ANALYSIS ON CARDIOVASCULAR DISEASE CLASSIFICATION USING MACHINE LEARNING FRAMEWORK

S. Kathiresan
Department of Computer Science, Bharathiyar University, India

Abstract

A big issue in the world of medical computing or clinical care is the classification of large medical records in particular cases of cardiac disease. Lack of a proper method for diagnosing cardiovascular disease results in a lack of early prediction. By designing machine-learning algorithms, a greater provision can be made for classifying patients on the basis of clinical records in the prediction of cardiovascular disease. In this article, we use a machine learning model to forecast cardiac rate at an earlier rate that enhances exam and assessment precision. This approach covers both cardiovascular disease surveillance, classification and estimation on a large dataset in real time. The experimental findings demonstrate the reliability of the proposed approach in real time datasets against existing methods and increase the precision in classification.

Keywords:

Cardiovascular disease, Classification, Machine Learning

1. INTRODUCTION

The people today is rising every day as a result of heritage. This not only generates time but also a great deal of knowledge, thus wasting health survey results. Now, however, there has been an inspection of results. The project was planned to generate a huge number of efficient data from hospitals [12]. During this time, cardiac disease is a deadly adversary. The condition should be treated as much as possible or this person would be infected. And diagnosing patients is the best time, and this is a challenging task. Hospital evaluations may be misdiagnosed often, resulting in bad names [9]. The condition is very hard to cure and most people cannot treat it.

In terms of the prevalence and severity of the risk of cardiovascular disease in older people, the degree of progress in recognising individuals at a greater or lesser risk of diabetes amongst this demographic and the degree to which they are able to distinguish. Indeed, it can be found that the risk of other risk factors in middle-aged diabetes adults ranges considerably [3] [4]. Some of the most widely used, including the Framingham Risk Score [5] and the Reynolds Risk Score [6], have been evaluated directly by diabetes-free individuals as well as, like so many others, by routinely eliminating older adults in derivative [10].

In the population of older adults, considering the prevalence and severity of the risk of cardiovascular disease (CVD) with diabetes the determinants of CVD have not been well clarified and the degree to which individuals with high or low risk are effectively identified. It is indeed apparent the risk ranges considerably based upon other risk factors in adults of Middle Ages with diabetes [3]-[5]. While many risk forecasting models are available, several of those most widely used have been studied primarily in adults without diabetes, such as the Framingham Risk Score [11] and Reynolds Risk Score [7], whereas, like many others, older adults were systemically omitted in their derivatives.

Two machine learning algorithms were used in paper [1] to forecast cardiovascular disease. In paper [2], the presence of heart disease is predicted by machine learning. A hybrid computer approach for the diagnosis of heart disease was used in document [8]. Data mining technology was used in paper [6] to detect and control cardiovascular disorders. The MLA method was used in paper [5] for atherosclerosis diagnosis and prediction.

2. PROPOSED METHOD

Cloud-based applications for tracking and forecasting heart attack are recommended for machine learning. This method lets cardiologists determine well and easily. This paper deals with many outstanding systems from different patient monitoring providers, to work with big data such as in-house data and health services. If this report does not deal with this high level of detail, some of the essential data features would be skipped. Real-time configurations are not used in most analyses. However, it just uses this file. And only consistency is the basis for such tests. However, this document deals with precision and real-time.

2.1. DATASET

For pre-processing, the final attributes are first chosen. This data set contains 300 people's information.

Table.1. Dataset

Index	Age	Sex	Fbs	Mhr	Eia	Cpt	Sc	Recg	Rbp	*Std*
1	64	0	0	151	1	2	234	0	140	2.4
2	68	0	1	109	0	3	287	0	165	1.6
3	68	0	1	130	0	3	230	0	125	2.7
4	38	0	1	188	1	4	251	2	120	3.
5	42	1	1	173	1	1	205	0	120	1.5
6	57	0	1	179	1	1	237	2	110	0.9
7	63	1	1	161	1	3	269	0	150	3.7
8	58	1	1	164	0	3	355	2	130	0.7
9	64	0	1	148	1	3	255	0	140	1.5
10	64	0	0	156	0	3	204	0	130	3.2

a) Architecture

Machine learning frameworks are built here, including the cloud. Four measures are here to monitor the occurrence of heart disease and to estimate it.

To monitor the health of patients, computers and sensors are used. In phase 1, this tracking gathers and integrates data from sensors and computers. In the second step, as the data are tracked and gathered every day, the data is increasing. To gather all this, it takes a large amount of computer. The cloud server is used to store a vast variety of details in real time. Step 3 creates a real-time design to diagnose coronary disease early. They use the MLA needed to predict and identify them. Step 4 sends users the final result. To do this, a request is used here. The four measures are structured in the Fig.1. Cloud server can accept and store requests from the monitoring system. If the current data is the same as the previously trained data, it is stored on a cloud server. If the latest data is above or below or does not equal the pre-trained data, this information is used as a user prompt. In this article, MLA, cloud storage, NoSQL databases and real-time predictive service offer a safer option for heart disease.

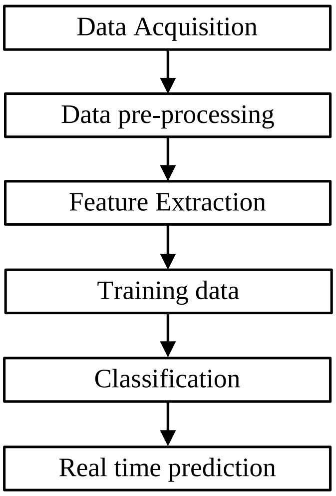

Fig.1. Proposed Architecture

The suggested protocol consists of four steps: data acquisition, data preparation, and research module and programme introduction, respectively. Data is obtained by means of devices and sensors to remove data from a given patient. Health tracking systems are connected into the human body to continually capture a single person health records. In addition, data are smoothly sent by tracking systems. This large number of data was made complicated by using conventional database methods and software. This proposed framework uses cloud and NoSQL databases to actively gather health records. Users can access their usage reports for the mobile application in the application phase.

2.2. FEATURE EXTRACTION

In this paper, we use the Laplacian filter to eliminate the noises from the image. The study uses Partial Differential Equation:

$$\frac{\partial u}{\partial t} = -sign\left(\nabla^2 u\right) \times \left|\nabla u\right| \tag{1}$$

The standard adaptive is utilized in this filter for forward differencing for expressing the upwind derivatives. Here, the Laplacian expression sign is represented as:

$$\frac{\partial u}{\partial t} = -\left|\nabla u\right| \tag{2}$$

In addition, the filter is applied with the minima and it suits if the Laplacian is negative:

$$\frac{\partial u}{\partial t} = \left|\nabla u\right| \tag{3}$$

The results of the filter offers dilation round the maxima region. Therefore the images tends to get sharper with the operations filter. The expression of disintegration and dilation operations is considered to be isolate and the expression of the Eq.(4) to Eq.(8):

$$\nabla u = \sqrt[2]{\left(u_x^2 + u_y^2\right)} \tag{4}$$

where

$$u_{x^2} = \left\{ \min\left(\frac{\left(u_{i,j} - u_{i,j-1}\right)}{h_x, 0} \right) \right\}^2 + \left\{ \max\left(\frac{\left(u_{i,j+1} - u_{i,j}\right)}{h_x, 0} \right) \right\}^2 \tag{5}$$

$$u_{y^2} = \left\{ \min\left(\frac{\left(u_{i,j} - u_{i,j-1}\right)}{h_y, 0} \right) \right\}^2 + \left\{ \max\left(\frac{\left(u_{i,j+1} - u_{i,j}\right)}{h_y, 0} \right) \right\}^2 \tag{6}$$

The Laplacian term sign is considered in Eq.(6) is negative i.e. the operation is considered dilation. If the operation is erosion, equations below is applied:

$$u_{x^2} = \left\{ \max\left(\frac{\left(u_{i,j} - u_{i,j-1}\right)}{h_x, 0} \right) \right\}^2 + \left\{ \min\left(\frac{\left(u_{i,j+1} - u_{i,j}\right)}{h_x, 0} \right) \right\}^2 \tag{7}$$

$$u_{y^2} = \left\{ \max\left(\frac{\left(u_{i,j} - u_{i,j-1}\right)}{h_y, 0} \right) \right\}^2 + \left\{ \min\left(\frac{\left(u_{i,j+1} - u_{i,j}\right)}{h_y, 0} \right) \right\}^2 \tag{8}$$

Discrete Laplacian solution is given below:

$$\nabla^2 u = \frac{\left\{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4 \times u_{i,j}\right\}}{h^2} \tag{9}$$

Combining the above expressions, the equations for filtering is expressed below:

$$\frac{\partial u}{\partial t} = \frac{\left(u_{i,j}^{k+1} - u_{i,j}^{k}\right)}{\nabla t} \tag{10}$$

2.3. GENETIC ALGORITHM

A selection of the best individuals from a population begins with the mechanism of natural selection. They build children that inherit parents' traits and are introduced to the next generation. If parents have increased fitness, their children would have a greater chance of survival than their parents. This method is still in progress and at the end there is a generation of the most adept people.

2.3.1. Initial Population:

The phase starts with a collection of people called a population. Each person solves the problem that he or she needs to solve. A human has a set of parameters (variables) called genes. Genes shape a chromosome into a series (solution).

The genes of a single person are represented by a string as an alphabet in a genetic algorithm. Binary values are commonly used (string of 1s and 0s). In a chromosome, we say we are coding the genes.

2.3.2. Fitness Function:

Fitness determines the fitness of an individual (the ability of an individual to compete with other individuals). Each individual receives a health score. The possibility of an individual being chosen to be repeated depends on their health.

2.3.3. Selection:

The selection process is planned to pick the best people to encourage them to disperse their genes to the next generation. Due to their fitness ratings, two pairs of individuals (parents) are chosen. High health individuals have the ability to be reproductively picked.

2.3.4. Crossover:

In a genetic algorithm the crossover is the most critical process. A crossing point is selected from within the genes randomly for each pair of parents to be matted. Offspring are formed by sharing parents' genes with each other before the crossroads are reached. The population is replaced by the new heirs.

2.3.5. Mutation:

Any of their genes may undergo a mutation with a low random chance in any new offspring. This ensures certain bits can be inverted in the bit series.

2.3.6. Termination:

If the population converges, the algorithm stops (does not produce offspring which are significantly different from the previous generation). The genetic algorithm is then said to have given a variety of solutions for our problem.

Pseudocode

```
START
    Generate the initial population
    Compute fitness
REPEAT
    Selection
    Crossover
    Mutation
    Compute fitness
UNTIL
    Population has converged
STOP
```

3. RESULTS AND DISCUSSION

For the determination of atherosclerosis, many approaches are used. Sensitivity is all these processes. It quantifies the percentage of people who have the disorder more reliably.

Table.2. CM for Binary Classification

Predicted Output (PO)	Actual Outputs (AO)	
	No disease	**Diseased**
Negative Test (NT)	TN	FN
Positive Test (PT)	FP	TP

The precision is to accurately measure the number of patients without a disorder. The performance of the algorithms used depends on the process. The correlation coefficient of Matthews tests the power of machine learning in binary classification. The CM is also called the error matrix of machine learning. For utility of the algorithms, refer to Table.2. The matrix comprises 2 types of details. This is the reliability and efficacy of the expected algorithms.

The True Positive (TP) case is that a disorder called atherosclerosis was properly identified by the patient. FP is an erroneous predictor of disease. The TN is a prediction that the patient is safe and exact. FN is an inaccurate hypothesis that the patient is safe. FN is a poor prediction.

The KNN and KMC approaches are used to provide the Therapeutic Diagnostic Support Framework for the Diagnosis of atherosclerosis diagnosis (AD). Two AD classes are open. Atherosclerosis is one healthy and the other. The NoSQL database includes 280 examples. 70% of the preparation data were used. And 15% of validation information was used. The majority is used for research purposes. For atherosclerosis, two forms of machine learning were used. The estimation and classification results of the proposed method are based on the uncertainty matrix in Table.3. Each uncertainty matrix cell includes a range of current and projected source values.

Table.3. uncertainty matrix of two MLA

ML Algorithms	FN	TN	FP	TP
KNN	7	42	6	28
KMC	10	25	23	25

The Table.4 explains the outcomes of success and assessment of the AD. The findings were found using KNN and KMC algorithms. The findings showed that 98% of the sensitivity of the proposed method was 97% and 97% were the good rates for specificity.

Table.4. Performance of MLA

Measures	KNN	KMC
MCC (%)	0.75	0.25
A (%)	89	61
SE (%)	85	76
SP (%)	92	54

The suggested machine output measure will, on the other hand, be the MCC search. The value of MCC is 0.94. If the outcome is one, then it has a nice and accurate preview of the proposed method. In order to improve the rate, the precision of the proposed method is organised.

4. CONCLUSION

In this article, supervised algorithms of machine learning are used to forecast heart disease. A real-time four-stage cloud system is built that incorporates the learning algorithm on a variety of parameters. The efficiency is calculated to be more than 10 times and cloud infrastructure for forecasting and diagnosing heart disease is integrated. These cloud systems have also been able to diagnose cardiac diseases by gathering information from the health centre. The cloud system will monitor and transmit alert messages to patients to detect cardiovascular diseases. Thus, this device can be used for heart disease prediction and tracking.

REFERENCES

[1] A. Liaw and Matthew Wiener, "Classification and Regression by Random Forest", *R News*, Vol. 2, No. 3, pp. 18-22, 2002.

[2] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", *Intelligent Computing in Signal Processing and Pattern Recognition*, Vol. 345, pp. 721-727, 2006.

[3] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", *International Journal of Biological, Biomedical and Medical Sciences*, Vol. 3, No. 3, pp. 1-8, 2008.

[4] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", *Studies in Health Technology and Informatics*, Vol. 107, No. 2, pp. 1256-1259, 2004.

[5] A. Dallali, A. Kachouri and M. Samet, "Classification of Cardiac Arrhythmia Using WT, HRV, and Fuzzy C-Means Clustering", *Signal Processing: An International Journal*, Vol. 5, No. 3, pp. 101-109, 2011.

[6] P. Radha and B. Srinvasan, "Hybrid Prediction Model for the Risk of Cardiovascular Disease in Type-2 Diabetic Patients", *Expert Systems with Applications*, Vol. 2, No. 10, pp. 23-29, 2014.

[7] J. Kim, H.S. Shin, K. Shin and M. Lee, Robust Algorithm for Arrhythmia Classification in ECG using Extreme Learning Machine", *Biomedical Engineering*, Vol. 8, No. 1, pp. 1-31, 2009.

[8] S.V. Arun Kumar and B.S. Harish, "A Modified Intuitionistic Fuzzy Clustering Algorithm for Medical Image Segmentation", *Journal of Intelligent Systems*, Vol. 27, No. 4, pp. 593-607, 2017.

[9] J.H. Morra, Z. Tu, L.G. Apostolova, A.E. Green, A.W. Toga and P.M. Thompson, "Comparison of Ada Boost and Support Vector Machines for Detecting Alzheimer's Disease through Automated Hippocampal Segmentation", *IEEE Transactions on Medical Imaging*, Vol. 29, No. 1, pp. 30-43, 2010.

[10] Nidhi Singh and Divakar Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", Ph.D Dissertation, Department of Computer Science and Engineering, Barkatullah University Institute of Technology, 2012.

[11] Ersen Yilmaz and Caglar Kilikcier, "Determination of Patient State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree", Master Thesis, Department of Electrical Electronic Engineering, Uludag University, 2013.

[12] K. Ming Leung, "Naive Bayesian Classifier", Master Thesis, Department of Computer Science and Engineering, Polytechnic University, 2007.