beginning of a decision tree is a tree with multiple nodes, each node represents a function (attribute), each link represents a decision called a rule, and each leaf in the tree represents a different known result. The idea of a category type or persistence value is to create a tree of the entire data and get the result from every leaf. Now I know a little bit more about decision trees. We will continue to discuss how to write a decision tree classifier. Decision trees can be built with two algorithms. One is CART (Classification and Regression Tree) and the other is ID3.

For ID3, first use the x and y values in the column. This value remains at the end of the column and only exists for "YES" or "NO" values. The chart above has x values (view, temperature, humidity, wind) and there are only two options at the end of the column: 'YES' or 'NO' or y values. Now we need to map x and y.

As you can see, this is a binary classification problem, so let's create a tree using the ID3 algorithm. To create a tree, you must first select the root node that will be the root node. The general rule of thumb is to first choose the root node as the function that most affects the y value. Next choose the most influential feature as the next node. Here we will use the concept of entropy. The entropy concept measures the degree of uncertainty in a data set. In binary classification problems, we need to calculate entropy for all category type values. In summary, the entropy of the data set must be calculated first. For all attributes / functions, first calculate the entropy for all category type values, then get the average value information entropy for the current attribute to calculate the amount you get for the current attribute. Then you need to select the highest gain property and iterate until you get the tree you want. This is the process of ID3.

As discussed above, the decision tree classifier is written with another algorithm called Cart, which represents classification and regression trees. This algorithm uses the Gini Index as the cost function used to evaluate segment anger in the data set. If the target variable is actually a binary variable, use two values (yes and no). As we all know, there can be four different combinations. Now, to get a good idea of how to divide your data, you need to understand your Gini score. If the Gini score is 0, the worst case scenario is a 50/50 split, but it is completely isolated. The problem is now how to calculate the Gini index value.

Even if the target variable is a categorical variable at another level, the Gini index is similar. So the step of this method is the first calculation of the Gini index for the dataset. Then you need to calculate the Gini indexes for all the category type values of all functions, then get the average information entropy for the current property and then calculate the Gini gain. Once you have done this, you can select the best Gini gain properties and repeat until you get the tree you want. This is how the decision tree algorithm works.

The DT classification method involves building a tree model consisting of a set of predictors. These predictors (attributes) within the training set are iteratively split until a pure subset is obtained. This iterative personal splitting process is influenced by the characteristics of certain entities, for example, customers. The basic structure of a DT consists of leaf nodes and decision nodes. Leaf nodes represent the predictor and the point where binary splitting occurs. Leaf nodes are also called internal nodes. Crystal nodes, also known as terminal nodes, represent output variables (binary result variables) and are graphically displayed at the end of a branch. In most cases, terminal nodes based on the Exit Forecast report category. According to existing literature, 1) classification and regression tree (CART) 2) C4.5 3) Chi-squared auto-interaction detection (CHAID) and 4) C5.0 are commonly used. DT serves as the basis for other tree methods, such as random forest and ensemble forest, and by default multiple decision trees must be aggregated.

The process of splitting a property into binary must choose the correct property to split. The correct characteristic choice depends on the choice of GART reference (CART), depending on the entropy measurement (C4.5) calculation or DT algorithm type. DT analysis is very famous for its simplicity, graphical layout and ease of interpretation. DT provides a suitable schematic for modeling quantitative and qualitative determination problems without the need to create dummy variables or transformations. DT can also monitor and calculate nonlinear gender. However, DT has some disadvantages. DT results are not always as predictable as other methods. Moreover, minor changes in the data set can lead to unexpected predictions. However, this classification technique was often used to model deviations.

3.3 Data Mining Models

To survive in the market, many companies are using data mining technology for decision making. Effective customer management requires a more effective and accurate decision prediction model. Constructed a decision prediction model using statistical and data mining techniques. Data mining techniques can be used to predict or classify behavior by discovering interesting patterns or relationships in data and fitting models based on available data. If your training dataset and test dataset are separated for machine learning, your test dataset must meet the following requirements: First you need to create a training data set and a test data set in the same format. Second, the test data set should not be included in the training data set. Third, the training dataset and the test dataset must match the data. However, creating test data sets that meet these requirements is very difficult. To solve this problem in data mining, various validation frame operations were developed using a single dataset. This study supports the use of the Split Validation operator provided by RapidMiner. To support performance evaluation, the operator divides the input data set into training and test data sets. In this study, we select relative segmentation in the segmentation method parameters of these operators and use 70% of the input data as training data.

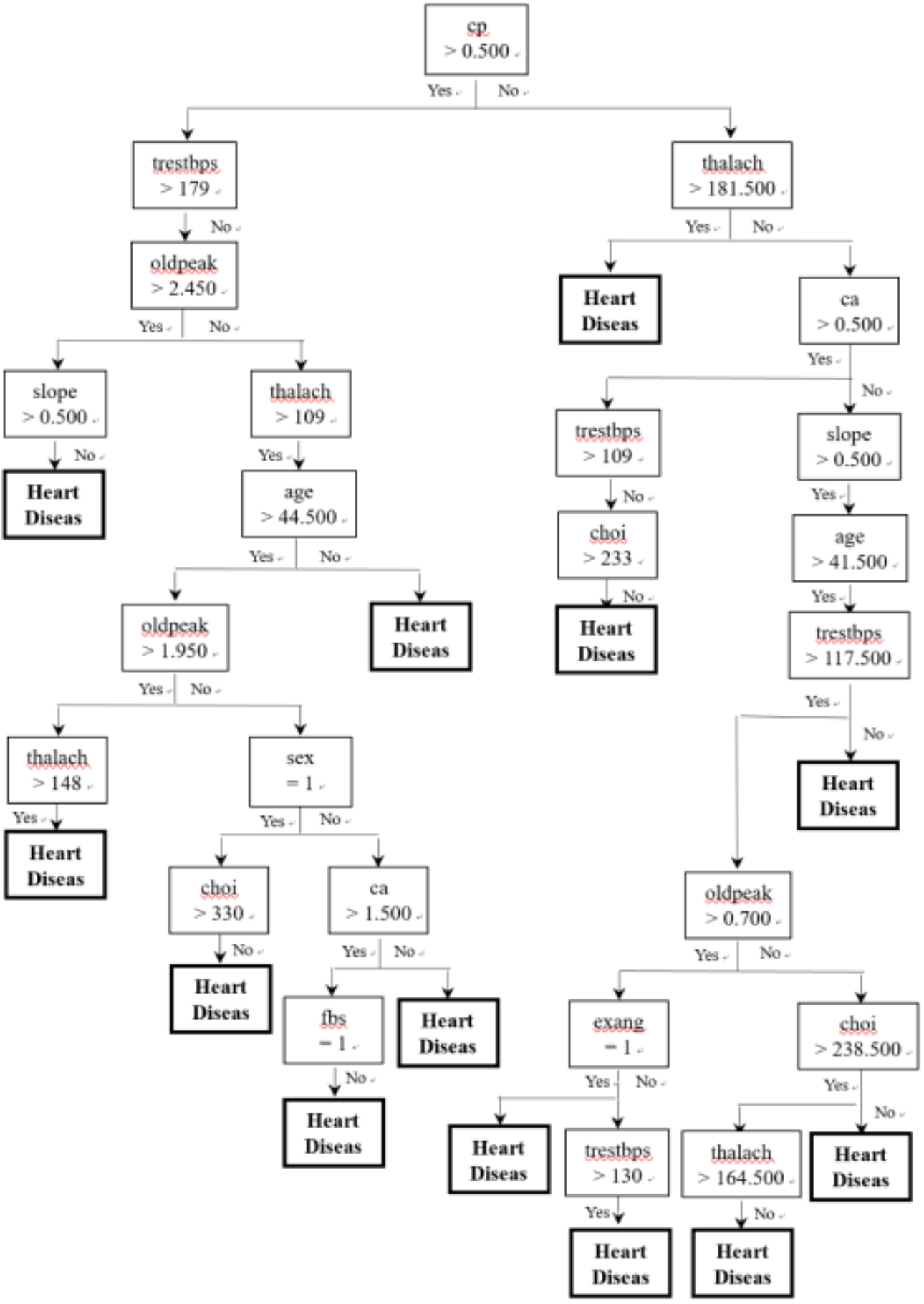
3.4 Performance Evaluation

Table 2. Key performance indicators

		Actual class (as determined by Gold Standard)	
		True	False
Predicted class	Positive	True Positive	False Positive (Type Ⅰ error)
	Negative	False Negative (Type Ⅱ error)	True Negative

Precision = TP/(TP+FP), Recall = TP/(TP+FN), True negative rate = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN), F-measure = 2·((precision·recall)/(precision + recall))

Performance assessment uses the training data to determine how well the model is built. Performance measurements can be divided into technical performance measurements and heuristic measurements. The technical performance measures used in this study show the performance results by building a model from the training data, processing the test data into a model, and comparing the class labels of the original validation



<Figure 1> Classification Tree for the Full Model

case with the predicted class labels. Measures of technical performance can be divided into supervised and unsupervised learning. The learning that is used in this study is classified and returned. All data used for this training and testing will have the original class values. Compare and analyze the original class values with the prediction results for performance.

Classification problems are the most common data analysis issues. Various indicators have been developed to measure the performance of classification models. Classification problems of category types are often used for accuracy, precision, recall, and f measurements. RapidMiner includes performance (classification) that measures performance metrics for common classification problems, and performance (differential classification) that provides performance metrics for binary classification problems. Table 2 shows how these indicators are calculated.

IV. RESULTS

Figure 1 shows the classification tree for the full model after pruning the tree using cross-validation to avoid overfitting. The key variables in the full model analysis consist of 14 ones, as shown below, based on the criterion established with each of these variables. In other words, the classifier has identified four potential questions along each of these variables and specific criteria as defined below to aid in the classification of unknown patients. Men seem to be more susceptible to heart disease than women. Age, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, and the number of major vessels also show increased odds of having heart disease.

Tables 3 illustrate each of the confusion matrix measures. For the full model, the accuracy rate is 0.873, which implies that the error rate is 0.127. Among the patients who predicted not to have heart disease, the accuracy that would not have heart disease was 85.43%, and the accuracy that had heart disease was 89.10% among the patients predicted to have heart disease.

Table 3. Performance evaluation

	True 1	True 0	Class precision
Pred. 1	129	22	85.43%
Pred. 0	17	139	89.10%
Class recall	88.36%	86.34%	

V.CONCLUSION

Heart disease is complex and causes many deaths every year. Ignoring the initial symptoms of heart disease can lead to drastic consequences for the patient in a short time. In this system, I used machine learning techniques to predict heart disease using Kaggle heart disease dataset. Medical-related information is huge, and in our application, this study has shown how to use huge data effectively to predict heart disease using the machine learning technique.

To recap, this study essentially had two primary goals. Firstly, this paper intends to understand the role of variables in heart disease modeling better. Secondly, the study seeks to evaluate the predictive performance of the decision trees. Based on the findings reported above, a series of implications are drawn.

Concerning the first goal, the findings of the study suggest that assessing the role of variables is complex and that their influences vary according to the classification methods employed. The decision tree methods highlight the explanatory power as most important to the analysis. Therefore, collectively no unanimous conclusions can be drawn about which explanatory variables are most critical to heart disease for all the methods employed in totality.

Yet, the findings of this study do shed some additional light on the patient's profile. The medical doctors should be seeking to predict heart disease on the classification methods employed. For example, first, men seem to be more susceptible to heart disease than women. Age, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, and the number of major vessels also show increased odds of having heart disease. Second, for the full model, the accuracy rate is 0.873, which implies that the error rate is 0.127. Among the patients who predicted not to have heart disease, the accuracy that would not have heart disease was 85.43%, and the accuracy that had heart disease was 89.10% among the patients predicted to have heart disease.

This study provides some research contributions and actual contributions. First, this study extends existing literature by experimentally examining the effects of variables on heart disease modeling. Heart disease has a big impact on the patient. Although much research has been done on heart disease, no one can say that it can create a universal human tool that can predict heart disease. Heart disease is so complex that it is associated with many factors, so researchers tend to use fewer and ignore the effects of other factors. Patient demographics are often constantly changing and monitored, which can cause hospital problems and compromise personal information. Some studies examined age, gender, and geographic location. But researchers are still unable to express cultural and behavioral factors that can affect heart disease. This study contributes to the literature on heart disease by providing a global model that summarizes heart disease determinants of patient factors. Second, the methodology used in this white paper can be viewed as a roadmap for applying the process for readers to follow the steps taken in this case study and to identify the causes of many other problems throughout the day. This paper seeks to present the best performance model for predicting heart disease based on a limited set of features, including patient factors. Mechanical learning techniques and functional importance analysis, including crystal trees and neural networks, are used to achieve the best results in terms of accuracy. With this methodology, the study identified patterns of heart disease that can predict patients' heart disease.

Practically, this application helps doctors manage patient health records and can speed up treatment if the user already has a patient report. Quick treatment saves lives. This application helps patients track their health records. Therefore, it is helpful to take care of your health regularly. Analyst reports help doctors predict heart disease easily and easily. The proposed system has a database that stores patient records, and as the number of patients increases, more data is generated and storage becomes an issue. Therefore, future releases will provide cloud capabilities to store all records in the cloud. So if you have the right to secure your data and access it, you can search it from anywhere. Smart devices will synchronize with applications in future releases. Therefore, the real-time health status of the patient is monitored and in case of an emergency, the user is alerted. This reduces the risk.

In the future, the machine running model will use a larger training data set using more than one million different data points maintained in electronic health recording systems. While calculations and software sophistication can be a major leap forward, systems operating in artificial intelligence allow doctors to determine the best care available for the patient concerned as soon as possible. Software APIs can be developed so that health websites and apps have free access to patients. Probability predictions will be performed with or without any delay in the processing.

REFERENCES

1. W. Guan, A. Gray, and S. Leyffer, “In mixed-integer support vector machine,” in Mini Symposia & Workshops NIPS, 2013, pp. 1–6.
2. S. Shilaskar, and A. Ghatol, “Feature selection for medical diagnosis: evaluation for cardiovascular diseases,” Expert. Syst. Appl., Vol. 40, 2013, pp. 4146–4153.
3. Y. E. Shao, C. D. Hou, and C. C. Chiu, “Hybrid intelligent modeling schemes for heart disease classification,” Appl. Soft. Comput., Vol. 14, 2014, pp. 47–52.
4. S. Rajathi, and G. Radhamani, “Prediction and analysis of Rheumatic heart disease using KNN classification with ACO,” International conference on data mining and advanced computing (SAPIENCE), Ernakulam, 2016, pp.68–73.
5. S. Bashir, U. Qamar, and M. Y. Javed, “An ensemble-based decision support frame work for intelligent heart disease diagnosis,” in Information Society International Conference, IEEE, 2014, pp. 259–264
6. S. U. Amin, K. Agarwal, and R. Beg, “Genetic neural network-based data mining in prediction of heart disease using risk factors," Information and Communication Technologies(ICT), IEEE, 2013, pp.1227–31
7. V. Khatibi, and G. A. Montazer, “A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment,” Expert Syst. Appl., Vol. 37, 2010, pp. 8536–8542.
8. F. Temurtas, and A. C. Tanrikulu, “An approach on probabilistic neural network for diagnosis of mesothelioma’s disease.” Comput. Electr. Eng., Vol. 38, 2012, pp. 75–81.
9. N. Yumusak, and F. Temurtas, “Chest diseases diagnosis using artificial neural networks. Expert. Syst. Appl., Vol. 37, 2010, pp. 7648–7655.
10. X. Liu, D. Tosun, M. W. Weiner, and N. Schuff, “Locally linear embedding for MRI based Alzheimer’s disease classification,” NeuroImage, Vol. 83, 2013, pp. 148–57.
11. S. Nashif, M. R. Raihan, M. R., Islam, and M. H. Imam, “Heart disease detection by using machine learning algorithms and a real-time Cardiovascular health monitoring system,” World J. Eng. Technol., Vol. 6, 2018, pp. 854–873.
12. P. Ghadge, V. Girme, K. Kokane,and P. Deshmukh, “Intelligent heart attack prediction system using Big data,” Int. J. Recent Res. Math. Comput. Sci. Inf. Technol., Vol. 2, 2016, pp. 73–77.

AUTHORS PROFILE

Jae Won Choi, Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, TX, USA. His research interests are data science, artificial intelligence, blockchain, game, and etc.

Young Keun Choi, Division of Business Administration, School of Business and Economics, Seoul, The republic of Korea. His research areas are business analytics, data science, entrepreneurship, technology management, and etc.