EasyChair Preprint

№ 11904

Predicting Coronary Heart Disease through Machine Learning Algorithms

Savina Mariettou, Constantinos Koutsojannis and Vassilios Triantafillou

February 13, 2024

Predicting Coronary Heart Disease through Machine Learning Algorithms

Savina Mariettou [1], Constantinos Koutsojannis[2], Vassilios Triantafillou[3]

[1] University of Peloponnese, Electrical and Computer Engineering Department, Patras, Greece
[2] Health Physics & Computational Intelligence Lab, Physiotherapy Department, School of Health Rehabilitation Sciences, University of Patras, Patras, Greece
[3] Network Technologies and Digital Transformation λab, Electrical and Computer Engineering Dpt., University of Peloponnese. Patras, Greece
s.mariettou@go.uop.gr, {ckoutsog1, vasdimtriantaf}@gmail.com

Abstract. Machine learning has gained popularity in medical fields due to the increasing availability of health data and the improvement of machine learning algorithms. It can be used to create predictive models that diagnose diseases, predict disease progression, tailor treatment to individual patient needs, and improve the functioning of medical systems. The right use of data can have a positive impact on improving the quality of patient care, reducing healthcare costs, and creating tailored and effective medical approaches. The healthcare sector benefits greatly from the accurate interpretation of medical data as it contributes to the early prediction of diseases in patients. Early detection of a disease can help control the symptoms and provide the correct treatment. In our work, we analyzed actual measurements from the Framingham Heart Study and we created a medical database with 78001 records. Our ultimate goal is to develop an expert Artificial Intelligence system and an Artificial Neural Network that can predict the development of coronary heart disease by employing intelligent knowledge-mining algorithms. We have created two intelligent systems that predict the progression of coronary heart disease using machine learning algorithms such as Random Forest, Decision Trees and Neural Networks. In our experimental analysis, the Decision Tree and Neural Network achieved an accuracy of 90.08% and 84.56% respectively.

Keywords: Machine learning, Big Data, Coronary Heart Disease, Intelligent Algorithms, Neural Networks.

1 Introduction

Cardiovascular disease, including conditions such as heart disease and stroke, remains the leading cause of death worldwide. More than half a billion people worldwide are still affected by cardiovascular disease, resulting in 20.5 million deaths in 2021, nearly a third of all deaths worldwide. This marks an overall increase in the estimated 121 million deaths from cardiovascular disease [1]. The cardiovascular system is a transportation system, which consists of a muscular pump, the heart, and a network of blood

vessels that contain blood. The three components that make up the cardiovascular system are the blood, the heart, and the vessels. Its main function is the transport of water, oxygen, carbon dioxide, fuel for energy production, electrolytes, hormones, and metabolic products, in particular, the transport of gasses and nutrients and the removal of waste substances [2]. Specifically, the roles of the cardiovascular system are the transport of oxygen from the lungs to the rest of the body, carbon dioxide from the tissues to the lungs, nutrient transport, thermoregulation, defense mechanisms, endocrine system functions, and fetal development, depending on the continuous flow of blood pumped from the heart to the capillary networks, where the exchange between tissues and blood takes place [3].

Coronary heart disease is directly related to the coronary arteries, which are the blood vessels that supply oxygen and blood to the heart [4]. From a medical perspective, coronary heart disease is caused by the narrowing of the coronary arteries, leading to an imbalance between the functional demands of the heart and the ability of the coronary arteries to supply blood and oxygen. The variation in coronary mortality is quite wide. Its factors vary, some are socio-economic, classic risks such as hypertension, diabetes, lifestyle, and family history. However, we also have factors such as emotional stress or acute physical exercise that can cause coronary events. The main cardiac symptoms are chest pain and shortness of breath. Early detection of coronary artery disease is essential. Specifically in patient survival, in easier clinical interventions, in reducing treatment costs as complications can be avoided and expensive therapeutic interventions can be prevented as well as in taking immediate measures to deal with dangerous situations [5]. Medical diagnosis by its nature is a complex and imprecise cognitive process as it relies on multiple elements. It usually requires the cooperation of several medical specialties such as patient history, clinical examinations for any signs and symptoms of heart failure, imaging tests, and laboratory tests. Early diagnosis improves the outlook and reduces the risk of death or complications [6].

1.1 Technology Approaches

The continuous evolution of technology has allowed the development of new methodologies based on Artificial Intelligence and Machine Learning. Health problems nowadays have increased consequently this has led to an increase in the production of big data. Their proper use requires the development of an automatic system for disease prediction by developing machine learning algorithms that can work effectively despite the challenges that may appear in the datasets [7].

Artificial Intelligence is an aspect of computer science that deals with the simulation of human intelligence with the help of a computer [8]. In studying the exact definition of artificial intelligence, most classify it as a system that can learn to make predictions and operate semi-autonomously. Artificial intelligence tools are based on expert systems (expert systems) and algorithms where they can be classified, interpreted, and synthesized advice and explanations on the collected data [9]. Although Artificial Intelligence is primarily related to computer science, it is also related to various fields of science such as mathematics, cognition, philosophy, psychology, and biology, and has recently been integrated into the field of engineering [8].

Machine learning uses statistical methodologies such as regression modeling, Bayesian probability, and others to predict the classification of data subjects from a dataset. It uses techniques such as thresh olding (for images), feature extraction and pattern recognition, and using a statistical model for prediction. Her field develops learning modes such as supervised learning or unsupervised learning. Supervised Learning is considered a process where the model is trained and uses the new data to predict the results. It is meant to infer the same answers from information as a human would (Classification, Prediction). In Unsupervised Learning, the algorithm builds a model for a given set of inputs in the form of observations without knowing the desired outputs (Clustering). It can also be used to find new patterns in data by inputting a training dataset without human interpretations of said data [9]. Basic machine learning categorization algorithms [10] are as follows.

1. Categorized by Bayes
Simplistic Bayes categorizer (naive Bayes classifier)
2. K-nearest neighbors' categorizer (KNN)
3. Categorizer with decision tree
ID ID3 algorithm, C4.5 Algorithm
4. Artificial Neural Networks

The integration of neural networks in the medical field has been remarkable. They are excellent at solving highly complex problems where traditional algorithmic solutions are insufficient or too complex. They have been successfully applied in medicine in various fields such as drug development, patient diagnosis, and image analysis. They make significant contributions to key areas such as the detection of coronary artery disease and the processing of Electroencephalography (EEG) signals. They offer improved capabilities for data analysis, pattern recognition, and decision-making, which leads to advances in medical research, diagnosis, and patient care [11].

At this point, since we understand the complexity of developing predictive methods for the diagnosis, prevention, and treatment of cardiovascular diseases, the ultimate goal of this work is to create an Artificial Intelligence system, which predicts the development of coronary heart disease. The rest of the paper is organized as follows: The second section describes Material / Methods. In the third section, the Results are described. In the fourth section, we have the summary as well as the future work we would like to achieve.

2 Material / Methods

The research investigates performance analysis for predicting coronary heart disease. Our database, CHD_DB, is based on actual measurements from one of the most famous cardiovascular disease studies (six-year follow-up), the Framingham Heart Study. There are more than 10,000 records available that are related to the development of coronary heart disease (CHD). Our database consists of four training datasets (Train_A, X, Y, and Z) and one test dataset (test set) (Test) [13]. The four training datasets are

designed by the researchers to have different proportions of CHD cases and non-CHD cases. Specifically, the six-year research base states that the records of Train A have 6,500 CHD cases and 6,500 Non-CHD cases, Train X 6,500 CHD cases and 13,000 Non-CHD cases, Train Y includes 6,500 CHD cases and 585,000 Non-CHD cases, Train Z approximates to the original data 4,000 CHD cases and 4,000 Non-CHD cases. Statistical analysis methods were used to validate the datasets. Our data items, the factors are eight items and they will be analyzed for their association with coronary heart disease and it will output whether it is positive for developing coronary heart disease or not. Specifically, we have the characteristics for each of these sets (Table 1).

Table 1. Input - Output parameters

Input parameters	*Output parameters*
1	ID
2	Coronary heart disease, CHD (0=non-CHD cases; 1=CHD)
3	Cholesterol, TC
4	Systolic blood pressure, SBP
5	Diastolic blood pressure, DBP
6	Left ventricular hypertrophy, LVH (0=negative; 1=definite or positive)
7	National origin, ORIGIN (0=native-born; 1=foreign-born)
8	Education, EDUCATE (0=grade school or less; 1=high school, not graduate; 2=high school, graduate; 3=college or more)
9	Smoking habit, TABACCO (0=never smoked; 1=stopped; 2=cigar or pipe;3=tobacco(<20/day); 4=tobacco(20/day=<))
10	Drinking habit, ALCOHOL

In this paper, we will deal with the implementation of two expert systems. Using Python's Scikit-Learn library and for the classification work the Jupyter Python Notebook, the first system, we use machine learning algorithms like Decision Trees, Naive Baye, and Random Forest. The second system is the deep learning system, the Neural Networks. Before training the algorithm, we went through some assumptions and changes. The serial number (ID) was not included as a class because it relates to the patients included in the particular database, and if it was taken into account, it would modify the predictive performance of the system. The second characteristic (development of coronary heart disease, CHD) was not included as an input class but as an output class. Using the function pandas.get_dummies to convert categorical variables to dummy variables. Specifically, it was applied to education and smoking, resulting in 14 from 8 input characteristics. The ultimate purpose of this change was to be able to compare similar things with each other.

Consequently, our training dataset, observed in (Fig. 1), has 14 input characteristics and one output characteristic for our system.

Show dataset with dummy categorical variables

```
pandas.DataFrame(X).head()
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.552707	0.531646	0.473282	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.360610
1	0.461538	0.594937	0.389313	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.196949
2	0.450142	0.666667	0.671756	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.159501
3	0.817664	0.383966	0.679389	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.374480
4	0.632479	0.569620	0.595420	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.417476

Show dataset output

```
pandas.DataFrame(y).head()
```

	0
0	0.0
1	0.0
2	1.0
3	1.0
4	1.0

Fig. 1. Input - Output parameters from Python

3 Results

Our first intelligent system is the Decision Tree, after our tests with the rest of the systems (Naive Baye, and Random Forest) and having the best prediction. As is known, the Decision Tree is one of the most frequently and widely used supervised machine learning algorithms. From the observation of the four sets, it was found that the accuracy of the Train A set from the test data set was 60.29 %. Train X had a percentage of 60.75 %. Train Y had 55.45 %. Finally, Train Z had a percentage of 56.08 %. Note that all ensembles were trained with a tree size of max_depth = 18 and random_state = 3. Finally, let's add that after tests we noticed that the performance did not differ whether we set the tree depth from 3 to 10. Remarkably we wanted to do an additional check regarding the accuracy of the Decision Tree. We used the Random Forest algorithm because this algorithm creates Decision Trees on randomly selected data samples, takes predictions from each tree, and selects the best solution through voting. Also, the reason we used it was that in relevant research it was stated that Random Forest gives the best result among various algorithms such as Naive Bayes, J48, etc. [7]. So, the training was done on the same data set, we kept the default values for the parameters and the accuracy ranged in the same percentages with very small deviations.

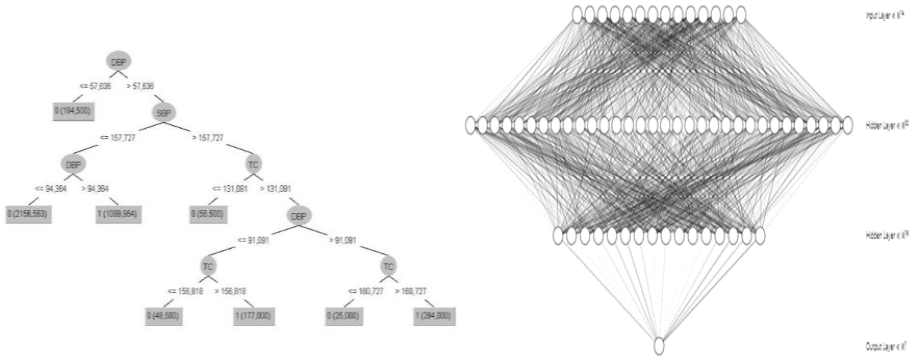

Fig. 2. Visualization of Decision Tree (left), Visualization of neural network (right)

Continuing with our second system, training the whole of us with our neural network produced the following results. Note as we can see in the image that we have the input neurons in the first level, and in the first hidden level, we have 32 neurons with an activation function: ReLU. In the second Hidden Level, we have 16 neurons with an activation function: ReLU. At the output level, we have 1 neuron as it is a binary classification problem (Fig. 2). We used the Activation function: Sigmoid (to output 0 to 1). Note that the activation functions (ReLU in the hidden layers and Sigmoid in the latter) help to introduce non-linearities in the model, while the Dropout layers help to avoid overfitting during training. The accuracy of the train A set from the test data set was 69.98 %. Train X had a rate of 73.82 %. Train Y had 90 %. Train Z had 90 %. Consequently, we understand that the accuracy of the Train A set is relatively the same as any supervised learning algorithm trained, but the percentages remain too small to be able to predict whether a patient is suffering from coronary heart disease. Then Train X, Y, and Z are already starting to notice how the initial modification of the sets by the researchers improves the percentage of the neural network but the accuracy of the Decision Tree remains constant (Fig. 3). It is worth noting that Train Y and Z seem to train our whole and display accuracy at 90%. So, observing this performance we went back to our original sets and combined the sets Train A and Train Y and Test and Train Y (figure 3). We observed that the accuracy rate is 84.56% in our neural network and 90.08% in the Decision Tree.

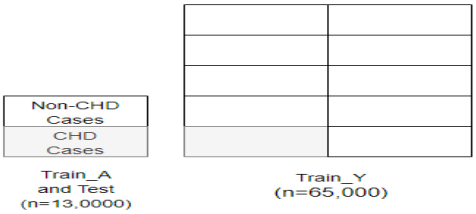

Fig. 3. Database structure. Figure adapted from Fig 3 in Ref. [13].

In Table 2, we compared the work done on [7, 12] with our proposed work. We could not find any article that trained the records using neural networks, so we did not make any additions to these fields. Upon analyzing the Decision Tree algorithm, we found that all the approaches showed similar accuracy rates with only a few deviations. We also observed that Python training brought shorter times compared to other methods.

Table 2. Experimental results

	Decision Tree	Calculation Time (sec)	Neural Networks	Calculation Taken (sec)
Proposed (Train AY)	90.08 %	**0.0111**	84.56 %	1.5067
Krishnani et al., 2019 [7]	92.45 %	0.8138	-	-
Rajliwall et al., 2017 [12]	90.00 %	77.4	-	-

4 Conclusions and future work

Few medical databases are available to researchers because it is impossible to distribute medical data without ensuring the privacy and confidentiality of the information they provide. The database that provides the data that we processed comes from the Framingham Heart Study for the purpose of evaluating prognostic systems [13]. This study is the first cardiovascular disease study, starting in 1948 under the direction of the National Heart Institute in the United States, participants were randomly selected from the city of Framingham, Massachusetts. The number of records and the ratio of CHD cases and non-CHD cases differ between the four training datasets. Two different forecasting systems were developed using the same data set. In other words, from the specific epidemiological data, that were validated using statistical analysis methods, we constructed two predictive systems using the training datasets, Train A, Train X, Train Y, and Train Z, and calculated their performance using the test dataset. By training our system, running the Scikit-Learn Python library, and for the classification task of the Jupyter Python Notebook, we ran enough tests to provide the optimal performance we recommend in this article. For the first system, we used the Decision Tree algorithm, which has an accuracy of 90.08% and a training time of 0.0111 sec. Your second system, which was trained with a neural network, has a prediction rate of 84.56% and a training time of 1.5067 sec. Finally, we should note that upon completing the creation and training of our systems, we compared our results with the work done in [7, 12], as shown in Table 2. The comparison focuses only on the results of the Decision Tree algorithm since no similar ones were found in studies with Neural Networks. In the Decision Tree algorithm, we observe that all approaches show relatively similar accuracy rates with few differences, and in terms of times, we observe that training with the Scikit-Learn Python library brings a shorter time.

Future work could consider combining data for different diseases. This approach would allow the detection of possible interactions between different diseases and the investigation of possible common risk factors. With this analysis, we could discover new information and gain a holistic understanding of the connections between health and disease. Finally, as an extension of this work, it's worthwhile to add to the existing data set factors from new research, that is, from national data to see the conclusions that will be drawn and how the performance will be modified. This particular study will bring benefits as it will test for risk factors that are common or different between the two populations such as the combination of different life contexts, dietary habits, genetic factors, and environmental influences.

References

1. World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland. World Heart Federation. 2023. https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf
2. Hodgson, R. D. et al., 2013, "Chapter 11 - The cardiovascular system: Anatomy, physiology, and adaptations to exercise and training", The Athletic Horse (Second Edition), Pages 162-173.

3. Mulroney, S. E., Myers, A. K., & Netter, F. H., 2009, “Netter's essential physiology”, Philadelphia, PA: Saunders/Elsevier
4. Felman A., 2019, “What to know about coronary heart disease”, Medical News Today, Medically reviewed by Debra Sullivan, Ph.D., MSN, R.N., CNE, COI από https://www.medicalnewstoday.com/articles/184130
5. Garavand, A., Behmanesh, A., Aslani, N., Hamidreza Sadeghsalehi, & Mustafa Ghaderzadeh. (2023). Towards Diagnostic Aided Systems in Coronary Artery Disease Detection: A Comprehensive Multiview Survey of the State of the Art. International Journal of Intelligent Systems, 2023, 1–19. https://doi.org/10.1155/2023/6442756
6. Bourazana, A., Xanthopoulos, A., Briasoulis, A., Magouliotis, D., Spiliopoulos, K., Athanasiou, T., Vassilopoulos, G., Skoularigis, J., & Triposkiadis, F. (2024). Artificial Intelligence in Heart Failure: Friend or Foe? Life, 14(1), 145. https://doi.org/10.3390/life14010145
7. Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019, October 1). Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms. IEEE Xplore. https://doi.org/10.1109/TENCON.2019.8929434
8. Lawal, A. I., & Kwon, S. (2021). Application of artificial intelligence to rock mechanics: An overview. Journal of Rock Mechanics and Geotechnical Engineering, 13(1), 248–266. https://doi.org/10.1016/j.jrmge.2020.05.010
9. Moxley-Wyles, B., Colling, R., & Verrill, C. (2020). Artificial intelligence in pathology: an overview. Diagnostic Histopathology, 26(11), 513–520. https://doi.org/10.1016/j.mpdhp.2020.08.004
10. Zaki, M. J., & Wagner Meira, J. (edited by: Megaloikonomou Vasileios, Makris Christos, translation: Stamou Giorgios), data mining and analysis: basic concepts and algorithms, Kleidarithmos, 2017
11. Hongmei, Y., Yingtao J. Zheng, J. Peng, C. Li, Q. (2006), “A multilayer perceptron-based medical decision support system for heart disease diagnosis”, Expert Systems with Applications, Volume 30, Issue 2, Pages 272-281
12. Rajliwall, N., Chetty, G., & Davey, R. (2017). Chronic Disease Risk Monitoring Based on an Innovative Predictive Modelling Framework: IEEE Symposium Series on Computational Intelligence 2017. 2017 IEEE Symposium Series on Computational Intelligence (SSCI), 1–8. https://doi.org/10.1109/SSCI.2017.8285257
13. Suka, M., Ichimura, T., & Yoshida, K. (2004). Development of Coronary Heart Disease Database. Lecture Notes in Computer Science, 1081–1088. https://doi.org/10.1007/978-3-540-30133-2_144