IOP Conference Series: Materials Science and Engineering

**PAPER** • **OPEN ACCESS**

# Heart disease prediction using machine learning algorithms

To cite this article: Harshit Jindal *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1022** 012072

View the article online for updates and enhancements.

You may also like

- A Classification Method of Heart Disease Based on Heart Sound Signal
Lizhiyaun and Liuhaikuan
- Determining which physical parameters are significant for heart disease
Zhuoning Li, Dingxin Tao, Jiaao Zheng et al.
- Prediction of Heart Diseases using Random Forest
Madhumita Pal and Smita Parija


This content was downloaded from IP address 103.144.120.205 on 15/08/2024 at 16:25

# Heart disease prediction using machine learning algorithms

**Harshit Jindal[1], Sarthak Agrawal[1], Rishabh Khera[1], Rachna Jain[2] and Preeti Nagrath[2]**

[1] Student, Dept. Of Electronics And Communication Eng. Bharti Vidyapeeth's College Of Engineering, New Delhi.
[2] Faculty, Dept. Of Computer Science & Engineering Bharti Vidyapeeth's College Of Engineering, New Delhi.
Email: harshit.jindal50@gmail.com

**Abstract.** Day by day the cases of heart diseases are increasing at a rapid rate and it's very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the .pynb format.

## 1.Introduction

"Machine Learning is a way of Manipulating and extraction of implicit, previously unknown/known and potential useful information about data" **[1].** Machine Learning is a very vast and diverse field and its scope and implementation is increasing day by day. Machine learning Incorporates various classifiers of Supervised, Unsupervised and Ensemble Learning which are used to predict and Find the Accuracy of the given dataset. We can use that knowledge in our project of HDPS as it will help a lot of people.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Cardiovascular diseases are very common these days, they describe a range of conditions that could affect your heart. World health organization estimates that 17.9 million global deaths from (Cardiovascular diseases) CVDs **[2].**

It is the primary reason of deaths in adults. Our project can help predict the people who are likely to diagnose with a heart disease by help of their medical history **[6].** It recognizes who all are having any symptoms of heart disease such as chest pain or high blood pressure and can help in diagnosing disease with less medical tests and effective treatments, so that they can be cured accordingly.

This project focuses on mainly three data mining techniques namely: (1) Logistic regression, (2) KNN and (3) Random Forest Classifier. The accuracy of our project is 87.5% for which is better than previous system where only one data mining technique is used. So, using more data mining techniques increased the HDPS accuracy and efficiency. Logistic regression falls under the category of supervised learning. Only discrete values are used in logistic regression.

The objective of this project is to check whether the patient is likely to be diagnosed with any cardiovascular heart diseases based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc. A dataset is selected from the UCI repository with patient's medical history and attributes. By using this dataset, we predict whether the patient can have a heart disease or not. To predict this, we use 14 medical attributes of a patient and classify him if the patient is likely to have a heart disease. These medical attributes are trained under three algorithms: Logistic regression, KNN and Random Forest Classifier. Most efficient of these algorithms is KNN which gives us the accuracy of 88.52%. And, finally we classify patients that are at risk of getting a heart disease or not and also this method is totally cost efficient.

## 2. Related Work

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient Cardiovascular disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives **[7].**

The model incorporating IHDPS had the ability to calculate the decision boundary using the previous and new model of machine learning and deep learning. It facilitated the important and the most basic factors/knowledge such as family history connected with any heart disease. But the accuracy that was obtained in such IHDPS model was far more less than the new upcoming model such as detecting coronary

heart disease using artificial neural network and other algorithms of machine and deep learning. The risk factors of coronary Heart disease or atherosclerosis is identified by McPherson et al**.,[8]** using the inbuilt implementation algorithm using uses some techniques of Neural Network and were just accurately able to predict whether the test patient is suffering from the given disease or not.

Diagnosis and prediction of Heart Disease and Blood Pressure along with other attributes using the aid of neural networks was introduced by R. Subramanian et al.,[**24].** A deep Neural Network was Built incorporating the given attributes related to the disease which were able to produce a output which was carried out by the output perceptron and almost included 120 hidden layers which is the basic and most relevant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. The supervised network has been advised for diagnosis of heart diseases [16]. When the testing of the model was done by a doctor using an unfamiliar data, the model used and trained from the previous learned data and predicted the result thereby calculating the accuracy of the given model.

## 3. Data Source

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions **[2].** Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. We take a data source which is comprised of medical history of 304 different patient of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 304 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 14 columns, where each row corresponds to a single record. All attributes are listed in ‘Table 1’.

**Table 1.** Various Attributes used are listed

| S. No | Observation | Description | Values |
|---|---|---|---|
| 1. | Age | Age in Years | Continuous |
| 2. | Sex | Sex of Subject | Male/Female |
| 3. | CP | Chest Pain | Four Types |
| 4. | Trestbps | Resting Blood Pressure | Continuous |
| 5. | Chol | Serum Cholesterol | Continuous |
| 6. | FBS | Fasting Blood Sugar | < ,or > 120 mg/dl |
| 7. | Restecg | Resting Electrocardiograph | Five Values |
| 8. | Thalach | Maximum Heart Rate Achieved | Continuous |
| 9. | Exang | Exercise Induced Angina | Yes/No |
| 10. | Oldpeak | ST Depression when Workout compared to the Amount of Rest Taken | Continuous |
| 11. | Slope | Slope of Peak Exercise ST segment | up/ Flat /Down |
| 12. | Ca | Gives the number of Major Vessels Coloured by Fluoroscopy | 0-3 |
| 13. | Thal | Defect Type | Reversible/Fixed/Normal |
| 14. | Num(Disorder) | Heart Disease | Not Present /Present in the Four Major types. |

## 4. Methodology

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model **[13].** The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology (Figure 1.) includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3$^{rd}$ is the preprocessing stage where we explore the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used **[15].** After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective **Heart Disease Prediction System**

**(EHDPS)** has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction **[17].**

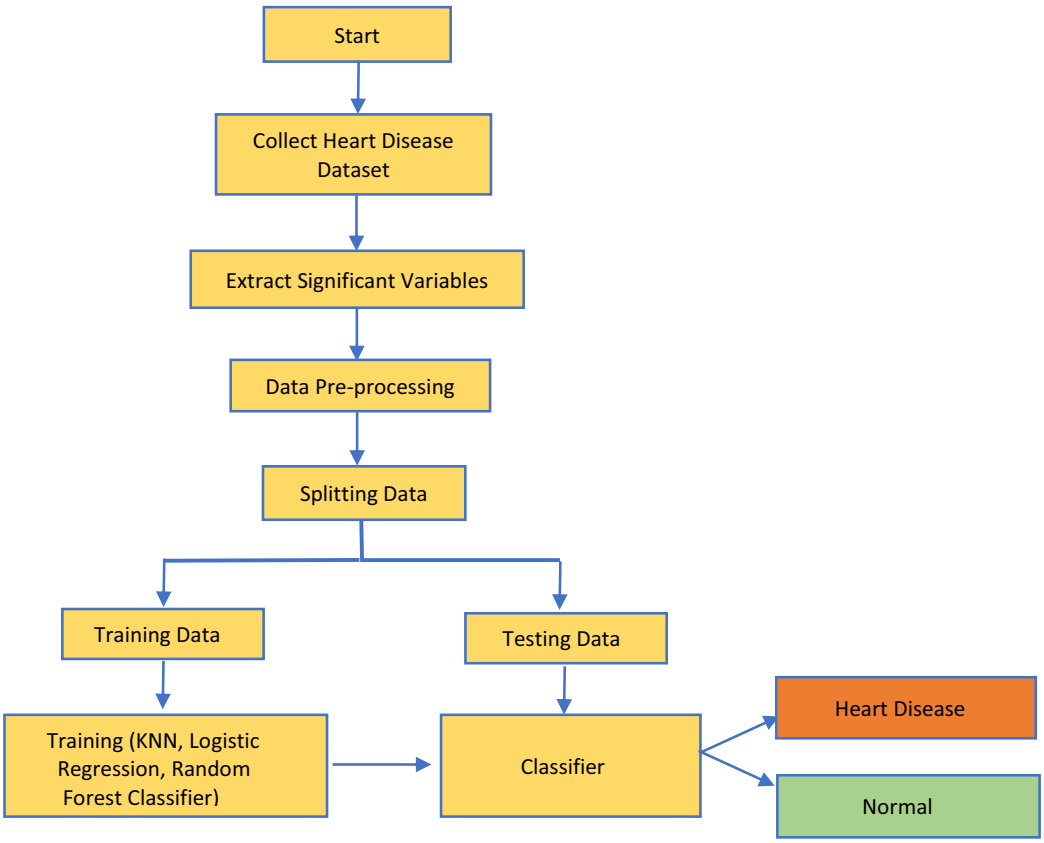


Figure 1. Proposed Model

## 5. Results & Discussions

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them **[23].** The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 88.5% which is greater or almost equal to accuracies obtained from previous researches. So, we summarize

that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following 'figure 2', 'figure 3', 'figure 4', 'figure 5' shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain:

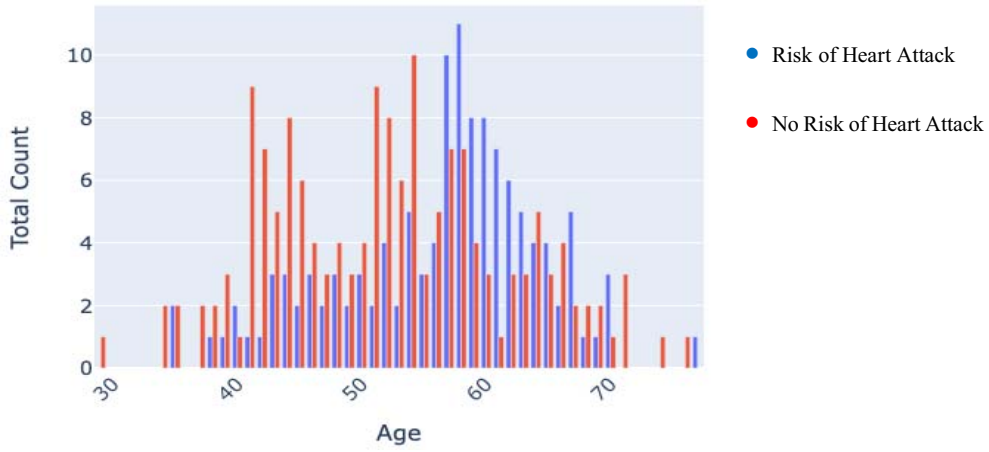


**Figure 2.** Shows the Risk of Heart Attack on the basis of their age.

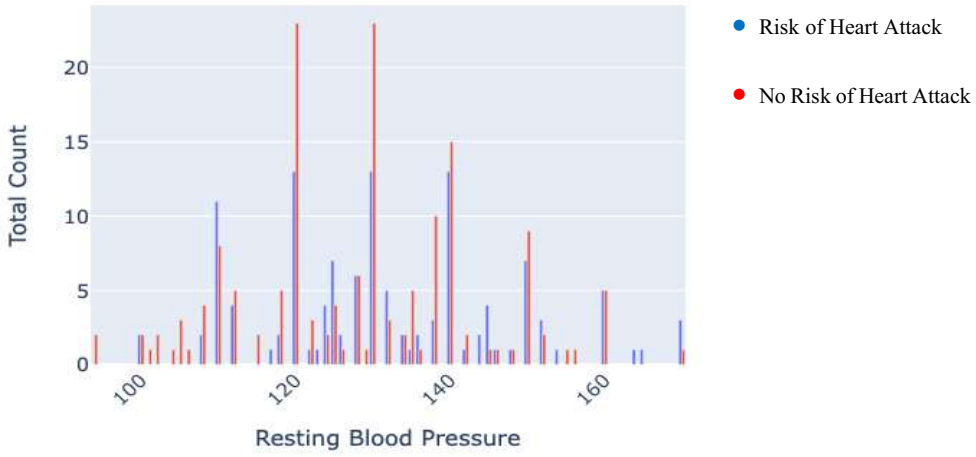


**Figure 3.** Shows the Risk of Heart Attack on the basis of their Resting Blood Pressure.

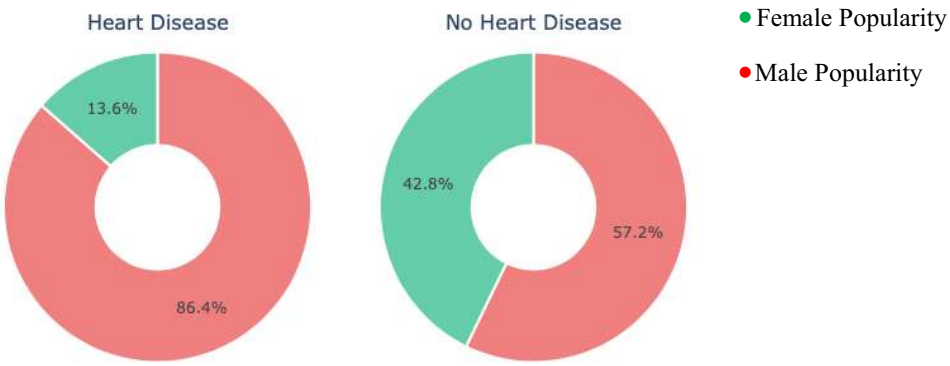


**Figure 4.** Shows the patients having or not having Heart Disease on the basis of Sex.

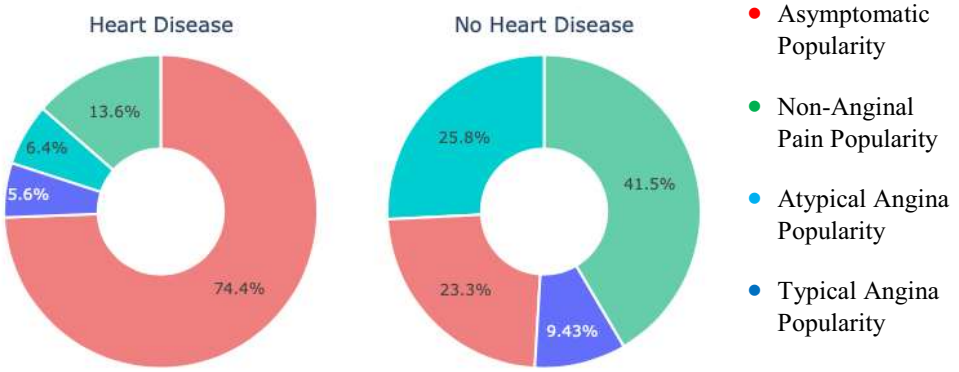


**Figure 5.** Shows the patients having or not having Heart Disease on the basis of type of Chest Pain.

## 6. Conclusion

A cardiovascular disease detection model has been developed using three ML classification modelling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression, Random Forest Classifier and KNN **[22].** The accuracy of our model is 87.5%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not **[9].** By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical

databases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors. Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying logistic regression and KNN to get an accuracy of an average of 87.5% on our model which is better than the previous models having an accuracy of 85%. Also, it is concluded that accuracy of KNN is highest between the three algorithms that we have used i.e. 88.52%. 'Figure 6' shows 44% of people that are listed in the dataset are suffering from Heart Disease.

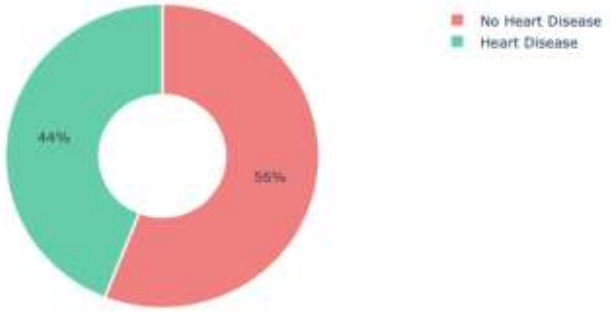


**Figure 6.** Shows the total number of patients having or not having Heart Disease.

## 7. References

[1] Soni J, Ansari U, Sharma D & Soni S (2011). *Predictive data mining for medical diagnosis: an overview of heart disease prediction.* International Journal of Computer Applications, ***17*(8), 43-8**

[2] Dangare C S & Apte S S (2012). *Improved study of heart disease prediction system using data mining classification techniques*. *International Journal of Computer Applications*, ***47*(10), 44-8**.

[3] Ordonez C (2006). *Association rule discovery with the train and test approach for heart disease prediction*. *IEEE Transactions on Information Technology in Biomedicine*, *10*(2), **334-43**.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). *An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm*. *International Journal of Computer Science and Information Technologies*, **6(1)**, 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). *An ensemble-based decision support framework for intelligent heart disease diagnosis*. In *International Conference on Information Society (i-Society 2014)* (pp. 259-64). **IEEE**.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). *A coronary heart disease prediction model: the Korean Heart Study*. *BMJ open*, ***4*(5)**, e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). *Multilocus genetic risk scores for coronary heart disease prediction*. *Arteriosclerosis, thrombosis, and vascular biology*, ***33*(9)**, 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). *Heart disease prediction using lazy associative classification*. In *2013 International Mutli-Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s)* (pp. 40- 6). **IEEE**.

[9] Dangare Chaitrali S and Sulabha S Apte. *"Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications* 47.10 (2012): 44-8.

[10] Soni Jyoti. "Predictive data mining for medical diagnosis: *An overview of heart disease prediction*." *International Journal of Computer Applications* 17.8 (2011): 43-8.

[11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). *HDPS: Heart disease prediction system*. In *2011 Computing in Cardiology* (pp. 557-60). **IEEE**.

[12] Parthiban, Latha and R Subramanian. *"Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences* 3.3 (2008).

[13] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). *Wireless body area network for heart attack detection [Education Corner]*. *IEEE antennas and propagation magazine*, ***58*(5)**, 84-92.

[14] Patel S & Chauhan Y (2014). *Heart attack detection and medical attention using motion sensing device -kinect*. *International Journal of Scientific and Research Publications*, ***4*(1)**, 1-4.

[15] Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). *U.S. Patent No. 8,014,863*. Washington, DC: U.S. Patent and Trademark Office.

[16] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). *Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design*. In *2016 19th International Conference on Computer and Information Technology (ICCIT)* (pp. 299-303). **IEEE**.

[17] Buechler K F & McPherson P H (1999). *U.S. Patent No. 5,947,124*. Washington, DC: U.S. Patent and Trademark Office.

[18] Takci H (2018). *Improvement of heart attack prediction by the feature selection methods*. *Turkish Journal of Electrical Engineering & Computer Sciences*, ***26*(1)**, 1-10.

[19] Worthen W J, Evans S M, Winter S C & Balding D (2002). *U.S. Patent No. 6,432, 124*. Washington, DC: U.S. Patent and Trademark Office.

[20] Acharya U R, Fujita H, Oh S L, Hagiwara Y, Tan J H & Adam M (2017). *Application of deep*

*convolutional neural network for automated detection of myocardial infarction using ECG signals*. *Information Sciences*, **415**, 190-8.

[21] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). *Inpatient deaths from acute myocardial infarction*, 1982-92: analysis of data in the Nottingham heart attack register. *BMJ*, ***315*(7101)**, 159-64.

[22] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). *Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone*. *Current controlled trials in cardiovascular medicine*, ***3*(1)**, 10.

[23] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). *Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women*. *International journal of epidemiology*, ***18*(2)**, 361-7.

[24] Kiyasu J Y (1982). *U.S. Patent No. 4,338,396*. Washington, DC: U.S. Patent and Trademark Office.