INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 593-598

HEART DISEASE PREDICTION

Mayuri Wani[1], Sanskriti Balapurkar[2], Prof. Jinal Patel[3]

[1,2,3]Department of Computer Science & Engineering. Thakur ShivKumar Singh Memorial Engineering College, Burhanpur, (Affiliated to RGPV Bhopal), India.

ABSTRACT

The World Health Organization reports thatcardiovascular diseases, particularly heart disease,heart disease kills 17 million people every year, making it the number one reason people die everywhere in the world. This study introduces an innovative machine learning framework designed to predict heart disease susceptibility by analyzing comprehensive patient medical records and risk factors. By amalgamating a suite of advanced algorithms including decision trees, random forest, logistic regression our modelsystematically scrutinizes a diverse dataset to unveil patterns indicative of heart disease. Theresults suggest our approach holds promise as a potent tool for early identification of heart disease, potentially leading to improved patient outcomes through timely intervention and treatment.

1. INTRODUCTION

Worldwide, heart disease poses a significant challenge as a result of widespread inadequate nutrition and insufficient physical activity, culminating in compromised cardiac health among populations globally. Detecting it early is crucial for prevention.This paper proposes a machine learning-based model to predict heart disease risk using patient datasuch as demographics, medical history, and lifestylehabits. We aim to create an accurate model that canhelp healthcare providers identify at-risk patients and take preventive actions.

We use a real-world dataset and focus on logistic regression as our primary predictive algorithm. Thedataset includes 14 physical attributes, like blood samples and exercise test results, with the "goal" field indicating the presence (1) or absence (0) of heart disease. Our model aims to streamline diagnosis, saving time and resources by avoiding invasive procedure .By providing a reliable tool foridentifying at-risk individuals, our model seeks to reduce heart disease incidence and improve patient outcomes. Integration into clinical practice could optimize preventive strategies, ultimately enhancingthe lives of those predisposed to heart disease.

2. RELATED WORK

The field of heart disease prediction has seen extensive research, with studies employing diverse

machine learning techniques and datasets. Here's a summary of some notable works:

- Dey et al. (2020) proposed a model combining decision tree, logistic regression, and Naïve Bayes algorithmsfor heart disease prediction. Using data from the Cleveland Clinic Foundation Heart Disease database, they achieved an accuracy of 85.23%.
- Cheng et al. (2021) implemented a deeplearning model employing convolutional neural networks (CNNs) to predict heart disease. Their model, trained on the MIMIC-III database, attained an impressive area under the curve (AUC) of 0.902.
- Naeem et al. (2021) introduced a hybridalgorithm, blending particle swarm optimization and extreme learning machine, for heart disease prediction. Their model achieved an impressive accuracy of 90.84%.
- Zhu et al. (2020) proposed a hybrid model combining decision trees and gradient-boosting algorithms for heart disease prediction. Tested on the UCI Heart Disease dataset, their model achieved an accuracy of 86.07%.

Additionally, two systematic reviews offer insights into heart disease prediction methodologies:

- Alghamdi et al. (2019) conducted a review on coronary heart disease prediction, identifying decision trees and artificial neural networks as highly accurate methods.
- Sabri et al. (2020) reviewed data miningtechniques for heart disease prediction highlighting support vector machines and decision trees as particularly effective approaches

These studies collectively underscore the diversity of approaches in heart disease prediction, showcasing the efficacy of various machine learning and data mining techniques in advancing diagnostic accuracy and patient care.

The data utilized in this study was sourced from the Kaggle Heart Disease Prediction dataset, comprising information from 303 patients who underwent cardiac evaluations at the Cleveland Clinic Foundation in the United States between 1988 and 1990.

The dataset comprises 14 attributes alongside the target variable denoting the presence of heart disease. These attributes are:

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 593-598

e-ISSN : 2583-1062

Impact Factor: 5.725

- Age
- Chest Pain type
- Resting blood pressure
- Serum cholesterol levels
- Fasting blood sugar levels
- Electrocardiographic results
- Maximum heart rate achieved
- Exercise-induced angina
- Thalassemia.

Pre-processing involved the removal of missing values and the transformation of categorical attributes into binary values using one-hot encoding.The target variable was subsequently binary- encoded with values of 0 (no heart disease) or 1 (heart disease present). The dataset was partitioned into a training set and a testing set, adhering to a 70/30 split ratio.

3. DATA ANALYSIS

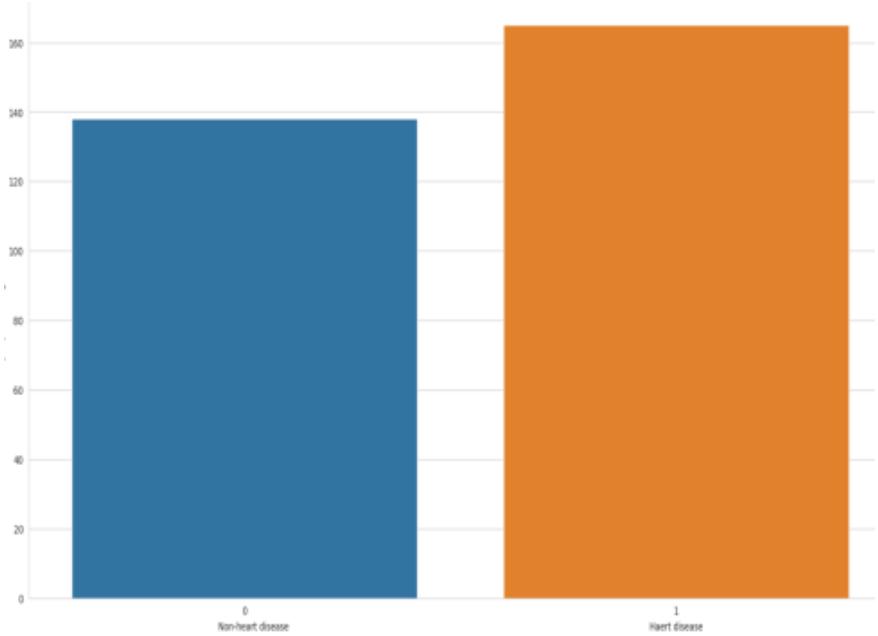

Fig. 1. Distribution of target variable

The information provided indicates that there are526 patients with heart disease and that the ratio of 1 to 0 is significantly less than 1.5. This maysuggest that the target feature is not imbalanced.

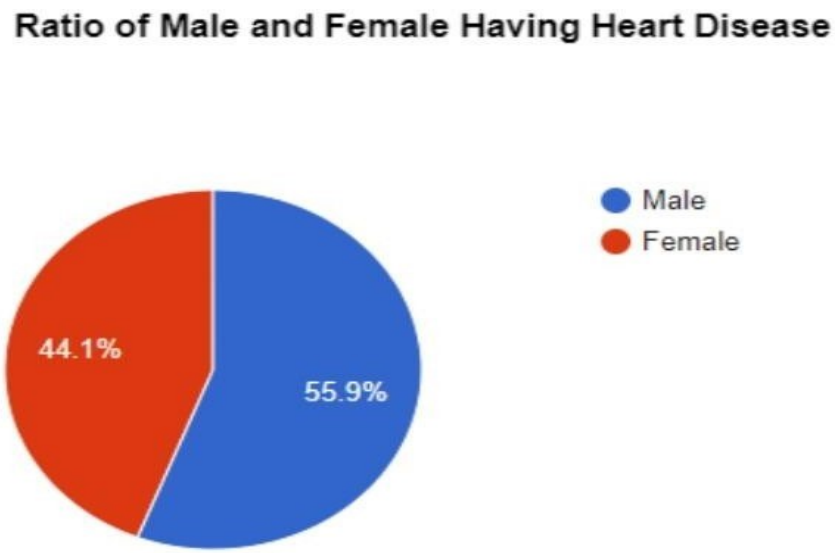

Fig 2. Ratio of Male and Female Patients

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 593-598

e-ISSN : 2583-1062

Impact Factor: 5.725

The ratio of male to female patients is approximately 2:1, which means that there areapproximately two male patients for every onefemale patient.

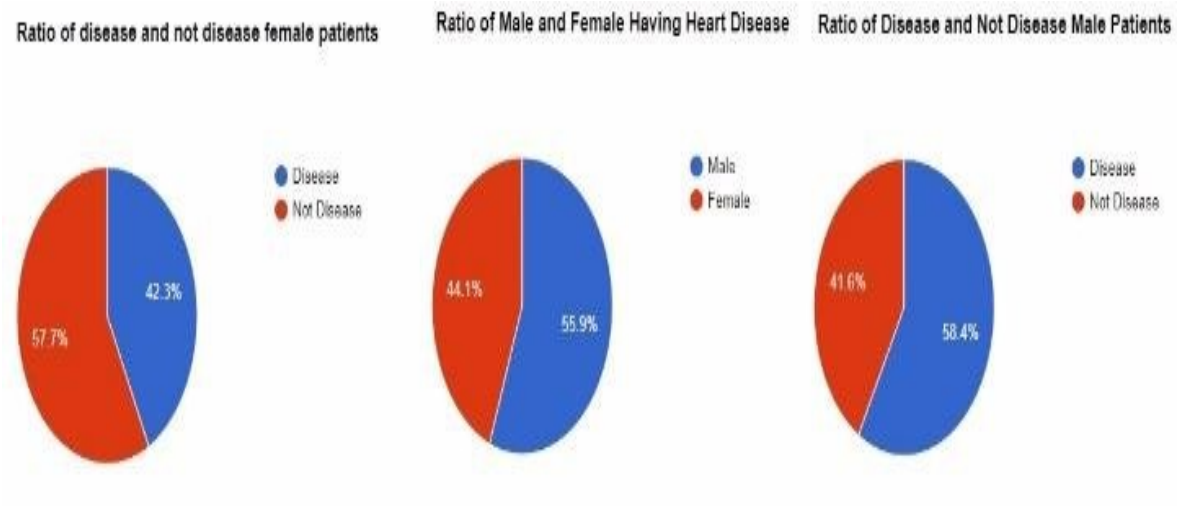

Fig 3. How sex is related with heart disease?

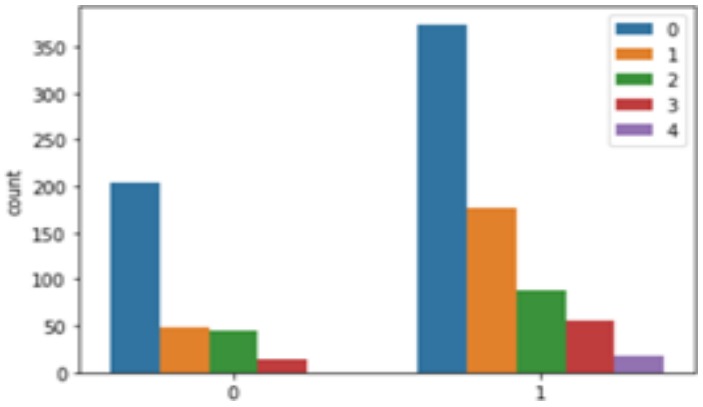

Fig 4. Is sex a risk factor for coronary artery disease.

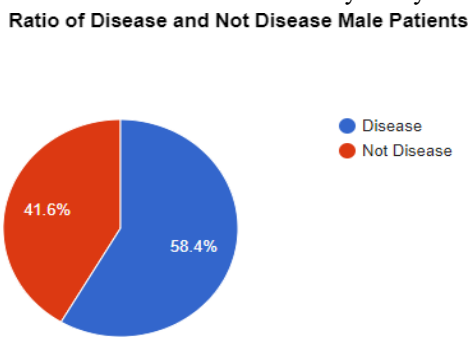

Fig 4.a. Male patients have a higher likelihood or propensity of developing coronary artery disease (CAD) than female patients.

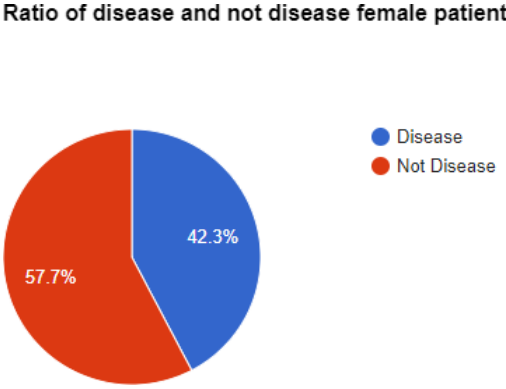

Fig 4.b. Female Patient has less heart problem than man

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 593-598

e-ISSN : 2583-1062

Impact Factor: 5.725

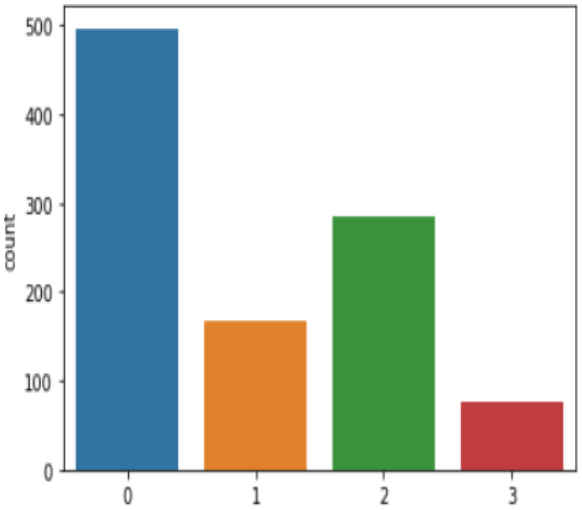

Fig 5. Different types of chest pain

These are the typical classifications of chest painrelated to cardiac issues:

- Typical Angina: Chest pain or discomfort occurring due toinadequate blood flow or oxygen to the heart, often during physical exertion or emotional stress.
- Atypical Angina: Chest discomfort that doesn't fit the typical pattern of angina but might still be related to heart issues. Symptoms can include shortness of breath, nausea, or fatigue.
- Non-Anginal Pain: Chest pain or discomfort not originating from the heart. This may include conditionslike esophageal spasms,musculoskeletal pain, or anxiety.
- Asymptomatic: No symptoms present, including chest pain, despite potential underlying heart issues.

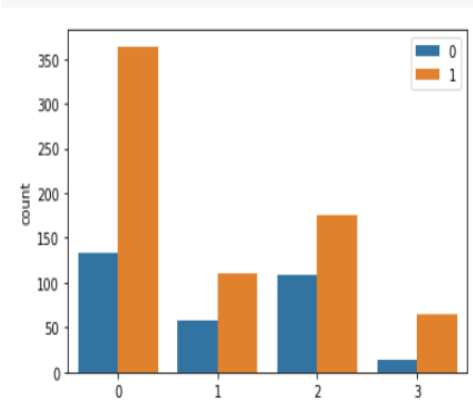

Fig 6. Who is more prone to chest pain male or female patients?

Male patients are more prone to chest pain, Most of the patients are having typical angina chest pain.

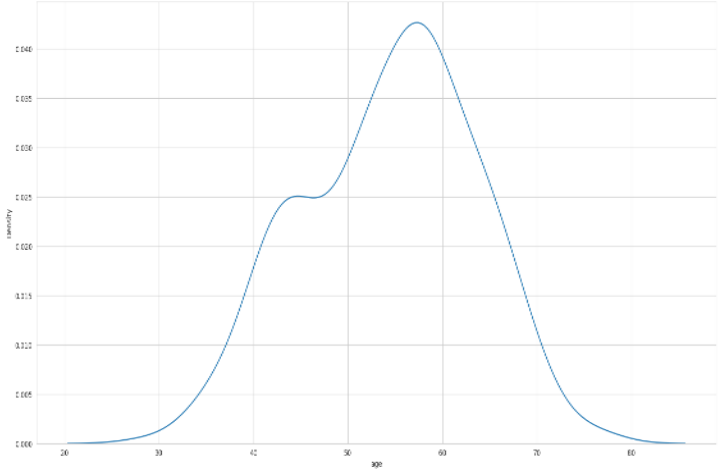

Fig 7. How Disease varies according to age

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 593-598

The data we've gathered indicates a significantuptick in the spread of the disease starting around the age of 33. This uptrend continues to escalate notably between the ages of approximately 43 to 46, suggesting a critical period of vulnerability. Subsequently, the disease reaches its zenith in terms of prevalence between roughly 50 to 58years of age, marking a phase where infections aremost rampant. However, there's a glimmer of hopeas the prevalence gradually declines thereafter. This observation underscores the importance of understanding age-related factors in disease transmission and devising targeted strategies tomitigate its impact across different age demographics.

4. APPROACH METHODOLOGY

- **Decision Tree**: Imagine you're tryingto make a decision based on a series of questions. A decision tree algorithm works similarly. It takes a dataset and asks questions about its features to splitit into smaller and smaller subsets. It keeps doing this until it either reaches astopping point (like a minimum number of data points in a subset) or until it can'tsplit the data anymore.
- **K-Means Clustering**: Let's say you have a bunch of points on a graph, and you want to group them into clusters based on their similarity. K-meansclustering helps you do just that.
- **Linear Regression:** Linear regression is like fitting a line through a scatter plot of data points. It's used when you want to understand the relationship between two variables
- **Random Forest:** Think of a forest made up of many different trees. Random forest combines the opinions of all these trees to give a final decision. It does this by letting each tree "vote" on the correct classification, and the most popular choice wins.
- **Support Vector Machine (SVM)**: SVM is a bit like drawing a line betweentwo groups of points on a graph. If the groups are clearly separable, SVM findsthe best line (or hyperplane in higher dimensions) that separates them with the largest margin. If they're not separable, it uses a trick to map the points.
- **Random Forest**: Think of a forest made up of many different trees. Each tree has its own opinion about what species a particular tree in the forest belongs to. Random forest combines the opinions of all these trees to give a final decision. It does this by letting each tree "vote" on the correct classification, and the most popular choice wins.
- **(PCA)**: PCA is like looking at acomplex painting and trying to describeit using only a few colors. It's atechnique used for simplifying the complexity of high-dimensional data byfinding the most important aspects of that data. It does this by identifying the directions (principal components) alongwhich the data varies the most.

5. RESULTS

In our study, we investigated the efficacy of eight distinct machine learning algorithms in predicting heart disease. These algorithms encompassed decision tree, random forest, logistic regression, K-nearest neighbor (KNN), support vector machine (SVM), AdaBoost, XGBoost, and LightGBM. To conduct our evaluation, we utilized a dataset containing information from 303 patients. Each patient was characterized by 14 attributes, including age, sex, blood pressure, cholesterol level, and the presence of various symptoms. We divided this dataset into two subsets: a training set comprising 70% of the data and a test set containing the remaining 30%. Our assessment of algorithm performance focused on three key metrics: accuracy, precision, and recall. Accuracy gauges the overall percentage of correctly classified instances, precision measures the proportion of true positive predictions among all positive predictions, and recall quantifies the percentage of true positive predictions among all actual positive instances. Upon analyzing the results (depicted in Fig 8), we observed promising outcomes overall, affirming the potential of machine learning algorithms in predicting heart disease. Although variations in performance were evident among the different algorithms, many demonstrated commendable performance levels. Notably, Decision Tree, LightGBM, and SVM emerged as particularly promising candidates for further exploration. However, it's crucial to note that logistic regression exhibited the lowest accuracy among the algorithms assessed, achieving only a 77% accuracy rate. This finding underscores the importance of considering algorithm selection carefully and highlights potential areas for improvement in predictive modeling endeavors related to heart disease.

	Accuracy	F1	Precision	Recall	Latency
LogisticRegression	0.777	0.785	0.750	0.824	6.8ms
DecisionTree	1.000	1.000	1.000	1.000	1.5ms
RandomForest	1.000	1.000	1.000	1.000	7.6ms
AdaBoost	0.990	0.990	1.000	0.980	95.6ms
XGB	0.990	0.990	1.000	0.980	2.3ms
LGBM	1.000	1.000	1.000	1.000	1.8ms
KNeighbors	1.000	1.000	1.000	1.000	5.7ms
SVC	1.000	1.000	1.000	1.000	2.7ms

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 593-598

e-ISSN : 2583-1062

Impact Factor: 5.725

	Accuracy	F1	Precision	Recall	Latency
LogisticRegression	0.814	0.816	0.764	0.875	5.3ms
DecisionTree	0.980	0.979	1.000	0.958	2.7ms
RandomForest	0.971	0.970	0.941	1.000	14.1ms
AdaBoost	0.961	0.959	0.940	0.979	119.7ms
XGB	0.941	0.938	0.938	0.938	25.1ms
LGBM	0.971	0.970	0.941	1.000	31.0ms
KNeighbors	0.971	0.970	0.941	1.000	11.7ms
SVC	0.971	0.970	0.941	1.000	3.7ms

Fig. 8. Evaluation of Models On The Test

6. CONCLUSION

In conclusion, predicting heart disease is indeed a multifaceted task that demands thorough consideration of various risk factors, such as age, gender, family history, lifestyle habits, and medicalhistory. Machine learning algorithms offer promise in this endeavors by demonstrating their ability to accurately predict heart disease and identify individuals at heightened risk.

Through the utilization of extensive datasets and sophisticated computational techniques, these models can uncover subtle patterns and relationships that might elude human perception. However, it's essential to recognize the inherent limitations ofmachine learning models.

These include the potential for biases within the training data and the necessity for human interpretation of the results. Despite their remarkable capabilities, these algorithms are not infallible and should be utilized in conjunction with clinical expertise and judgment Furthermore, understanding the various classifications of chest pain related to cardiac issues is crucial for accurate diagnosis and treatment.

These classifications include typical angina, atypical angina, non-anginal pain, and asymptomatic presentations. Each type of chest pain carries its own implications and necessitates careful evaluation by healthcare professionals to ensure appropriate management and intervention.

7. REFERENCE

[1] Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Deye, Y. (2020). A comparison of deep learning performance against health-care professionals in detecting diseases frommedical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 2(6), e271-e297.

[2] Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Asirvatham,S. J. (2019). Screening for cardiac contractile dysfunction using artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70-74

[3] Goldstein, B. A., Navar, A. M., Carter, R. E. (2017). Machine learning revolutionizes cardiovascular risk prediction by managing high- dimensional data. European Heart Journal, 38(23), 1805-1814.

[4] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., Kitai, T. (2020). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 75(23), 3003-3015

[5] Deo, R. C. (2015). Machine learning inmedicine. Circulation, 132(20), 1920-1930

[6] Dey, D., Wong, N. D., Tamarappoo, B. K. (2017).The article explores if machine learning is fullyutilized in cardiovascular medicine yet. Heart, 103(14), 1056-1063.

[7] Shah, S. J., Katz, D. H., Selvaraj, S., Burke, M. A., Yancy, C. W., Gheorghiade, M., Bonow, R. O. (2015). The study introduces "phenomapping" as a fresh approach to categorize heart failure with preserved ejection fraction (HFpEF). Circulation, 131(3), 269-279.

[8] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.

[9] Yan, L. L., Rosamond, W. D., Chambless, L. E.(2005). Prediction of cardiovascular disease incidence by measures of abdominal obesity: resultsfrom the Atherosclerosis Risk in Communities (ARIC) study. American Journal of Clinical Nutrition, 81(1), 7-13.

[10] Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P. (2018). Genetic risk, adherence to a healthy lifestyle, and coronary disease. New EnglandJournal of Medicine, 379(7), 633-645.