Machine Learning Improves Accuracy of Coronary Heart Disease Prediction

Savina Mariettou, Constantinos Koutsojannis and Vassilios Triantafillou

January 29, 2024

Machine Learning Improves Accuracy of Coronary Heart Disease Prediction

Savina Mariettou*, Constantinos Koutsojannis, Vassilios Triantafillou*****

**University of Peloponnese, Electrical and Computer Engineering Department, Patras, Greece*

***Professor of Medical Physics & Electrophysiology, Director of Health Physics & Computational Intelligence Laboratory, Physiotherapy Department, School of Health Rehabilitation Sciences, University of Patras, Patras, Greece*

****Professor of Network Technologies and Digital Transformation lab, Electrical and Computer Engineering Dpt., University of Peloponnese. Patras, Greece*

Abstract

In recent years, there has been a growing expectation to analyze the available data, and obviously with the contribution of machine learning. Machine learning is increasingly applied in medical specialties. The impact of using data properly can have a positive impact on improving people's lives. The health sector is of particular interest. Initially, it is because it has been proven that solutions are given to health care problems as well as with the accurate interpretation of medical data it contributes to the early prediction of a disease that the patient has. It is still possible to detect early signs of the disease, which can be useful in controlling the symptoms as well as in the correct treatment. Our work is based on the medical database with actual measurements from the Framingham Heart Study. The final set that was created with the help of the study has 78001 records. The ultimate goal of this work is, through the help of intelligent knowledge mining algorithms, to create an expert Artificial Intelligence system which predicts the development of coronary heart disease. We created two intelligent systems that predict the progression of coronary heart disease. Specifically, we trained machine learning algorithms such as Random Forest from Decision Trees, as well as 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

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 [World Heart Report, 2023]. 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 [Hodgson R. D., 2013]. 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, depends on the continuous flow of blood pumped from the heart to the capillary networks, where the exchange between tissues and blood takes place [S. Mulroney et al Myers, 2009].

Coronary heart disease is directly related to the coronary arteries, which are the blood vessels that supply oxygen and blood to the heart [Felman, 2019]. 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 [Garavand et al., 2023]. 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, laboratory tests. Early diagnosis improves the outlook and reduces the risk of death or complications [Bourazana et al., 2024].

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 the purpose of disease prediction by developing machine learning algorithms that can work effectively despite the challenges that may appear in the datasets [Krishnani et al., 2019]. Artificial Intelligence is an aspect of computer science that deals with the simulation of human intelligence with the help of a computer [Lawal & Kwon, 2021]. 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 [Moxley-Wyles et al., 2020]. 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 [Lawal & Kwon, 2021].

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 thresholding (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 [Moxley-Wyles et al., 2020]. Basic machine learning categorization algorithms [Zaki et al., 2017] 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*

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 fifth section we have the summary as well as the future work we would like to achieve.

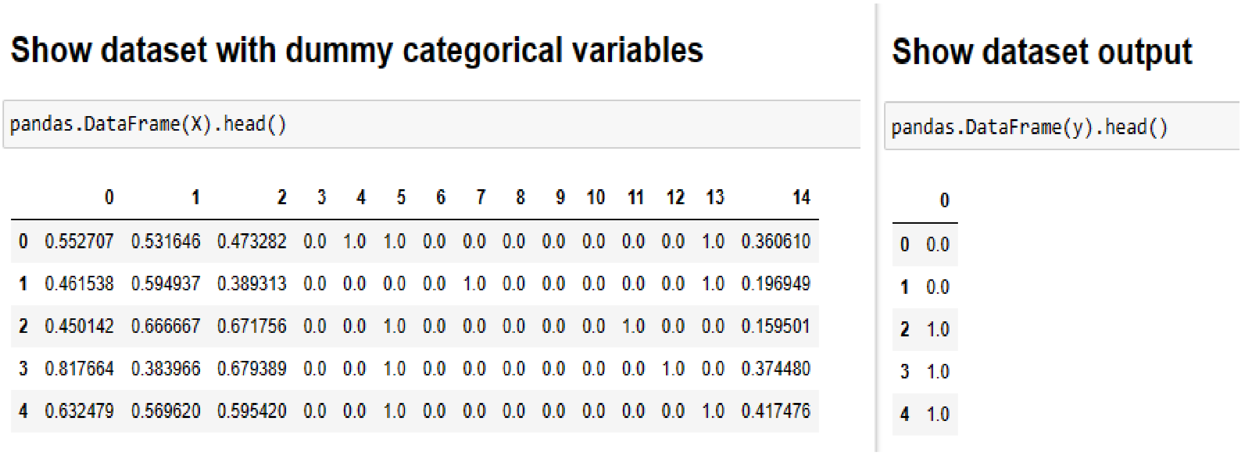
2. Material / Methods

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, the Framingham Heart Study. Belonging to the area of big data, it includes over 10,000 records 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). The four training datasets are designed by the researchers to have different proportions of CHD cases and non-CHD cases. Our data items, the factors are eight items and they will be analyzed for their association with coronary heart disease and then fed into an artificial intelligence tool 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).

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*

Table 1. Input - Output parameters

In this paper, we will deal with the implementation of two expert systems. Our tests were implemented using Python's Scikit-Learn library and the classification work was performed on the Jupyter Python Notebook. First, 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 (Fig 1). The ultimate purpose of this change was to be able to compare similar things with each other.



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

Figure 1. Input - Output parameters from Python

3. Results

Our first intelligent system is the Decision Tree. As is known, it 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 as in relevant research it was stated that Random Forest gives the best result among various algorithms such as Naive Bayes, J48, etc. [Krishnani et al., 2019]. 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.

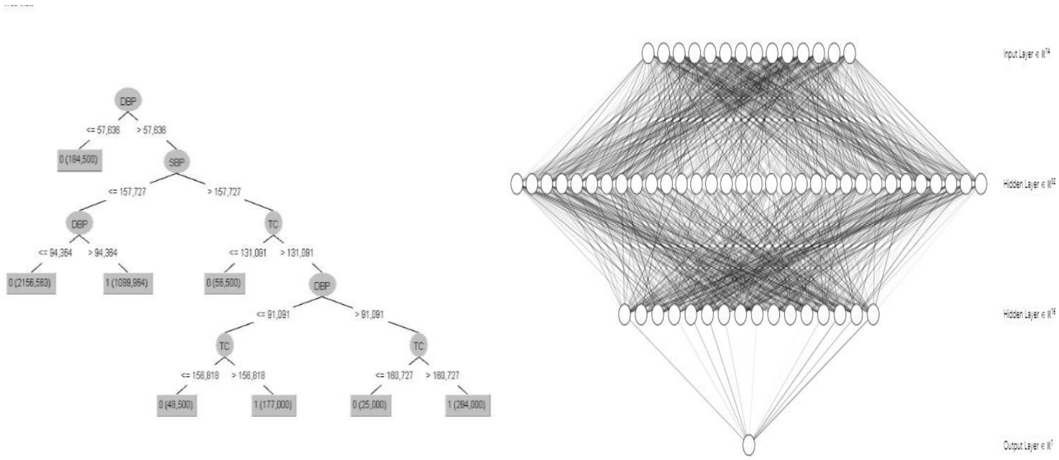


Figure 2. a. Visualization of Decision Tree, b. Visualization of neural network

Continuing with our second system, the training of the whole of us with our neural network produced the following results. Note as we can see in the image that in the first level we have the input neurons, in the first hidden level we have 32 neurons with an activation function: ReLU. In the second Hidden Level we have 16 neurons with 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 as well as 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.

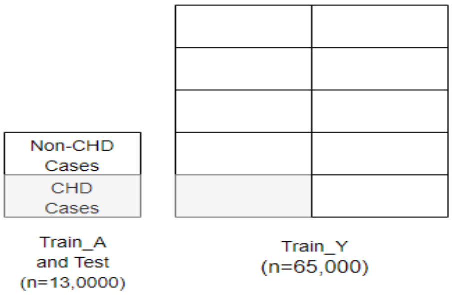


Figure 3. Database structure

Table 2 points to the comparison of work done in [Krishnani et al., 2019] and of our proposed work. We didn't find an article that trained these records with neural networks, so we didn't add anything to these fields. The Decision Tree algorithm we observe that all approaches show similar accuracy rates with few deviations and regarding the times we observe that python training brings shorter time.

	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	92.45 %	0.8138	-	-
Rajliwall et al., 2017	90.00 %	77.4	-	-

Table 2. Experimental results

4. Discussion

The database that provides the data and that we processed comes from the Framingham Heart Study. 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. 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. et al., 2004]. The researchers created a database of coronary heart disease (CHD_DB), which includes over 10,000 records related to the development of coronary heart disease (CHD). The database consists of four training data sets and one test data set (control set). The purpose is to apply certain data mining methods to these sets. The effectiveness of the data mining method should be evaluated using the test dataset with regard to the accuracy of the classification. The number of records and the ratio of CHD cases and non-CHD cases differ between the four training datasets. We should therefore consider which of the sets will be suitable for data mining and / or knowledge acquisition as well as we should develop a predictive system that outputs whether each record is a CHD case or a non-CHD case [Machi S. et al., 2004]. Two different forecasting systems were developed using the same data set. In other words, from the specific epidemiological data we constructed two predictive systems using the training datasets and calculated their performance using the test dataset. At an early stage the performance of Train_a, Train_X, Train_Y and Train_Z sets fluctuates in an accuracy rate

5. Conclusions and future work

Accurately predicting the presence or absence of heart disease using machine learning is central to the healthcare industry. To achieve this goal, we take advantage of machine learning techniques and advanced prediction systems. The use of these systems is expected to contribute to health promotion and early diagnosis of possible heart diseases. 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 in order 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

Journals

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

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

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

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

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

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

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

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

Books

Hodgson, R. D. et al., 2013, "Chapter 11 - The cardiovascular system: Anatomy, physi-ology, and adaptations to exercise and training", The Athletic Horse (Second Edition), Pages 162-173.

Mulroney, S. E., Myers, A. K., & Netter, F. H., 2009, "Netter's essential physiology", Philadelphia, PA: Saunders/Elsevier.

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

Reports

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