# Heart Disease Prediction Using Logistic Regression Machine Learning Model

**Conference Paper** *in* IFMBE proceedings · January 2024

DOI: 10.1007/978-3-031-49062-0_69



**3 authors:**

Faris Hrvat
Humanitas University


Lemana Spahić
Research Institute Verlab


Amina Aleta
Humanitas University

# Machine learning regression model for prediction risk of heart diseases

Faris Hrvat[1] and Lemana Spahić[2] Amina Aleta[3]

[1] Dept.of Endocrine and Metabolic Diseases and Lab. of Endocrine and Metabolic Research, IRCCS Istituto Auxologico Italiano, Milan, Italy
[3] Dept. of Medical Biotechnology and Translational Medicine, University of Milan, Milan, Italy,
aminaaleta@gmail.com

**Abstract.** Heart disease is a significant global health issue responsible for millions of deaths annually. Modifiable risk factors such as high cholesterol, smoking, physical inactivity, and high blood pressure can be tackled through lifestyle changes and medical interventions. Machine learning and artificial intelligence have the potential to enhance disease prediction and management, leading to better outcomes for patients. Python, along with its libraries such as Scikit-Learn and TensorFlow, provides a versatile platform for developing and deploying machine learning models.

In this study, the logistic regression model from Scikit-Learn was employed to predict the likelihood of cardiovascular disease based on various risk factors. The Pickle library was used to store the trained model for future use. Future research could enhance the model's efficacy by developing a user-friendly graphical user interface, implementing more advanced machine learning techniques, expanding the database used for training, and incorporating additional risk factors.

The study demonstrates the potential of machine learning models for predicting and managing cardiovascular disease, highlighting the need for further research and development to improve accuracy, applicability, and clinical utility. The development of effective prevention and treatment strategies using advanced technologies such as AI and machine learning is crucial in reducing the impact of this disease on individuals and society as a whole.

**Keywords:** Heart disease, risk prediction, Artificial Intelligence, Python, Logistic regression.

## 1 Introduction

### 1.1 Heart diseases

Heart disease encompasses a range of conditions that impact the structure and function of the heart, resulting in decreased blood flow, heart attacks, and other cardiovascular events. Despite significant progress in the fields of medicine and healthcare, heart disease continues to be the primary cause of mortality worldwide. In the United States, cardiovascular disease, which includes stroke, is responsible for the majority of illnesses and fatalities. It affects approximately 62 million individuals, with 50 million people having high blood pressure (Nabel, 2003). In the year 2000, cardiovascular disease caused nearly one million deaths in the United States, accounting for 39% of all fatalities. Many researches has demonstrated that coronary heart disease can be primarily preventable, although genetics may also play a role in the development of the disease. Understanding the root causes and mechanisms of heart disease is vital for the development of effective prevention and treatment strategies (Benjamin et al., 2019), (Nabel, 2003).

The heart's primary function is to circulate blood throughout the body, and any conditions that impede this process are referred to as heart disease. There are various types of heart disease, with the most prevalent ones being coronary artery disease (CAD) and heart failure (HF) (Jilani et al., 2021). CAD occurs when the arteries that supply blood to the heart become narrowed or blocked due to plaque accumulation. Plaque is a substance composed of cholesterol, fat, calcium, and other materials that can gather on the inner walls of arteries (Nabel, 2003). Over time, this buildup can lead to a reduction in blood flow to the heart, causing chest pain (angina), heart attacks, and other cardiovascular complications. HF arises when the heart is unable to efficiently pump blood to meet the body's needs. This can occur due to weakened or stiffened heart muscle or damage to the heart valves. HF can be triggered by various factors, such as CAD, high blood pressure, diabetes, and congenital heart defects. Heart failure risk factors can be divided into two categories: non-modifiable factors such as family history, sex, and age, and modifiable factors, including high cholesterol, smoking, physical inactivity, and high blood pressure (Benjamin et al., 2019), (Jilani et al., 2021).

## 1.2 Machine Learning use in medical field

Machine learning is a specialized area of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to learn from data and make predictions or decisions (Esteva et al., 2019). The key concept of machine learning involves the ability of computers to automatically identify patterns in large datasets and use these patterns to make predictions or decisions. Machine learning has diverse applications in healthcare, particularly in disease prediction and management (Esteva et al., 2019). For example, researchers have successfully used machine learning algorithms to develop models that accurately predict the risk of developing several diseases such as diabetes, cancer, and cardiovascular diseases. These models utilize various risk factors, including age, sex, family history, lifestyle, and biomarkers, to identify patterns and predict disease risk (Rajkomar et al., 2019). Additionally, machine learning algorithms can aid in improving disease management by identifying patients who are most likely to respond to particular treatments. By analyzing patient data, such as genetic profiles, disease history, and treatment outcomes, personalized treatment plans can be developed (Rajkomar et al., 2019).

Machine learning's growing popularity in healthcare is due to its potential to enhance disease prediction and management. Several studies have demonstrated that machine learning algorithms outperform traditional statistical models in predicting disease risk and treatment outcomes (Krittanawong et al., 2019). With the availability of open-source tools and platforms for developing and deploying machine learning models, it is becoming more accessible. It has the potential to revolutionize healthcare by leveraging large datasets and advanced algorithms to identify patterns and make predictions or decisions that can ultimately improve patient outcomes (Krittanawong et al., 2019).. To address the problem of providing quality care to patients, the use of artificial intelligence (AI) tools and machine learning can be explored. AI has already been effectively utilized in healthcare for disease diagnosis and treatment, as well as other applications. Expert systems, powered by AI, can solve problems in specific domains that typically require human expertise. Implementing such systems could be particularly advantageous in remote and rural areas as well as city areas where medical professionals with may not be immediately available (Aleta et al., 2020). Such system have also be tested in different fields such as device performance in medical field, and enhancing the safety and precision of patients' diagnosis and treatment is of utmost significance (Hrvat et al., 2020). While the field is relatively new, it is gaining traction in healthcare research and clinical practice (Krittanawong et al., 2019).

## 2 Methods

Through method section we explain how we developed model for predicting risk for developing heart diseases, dividing each section as we divided part of our model, with presentation of our model as last part of this section.

### 2.1 Programming language and libraries used

The programming Language that we used for developing this model is Python, it is a widely used programming language for machine learning due to its simplicity, flexibility, and availability of various libraries and tools for data analysis and visualization. Python allows for easy integration of various machine learning algorithms and is supported by popular open-source libraries such as Scikit-learn and TensorFlow.

Logistic Regression Model: Logistic Regression is a popular statistical method used for analyzing data and building predictive models. The model is used to estimate the probability of a binary outcome (i.e., a binary response variable) based on one or more predictor variables. The outcome variable can take only two possible values, such as success or failure, healthy or diseased, or yes or no. Logistic regression is widely used in various fields of research, including medical, social, and biological sciences, for its simplicity, interpretability, and efficiency. One of the main advantages of logistic regression is its ability to model the probability of an event occurring based on multiple predictor variables (Zou & Hastie, 2005). This makes it useful for predicting the occurrence of a disease or health condition based on various risk factors or biomarkers. For instance, logistic regression can be used to predict the risk of developing cancer based on age, sex, lifestyle habits, genetic factors, and other clinical variables. It can also be used to predict the outcome of a medical treatment based on the patient's characteristics and medical history (Cucchiara, 2019). Logistic regression also allows for the identification of significant predictors and their relative contributions to the outcome variable. In our study, we used the logistic regression model to predict the likelihood of cardiovascular disease based on several risk factors, including age, sex, blood pressure, chest pain type, and cholesterol levels. This model is widely used in healthcare applications due to its simplicity and interpretability (Cucchiara, 2019).

Pickle Library: Pickle is a Python library used for serializing and deserializing Python objects. In our study, we used the pickle library to save our trained logistic regression model as a file. This allowed us to reuse the model in the future without the need to retrain it every time. Which in the end, allowed next user ease of use model for prediction without training model (Pickle, n.d.).

Scikit-Learn Library: Scikit-Learn is a popular open-source library for machine learning in Python. We utilized the Scikit-Learn library for several tasks in our study, including data preprocessing, feature selection, model training, and evaluation. The library provides a wide range of machine learning algorithms and tools, making it easy to develop and deploy machine learning models (Pedragosa, et.al., 2011).

## 2.2 Data Used

The dataset available at Kaggle.com titled "Predicting Heart Disease Risk Using Clinical Variables", the dataset comprises 13 independent predictive variables and information from 270 patients. To make the most of this dataset, it is recommended to begin by comprehending the attributes of the columns and their significance in forecasting the risk of heart disease. The attributes from which dataset is consisted are Age, Sex, Chest Pain Type, BP (blood pressure), Cholesterol level, FBS over 120 (fasting blood sugar), EKG Results (electrocardiogram results), Max HR (maximum heart rate), Exercise Angina status, ST Depression ( depression of ST segment on ECG) , Slope of ST(slope of the ST segment on the EKG), Number of Vessels Fluoroscopy (number of vessels seen on fluoroscopy) and Thallium Stress test results. Dataset is available freely for purpose of this model thanks to Robert Hoyt MD.

**Table 1.** Dataset variables used explanation

| Data type | Explanation of variable |
|---|---|

| | |
|---|---|
| Age | The risk of developing heart disease increases as age advances, particularly for men above the age of 45 and women above the age of 55. As per the American Heart Association (AHA), the risk of heart disease doubles with each decade of life after 55 years of age for women and after 45 years of age for men (Benjamin et al., 2019). |
| Sex | According to the American Heart Association (AHA), men have a higher risk of developing heart disease than women.. However, women are at an increased risk after menopause, and the risk becomes similar to men's risk (Benjamin et al., 2019). |
| Chest Pain Type | Chest pain is one of the common symptoms of heart disease, and its type can be used to identify the type of heart disease. There are generally four types of chest pain associated with heart disease:<br>1. Stable angina: discomfort or pressure in the chest during physical activity or stress, which usually goes away with rest or medication.<br>2. Unstable angina: unexpected pain or pressure in the chest that occurs at rest or with minimal exertion, which is a sign of an impending heart attack.<br>3. Variant (Prinzmetal's) angina: sudden chest pain caused by a spasm in a coronary artery that occurs when resting, usually at night.<br>4. Non-cardiac chest pain: pain in the chest that is not caused by heart disease, but rather by other factors such as gastroesophageal reflux disease (GERD), anxiety, or muscle strain.<br>The presence of chest pain, along with its type, can help in determining the likelihood of heart disease (American Heart Association , 2022). |
| Blood Pressure | Blood pressure is an important factor in determining the risk of developing heart disease. High blood pressure, also known as hypertension, can damage the arteries and make the heart work harder, increasing the risk of heart disease, heart attack, and stroke (Lewington,S., et.al., 2002). |
| Cholesterol Level | Cholesterol is a type of fat that is essential for the proper functioning of the body. However, high levels of cholesterol in the blood can increase the risk of developing heart disease by causing the buildup of plaque in the arteries, which can lead to a heart attack or stroke (Baigent, C., et.al., 2005). |

| | |
|---|---|
| FBS over 120 | Elevated fasting blood sugar levels have been linked to an increased risk of heart disease. A study published in the New England Journal of Medicine in 2011 found that individuals with impaired fasting glucose levels (fasting blood sugar levels between 100 and 125 mg/dL) had an increased risk of cardiovascular disease compared to those with normal fasting blood sugar levels (<100 mg/dL) (McDermott & Criqui, 2016). |
| EKG results | Electrocardiogram (EKG or ECG) results can be useful in diagnosing and evaluating heart disease. Abnormal EKG findings can indicate a variety of heart conditions, including ischemia, arrhythmias, and hypertrophy (Sheikh Beig Goharrizi et al., 2023). |
| Maximum HR | Several studies have shown that maximum heart rate (HR) during exercise testing is a strong predictor of cardiovascular mortality and morbidity. One such study conducted by researchers at the University of Oslo found that maximal HR was an independent predictor of mortality and major cardiovascular events in both men and women with suspected coronary artery disease (Haugaa et al., 2012). |
| Exercise Angina Status | Exercise-induced angina, which is chest pain or discomfort that occurs during physical activity or exertion, is a known risk factor for heart disease. In a study published in the Journal of the American College of Cardiology, it was found that patients with exercise-induced angina had a significantly higher risk of cardiovascular events such as myocardial infarction (heart attack) and death compared to those without angina (Cavender et al., 2018). |
| ST Depression | ST depression, which is a common EKG finding, is a significant predictor of coronary artery disease and is associated with an increased risk of myocardial infarction and cardiac death. ST segment depression during exercise stress testing has been shown to be a sensitive marker for coronary artery disease, especially in patients with an intermediate pretest probability of disease (Gibbons et al., 2002). |
| Slope of ST | The slope of the ST segment on an electrocardiogram (ECG) is one of the measurements used to assess the |

| | |
|---|---|
| | severity of heart disease. Avarious studies that were diag-nosing with these guidelines found that a downsloping ST segment was associated with an increased risk of cardiovascular events, while an upsloping ST segment was associated with a lower risk (Gibbons et al., 2002). |
| Number of vessels on Fluoroscopy | The number of vessels seen on Fluoroscopy is a measure of the severity of coronary artery disease, with a higher number indicating a greater degree of occlusion. This is because coronary artery disease is characterized by the build-up of plaque within the coronary arteries, which can lead to the narrowing or complete blockage of these vessels (Saeed,M., et.al., 2012). |
| Thallium Stress Test | The thallium stress test is a nuclear medicine imaging technique used to assess blood flow to the heart muscle. It involves administering a small amount of radioactive thallium to the patient, which is taken up by the heart muscle. The patient then exercises on a treadmill or bicycle to increase the workload on the heart, and images of the heart are taken before and after exercise to assess blood flow to the heart muscle (Gibbons et al., 2002). |

### 2.3 Model for predicting risk of a patient having heart disease

Model that we developed, implements a logistic regression model for predicting the risk of heart disease based on various patient characteristics. The code first loads the heart disease dataset and defines the target variable and features. It then splits the data into training and testing sets and uses grid search to find the best hyperparameters for the logistic regression model. The code then fits the model on the training data and evaluates its performance on the test data. The trained model is then saved to disk using the pickle library. The code provides an option for the user to either train a new model or make predictions with the saved model. If the user chooses to make predictions, they are prompted to input various patient characteristics, and the saved model is used to predict the patient's risk of heart disease. In this code, a logistic regression model is built to predict the risk of heart disease based on several patient characteristics. The model is trained on a dataset that contains information 270 patients which was available freely for purpose of this model thanks to Robert Hoyt MD., and it is evaluated using a test set that contains 20% of the data. To build the model, several Python libraries are used. The pandas library is used to load and manipulate the dataset, while the scikit-learn library is used to build and evaluate the logistic regression model. Specifically, the following scikit-learn modules are imported: LogisticRegression, accuracy_score, GridSearchCV, and train_test_split. Finally, the pickle library is used to save and load the trained model. In addition to the model itself, hyperparameter tuning is used to optimize the performance of the logistic regression algorithm. Hyperparameters are adjustable parameters that are not learned during training, but are set prior to training. In this code, the hyperparameters that are tuned include the inverse regularization strength (C), the maximum number of iterations for the solver to converge (max_iter), and the algorithm used to solve the optimization problem (solver). The GridSearchCV function is used to perform a grid search over a range of hyperparameters to find the best combination of hyperparameters for the model. Specifically, the hyperparameter grid that is searched over includes five values for C (0.01, 0.1, 1, 10, and 100), six values for max_iter (100, 250, 500, 750, 1000, and 10000), and three values for solver ('lbfgs', 'liblinear', and 'saga').

# 3 Results

## 3.1 Correlation of the data in dataset

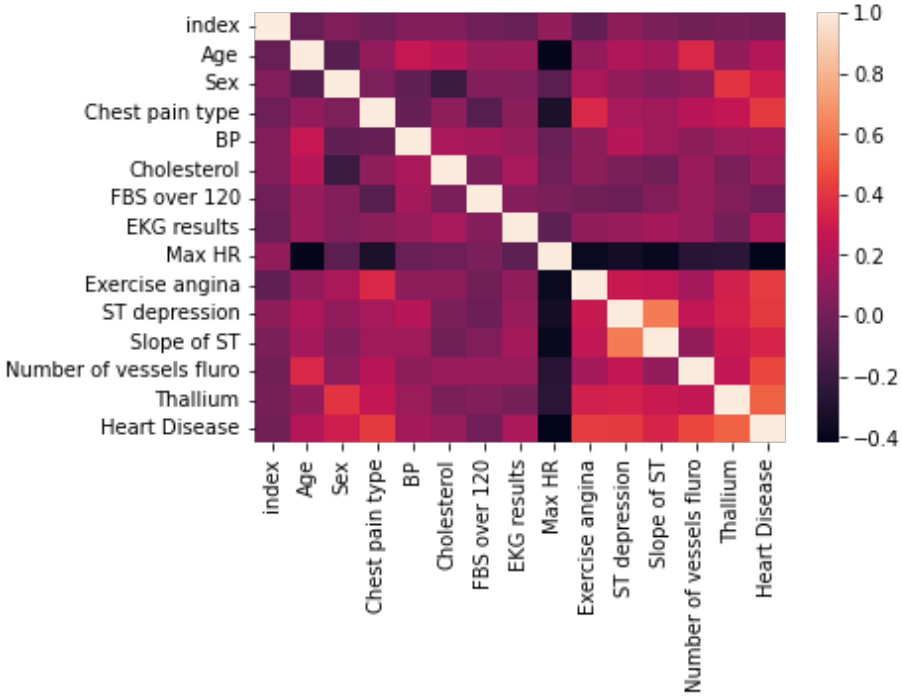


**Fig. 1.** Heatmap of correlation of data inside the dataset

The correlation heatmap indicates a relatively low correlation between all predictor variables, with the lowest correlation being observed between maximum heart rate and age, the slope of ST, exercise angina, ST depression, Number of vessels observed by fluoroscopy, and thallium. The correlations of individual variables with heart disease are also not pronounced, but the highest correlation is observed for thallium, a number of vessels observed in fluoroscopy, ST depression, exercise angina, and chest pain type. The highest correlation observed between individual predictor variables was observed between ST depression and slope of ST, a number of vessels observed with fluoroscopy and age as well as thallium and sex.

### 3.2 ROC of a model

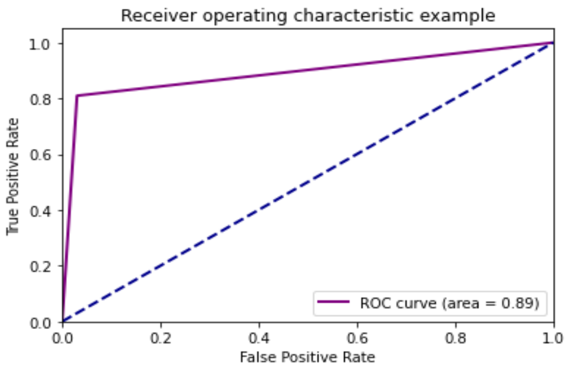


**Fig. 2.** Representation of ROC for a given model

The AUC indicates the accuracy of the model of 0.89. The sharp elbow of the ROC curve is due to the binary separation of classes, and this indicates near-perfect separation.

### 3.3 Relationship between variables

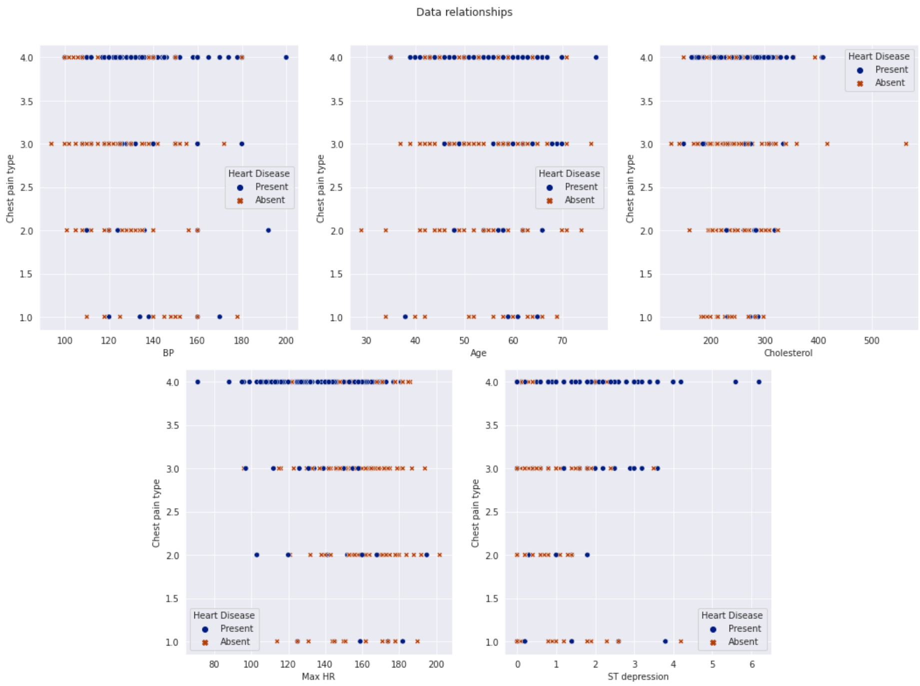


**Fig. 3.** Relationships between individual continuous variables and chest pain type

The relationship between individual continuous variables and chest pain type, which is the most common observable symptom of heart disease, was evaluated with respect to the presence and absence of heart disease. As can be seen, the increase in blood pressure in combination with the severity of chest pain was observed in a very significant subset of instances where heart disease was present. The same can be deduced for all other continuous variables such as age, cholesterol levels, maximum heart rate, and ST depression.

## 4 Conclusion

Heart disease continues to be a significant global health issue and is responsible for causing millions of deaths annually. Although genetics may contribute to the development of heart disease, controllable risk factors like high cholesterol, smoking, physical inactivity, and high blood pressure can be tackled by modifying one's lifestyle or through medical interventions. Machine learning and artificial intelligence (AI) have the potential to enhance disease prediction and management, leading to better outcomes for patients. Python, along with its widely used libraries such as Scikit-Learn and

TensorFlow, offers a versatile and powerful platform for developing and deploying machine learning models.

In this study, we employed the logistic regression model from Scikit-Learn to predict the likelihood of cardiovascular disease based on various risk factors. We also utilized the Pickle library to store the trained model for future use. By comprehending the underlying mechanisms and causes of heart disease and leveraging advanced technologies like AI, we can work towards effective prevention and treatment strategies to reduce the impact of this disease on individuals and society as a whole.

However, there are several avenues for future research and development that could enhance the efficacy of the model and its applications.
One potential area for future research is the development of a user-friendly graphical user interface (GUI) for the model. This would enable healthcare professionals to easily input patient data and obtain risk predictions, allowing for quicker and more efficient diagnosis and treatment. Another avenue for further development is the implementation of more advanced machine learning techniques, such as deep learning and neural networks. These approaches may provide more accurate predictions and improve the overall performance of the model. Expanding the database used for training the model could also improve its accuracy and applicability to different populations. Including data from diverse patient populations, particularly those underrepresented in current research, could enhance the generalizability of the model.

In addition, further research could explore the incorporation of additional risk factors into the model. This would provide a more comprehensive assessment of cardiovascular disease risk and enable targeted interventions for individuals with specific risk profiles. Overall, the present study demonstrates the potential of machine learning models for predicting and managing cardiovascular disease. Future research and development could lead to improved accuracy, broader applicability, and enhanced clinical utility, ultimately leading to better outcomes for patients.

## References

1. Aleta, A., Džuho, A., & Hrvat, F. (2020). Detection of acute inflammation of urinary bladder and acute nephritis of renal pelvis origin using artificial neural network. *8th European Medical and Biological Engineering Conference*, 363–371. https://doi.org/10.1007/978-3-030-64610-3_42
2. American Heart Association. (2022, December 5). *Angina (chest pain)*. www.heart.org. Retrieved February 25, 2023, from https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain
3. Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Jordan, L. C., Khan, S. S., Kissela, B. M., Knutson, K. L., … Virani, S. S. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. *Circulation*, *139*(10). https://doi.org/10.1161/cir.0000000000000659

4. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in Healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z
5. Haugaa, K. H., Goebel, B., Dahlslett, T., Meyer, K., Jung, C., Lauten, A., Figulla, H. R., Poerner, T. C., & Edvardsen, T. (2012). Risk assessment of ventricular arrhythmias in patients with nonischemic dilated cardiomyopathy by strain echocardiography. *Journal of the American Society of Echocardiography*, *25*(6), 667–673. https://doi.org/10.1016/j.echo.2012.02.004
6. Hrvat, F., Spahic, L., Pokvic, L. G., & Badnjevic, A. (2020). Artificial Neural Networks for prediction of Medical Device Performance Based on Conformity Assessment Data: Infusion and perfusor pumps case study. *2020 9th Mediterranean Conference on Embedded Computing (MECO)*. https://doi.org/10.1109/meco49872.2020.9134359
7. Jilani, M. H., Javed, Z., Yahya, T., Valero-Elizondo, J., Khan, S. U., Kash, B., Blankstein, R., Virani, S. S., Blaha, M. J., Dubey, P., Hyder, A. A., Vahidy, F. S., Cainzos-Achirica, M., & Nasir, K. (2021). Social determinants of health and cardiovascular disease: Current State and future directions towards healthcare equity. *Current Atherosclerosis Reports*, *23*(9). https://doi.org/10.1007/s11883-021-00949-w
8. Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., Min, J. K., Tang, W. H., Halperin, J. L., & Narayan, S. M. (2019). Deep Learning for Cardiovascular Medicine: A Practical Primer. *European Heart Journal*, *40*(25), 2058–2073. https://doi.org/10.1093/eurheartj/ehz056
9. McDermott, M. M. G., & Criqui, M. H. (2016). Reducing cardiovascular risk with ankle brachial index screening. *Journal of the American College of Cardiology*, *67*(6), 641–643. https://doi.org/10.1016/j.jacc.2015.12.018
10. Nabel, E. G. (2003). Cardiovascular disease. *New England Journal of Medicine*, *349*(1), 60–72. https://doi.org/10.1056/nejmra035098
11. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/nejmra1814259

Sheikh Beig Goharrizi, M. A., Teimourpour, A., Falah, M.,Hushmandi, K., & Isfeedvajani, M. S. (2023). Multi-lead ECG Heartbeat Classification of heart disease based on Hog Local Feature descriptor. *Computer Methods and Programs in Biomedicine Update*, *3*, 100093. https://doi.org/10.1016/j.cmpbup.2023.100093

12. Lewington, S., Clarke, R., Qizilbash, N., Peto, R., Collins, R., & Prospective Studies Collaboration (2002). Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet (London, England)*, *360*(9349), 1903–1913. https://doi.org/10.1016/s0140-6736(02)11911-8
13. Baigent, C., Keech, A., Kearney, P. M., Blackwell, L., Buck, G., Pollicino, C., ... & Collins, R. (2005). Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. The Lancet, 366(9493), 1267-1278.
14. Saeed, M., Hetts, S. W., English, J., & Wilson, M. (2012). MR fluoroscopy in vascular and cardiac interventions (review). *The international journal of cardiovascular imaging*, *28*(1), 117–137. https://doi.org/10.1007/s10554-010-9774-1
15. Gibbons, R. J., Balady, G. J., Timothy Bricker, J., Chaitman, B. R., Fletcher, G. F., Froelicher, V. F., Mark, D. B., McCallister, B. D., Mooss, A. N., O'Reilly, M. G., Winters, W. L., Gibbons, R. J., Antman, E. M., Alpert, J. S., Faxon, D. P., Fuster, V., Gregoratos, G., Hiratzka, L. F., Jacobs, A. K., … Smith, S. C. (2002). ACC/AHA 2002 guideline Update for

exercise testing: Summary article. *Journal of the American College of Cardiology*, *40*(8), 1531–1540. https://doi.org/10.1016/s0735-1097(02)02164-2
16. Pickle. In: Python Software Foundation. Python Software Foundation. https://docs.python.org/3/library/pickle.html. Accessed February 23, 2023.
17.