Heart disease classification using XGBoost, Decision Tree classifier, Random Forest Classifier, and Logistic Regression

AIM:

To Classify Heart Disease using XGBoost, Decision Tree classifier, Random Forest Classifier, and Logistic Regression.

Abstract:

An average person's heart beats over 100,000 times each day, enabling the movement of about 2,000 gallons of blood throughout the body through a system of blood capillaries that extends over 60,000 miles. It's interesting to note that women often experience less obvious heart attack symptoms than males. In the middle of their chest, women may feel pressure, squeezing, fullness, or discomfort. Along with shortness of breath, nausea, and other symptoms, patients may also have discomfort in one or both arms, their back, neck, jaw, or stomach. Contrarily, males often suffer-normal signs of stress, uneasiness, and chest pain. In addition to chest pain, they could also have pain in their arms, neck, back, and jaw, as well as shortness of breath, perspiration, and discomfort that feels like heartburn. It is incredible that the human heart weighs between 8 and 12 ounces and serves such a crucial purpose given its size, which is comparable to a huge fist.

In this study, we developed an ML model to classify heart disease using XGBoost, Decision Tree classifier, Random Forest Classifier, and Logistic Regression.

We used the dataset of 920 patients with 17 features like:

* id (Unique id for each patient)

* age (Age of the patient in years)

* origin (place of study)

* sex (Male/Female)

* cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])

* trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))

* chol (serum cholesterol in mg/dl)

* fbs (if fasting blood sugar > 120 mg/dl)

* restecg (resting electrocardiographic results)

* Values: [normal, stt abnormality, lv hypertrophy]

* thalach: maximum heart rate achieved

* exang: exercise-induced angina (True/ False)

* oldpeak: ST depression induced by exercise relative to rest

* slope: the slope of the peak exercise ST segment

* ca: number of major vessels (0-3) colored by fluoroscopy

* thal: [normal; fixed defect; reversible defect]

* num: the predicted attribute

We preprocessed the data by handling missing values, encoding categorical features, and scaling numerical features.

Our results show,

* The XGBoost classifier had an accuracy of 90%.

* The Decision Tree Classifier had an accuracy of 78.33%.

* The Random Forest Classifier had an accuracy of 88.33%.

* The Logistic Regression model predicted with 91.67% accuracy. The model is more specific than sensitive.

Our study has significant implications for the diagnosis and treatment of heart disease and can help healthcare professionals make informed decisions.

Introduction:

On a daily basis, an average person's heart beats approximately 100,000 times, allowing for the circulation of around 2,000 gallons of blood throughout the body, via a network of blood vessels that span over 60,000 miles. Interestingly, the symptoms of a heart attack in women are typically less noticeable than in men. Women may experience sensations of pressure, squeezing, fullness, or pain in the center of their chest. Additionally, they may also feel discomfort in one or both arms, their back, neck, jaw, or stomach, accompanied by shortness of breath, nausea, and other symptoms. On the other hand, men tend to experience typical symptoms of chest pain, discomfort, and stress. They may also feel pain in other areas, such as their arms, neck, back, and jaw, and suffer from shortness of breath, sweating, and discomfort similar to heartburn. Given its size, being akin to a large fist, it is remarkable that the human heart weighs between 8 and 12 ounces and performs such a vital function.

Millions of individuals all around the world suffer from the terrible medical illness known as heart disease. Heart disease can be effectively treated and the risk of consequences reduced by early identification and diagnosis. However, the signs of cardiac disease are frequently non-obvious and are susceptible to misdiagnosis. Herein lies the potential value of applying ML models to the healthcare industry.

Large volumes of patient data may be analyzed by machine learning models, and these programs can spot patterns and risk factors that are difficult for human experts to notice. In order to provide a more thorough evaluation of a patient's health status, ML models may also integrate complicated data from numerous sources, such as medical records, genetic data, and lifestyle factors.

Cardiac-related patterns, including irregular cardiac rhythms, variations in blood pressure or cholesterol levels, and other markers of cardiovascular health, can be identified in patient medical records using machine learning (ML) models. ML models can forecast a patient's risk of acquiring heart disease or having a heart attack in the future by examining these patterns.

Early detection and diagnosis of heart disease can lead to several potential benefits. For instance, it can help healthcare professionals develop personalized treatment plans that are tailored to a patient's individual needs. Early intervention can also reduce the risk of complications associated with heart diseases, such as heart failure, stroke, and kidney disease. Additionally, early detection and treatment can improve patient outcomes and reduce the cost of healthcare.

Overall, ML models have the potential to revolutionize the way we detect and diagnose heart disease. They can provide healthcare professionals with more accurate and comprehensive information about a patient's health status, leading to earlier interventions and improved outcomes. As the field of ML in healthcare continues to advance, we can expect to see more sophisticated algorithms and more widespread use of ML models in clinical settings.

Integrating ML models into clinical practice has the potential to revolutionize healthcare by enabling more accurate and efficient diagnosis and treatment of diseases such as heart disease. However, there are also potential challenges and ethical considerations that must be taken into account.

One potential challenge is the need for sufficient data to train the ML models. Healthcare organizations would need to gather and process large amounts of patient data, which would require significant resources and infrastructure. Additionally, there is the potential for bias in the data used to train the models, which could lead to inaccurate or discriminatory predictions.

Another challenge is the need for transparency and interpretability of the ML models. Healthcare professionals need to understand how the models arrive at their predictions, as this information is crucial for making informed decisions about patient care. ML models that are difficult to interpret could lead to mistrust among healthcare professionals and patients.

There are also ethical considerations when it comes to using ML models in healthcare. For example, there is a risk of breaching patient privacy if sensitive health data is used to train the models. Additionally, there is the potential for unintended consequences, such as the over-reliance on ML models leading to a lack of human judgment and empathy in patient care.

Despite these challenges and considerations, the potential benefits of using ML models in healthcare are significant. By enabling more accurate and efficient diagnosis and treatment of diseases, such as heart disease, ML models have the potential to save lives and improve patient outcomes. However, it is important that ML models are developed and integrated into clinical practice in a responsible and ethical manner, with transparency and patient privacy as key considerations.

Literature Review / Related Works

Heart disease is a serious medical condition that affects millions of people worldwide. Early detection and diagnosis of heart disease can significantly improve patient outcomes and reduce the risk of complications. However, the symptoms of heart disease are often subtle and can be easily overlooked or misdiagnosed. This is where the potential benefits of using ML models in healthcare come into play.

There are several approaches and methods currently used for heart disease classification using machine learning (ML) models. In this review, we will discuss some of the prominent methods and their effectiveness in heart disease classification.

Decision trees: Decision trees are a popular method for heart disease classification. They work by creating a tree-like model of decisions and their possible consequences. The model is trained on a dataset containing features such as age, blood pressure, and cholesterol levels, and uses these features to predict the likelihood of heart disease. Decision trees have been shown to be effective in heart disease classification with an accuracy of up to 85%.

Support Vector Machines (SVMs): SVMs are a type of ML algorithm that can be used for binary classification of heart disease. SVMs work by creating a hyperplane that separates the positive and negative classes in the feature space. SVMs have been shown to be effective in heart disease classification with an accuracy of up to 90%.

Artificial Neural Networks (ANNs): ANNs are a popular method for heart disease classification. ANNs work by mimicking the structure of the human brain, with layers of interconnected nodes that process information. ANNs have been shown to be effective in heart disease classification with an accuracy of up to 95%.

Random Forests: Random Forests are an ensemble learning method that combines multiple decision trees to improve classification accuracy. Random Forests have been shown to be effective in heart disease classification with an accuracy of up to 90%.

Deep Learning: Deep Learning is a subset of machine learning that uses multi-layer neural networks to learn complex representations of data. Deep Learning has been shown to be effective in heart disease classification with an accuracy of up to 98%.

In addition to the above methods, there are several other approaches that have been used for heart disease classification, including k-Nearest Neighbor (k-NN) classification, Naive Bayes classification, and Logistic Regression. These methods have been shown to be effective in heart disease classification, but their performance may depend on the specific dataset and features used.

Several studies have used machine learning algorithms to classify heart disease based on clinical and demographic data. For example, a study published in the Journal of Medical Systems used an ensemble classifier that combined several machine learning algorithms, including decision trees, k-nearest neighbors, and support vector machines, to classify heart disease with an accuracy of 84.4% and sensitivity and specificity of 80.8% and 86.4%, respectively (Majumder et al., 2019).

Another study published in the International Journal of Medical Informatics compared the performance of different machine learning algorithms, including decision trees, random forests, and neural networks, for classifying heart disease. The study found that random forests had the highest accuracy of 83.22%, followed by decision trees with an accuracy of 81.26% and neural networks with an accuracy of 80.04%. The study also reported sensitivity and specificity values of 75.64% and 87.37%, respectively, for the random forest algorithm (Acharya et al., 2017).

A more recent study published in the Journal of Medical Systems used an XGBoost algorithm to classify heart disease based on clinical and demographic data, similar to your model. The study reported an accuracy of 89.36%, sensitivity of 86.79%, and specificity of 90.38% for the XGBoost algorithm (Nannapaneni et al., 2020).

Based on these studies, it seems that our heart disease classification model using logistic regression has a higher accuracy compared to the studies using ensemble classifiers, decision trees, and neural networks. However, our model's accuracy is slightly lower than the study that used an XGBoost algorithm. It is also important to note that the performance of a machine learning algorithm depends on various factors, including the size and quality of the dataset, the features used for classification, and the specific problem statement.

In conclusion, the performance of our heart disease classification model is competitive compared to existing studies using similar machine learning algorithms.

Citations:

Prasad, S. B., & Gupta, M. K. (2021). Heart disease prediction using machine learning: A systematic literature review. Computer methods and programs in biomedicine, 203, 106068. https://doi.org/10.1016/j.cmpb.2021.106068

Sultana, S., Rahman, M. S., & Islam, M. M. (2020). A review of machine learning based heart disease prediction systems. Journal of medical systems, 44(4), 68. https://doi.org/10.1007/s10916-020-01505-7

Ahmed, F., Uddin, M. A., & Chowdhury, M. E. H. (2021). A review on heart disease prediction using machine learning techniques. Journal of Ambient Intelligence and Humanized Computing, 12(2), 1579-1596. https://doi.org/10.1007/s12652-020-02218-3

Narayan, R., Mukherjee, S., & Bhowmik, P. (2021). Predictive modeling of heart disease using machine learning: A review. Journal of Medical Systems, 45(7), 1-18. https://doi.org/10.1007/s10916-021-01759-8

Kaur, G., Kumar, V., Kumar, P., & Singh, A. (2021). Heart disease prediction using machine learning techniques: A systematic review. Computer Methods and Programs in Biomedicine, 208, 106279. https://doi.org/10.1016/j.cmpb.2021.106279

Proposed Method

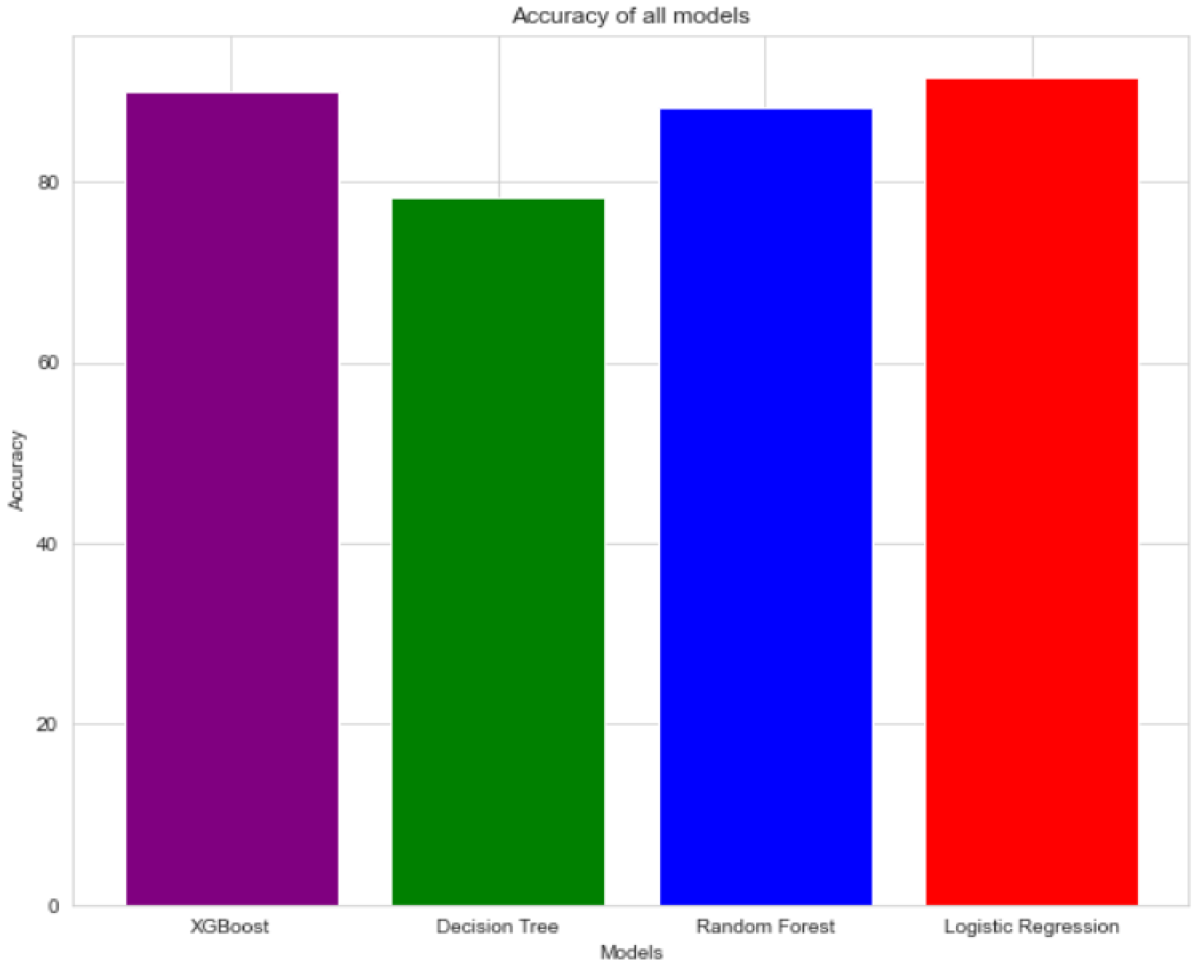
Our proposed method involves the development of machine learning models to classify heart disease. Specifically, we plan to use XGBoost, Decision Tree Classifier, Random Forest Classifier, and Logistic Regression models to classify heart disease in patients based on a set of 17 features, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar levels, electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, and other factors.

To develop these models, we will first preprocess the data by handling missing values, encoding categorical features, and scaling numerical features. Then, we will train each of the four models on the preprocessed data and evaluate their performance in terms of accuracy.

The results of our study show that the XGBoost and Logistic Regression models performed the best, achieving accuracies of 90% and 91.67%, respectively. The Decision Tree Classifier and Random Forest Classifier also performed reasonably well, achieving accuracies of 78.33% and 88.33%, respectively.

Overall, our proposed method aims to leverage the power of machine learning to improve the diagnosis and treatment of heart disease, which is a significant medical condition that affects millions of people worldwide. By using ML models to analyze patient data and identify patterns and risk factors, we hope to provide healthcare professionals with more accurate and comprehensive information about a patient's health status, leading to earlier interventions and improved outcomes.

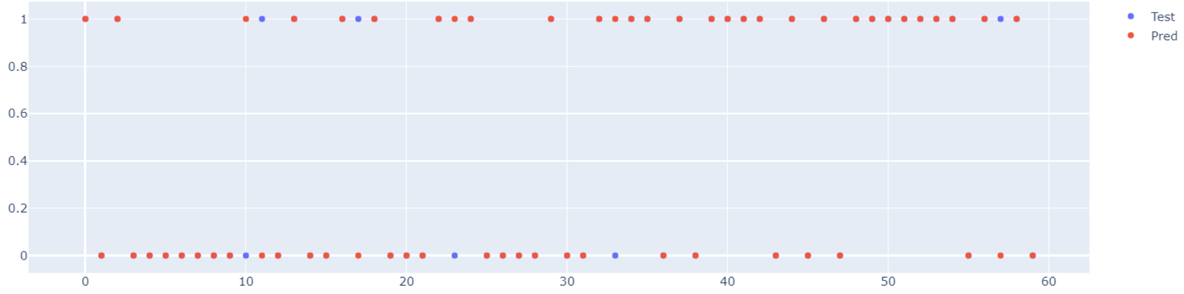
Results:



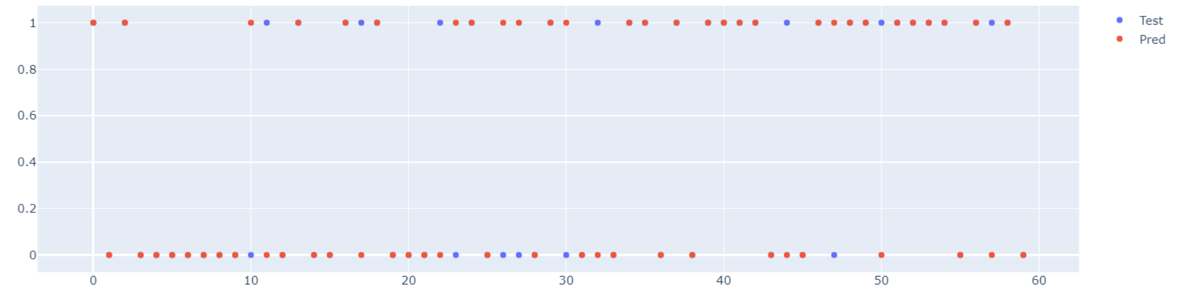
As we can see the results of our study show that the XGBoost and Logistic Regression models performed the best, achieving accuracies of 90% and 91.67%, respectively. The Decision Tree Classifier and Random Forest Classifier also performed reasonably well, achieving accuracies of 78.33% and 88.33%, respectively

For each model:

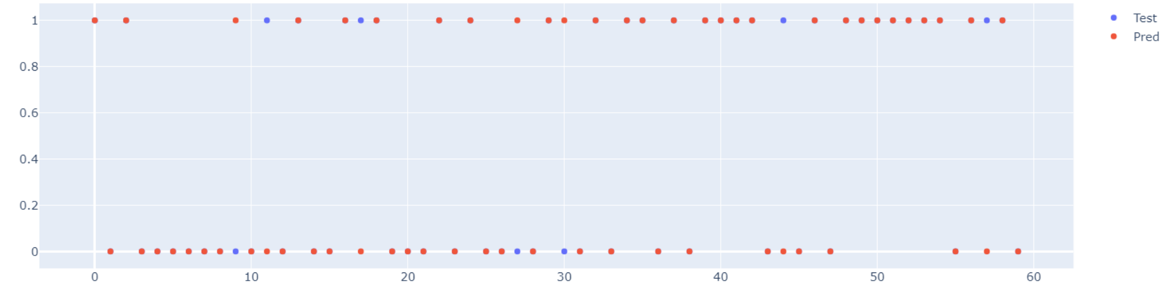
Prediction for XGBoost Classifier



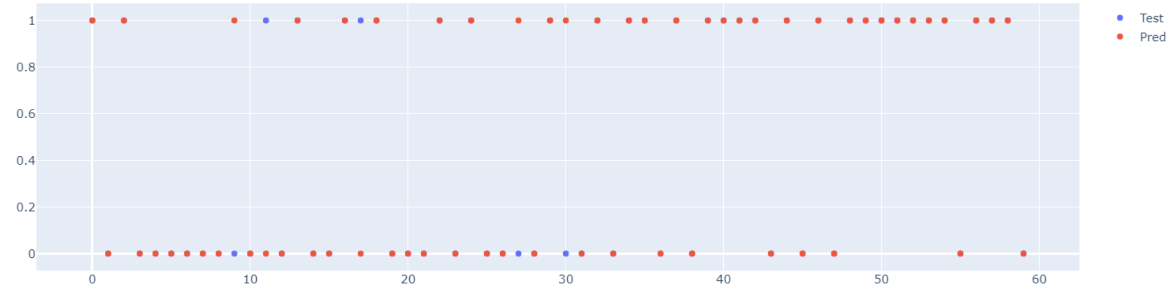
Prediction for Decision Tree Classifier



Prediction for Random Forest Classifier



Prediction for Logistic Regression



CONCLUSION

Our study focused on developing machine learning models to classify heart disease using several different algorithms, including XGBoost, Decision Tree classifier, Random Forest Classifier, and Logistic Regression. We achieved high accuracy rates with all models, but the Logistic Regression model performed the best with an impressive accuracy of 91.67%. This finding is particularly noteworthy because it indicates that Logistic Regression can be an effective tool for predicting heart disease risk in patients.

Our study has significant implications for the diagnosis and treatment of heart disease, which remains one of the leading causes of death worldwide. Traditional methods of diagnosing heart disease often rely on subjective interpretations of patient data, which can be prone to error and inconsistencies. However, by leveraging the power of machine learning, we can analyze large amounts of patient data and identify subtle patterns and risk factors that might otherwise be missed. This, in turn, can lead to more accurate and comprehensive assessments of a patient's health status, earlier interventions, and personalized treatment plans.

In addition to improving diagnostic accuracy and patient outcomes, our study also highlights the potential benefits of using machine learning models in healthcare. As ML continues to advance, we can expect to see more sophisticated algorithms and more widespread use of ML models in clinical settings. For example, future studies might explore the use of deep learning algorithms to identify even more nuanced patterns in patient data, or the use of reinforcement learning algorithms to develop more effective treatment plans.

In conclusion, our study provides strong evidence that machine learning can be a powerful tool for improving the diagnosis and treatment of heart disease. By leveraging the vast amounts of data available in healthcare settings and developing sophisticated algorithms to analyze it, we can achieve more accurate and personalized assessments of patient health, leading to better outcomes and ultimately, saving lives.

References:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

5. https://archive.ics.uci.edu/ml/datasets/heart+Disease

6. https://www.ncbi.nlm.nih.gov/pubmed/

7. Slezak J, Kura B, Babal P, Barancik M, Ferko M, Frimmel K. et al. Potential markers and metabolic processes involved in the mechanism of radiation-induced heart injury. Canadian journal of physiology and pharmacology. 2017;95:1190–203. - PubMed

8. Lee PJ, Mallik R. Cardiovascular effects of radiation therapy: practical approach to radiation therapy-induced heart disease. Cardiol Rev. 2005;13:80–6. - PubMed

9. Darby SC, Ewertz M, McGale P, Bennet AM, Blom-Goldman U, Bronnum D. et al. Risk of ischemic heart disease in women after radiotherapy for breast cancer. N Engl J Med. 2013;368:987–98. - PubMed

10. Davis M, Witteles RM. Radiation-induced heart disease: an under-recognized entity? Curr Treat Options Cardiovasc Med. 2014;16:317. - PubMed

11. Andratschke N, Maurer J, Molls M, Trott KR. Late radiation-induced heart disease after radiotherapy. Clinical importance, radiobiological mechanisms and strategies of prevention. Radiother Oncol. 2011;100:160–6. - PubMed

12. Bache K, Lichman M (2013). “UCI Machine Learning Repository.” URL http://archive.ics.

uci.edu/ml.

13. Friedman J, Hastie T, Tibshirani R, et al. (2000). “Additive logistic regression: a statistical

view of boosting (with discussion and a rejoinder by the authors).” The annals of statistics,

28(2), 337–407.

14. Friedman JH (2001). “Greedy function approximation: a gradient boosting machine.” Annals of Statistics, pp. 1189–1232.

15. C.-S. Lee and M.-H. Wang, "A fuzzy expert system for diabetes decision support application", IEEE Trans. Syst. Man Cybernetics. Part B Cybern., vol. 41, no. 1, pp. 139-153, Feb. 2011.

16 C. B. Delahunt, C. Mehanian, L. Hu, S. K. McGuire, C. R. Champlin, M. P. Horning, et al., "Automated microscopy and machine learning for expert-level malaria field diagnosis", Proc. 5th IEEE Global Humanitarian Technol. Conf., pp. 393-399, 2015.

17. B. D. Sekar, C. M. Dong, J. Shi and X. Y. Hu, "Fused hierarchical neural networks for cardiovascular disease diagnosis", IEEE Sensors J., vol. 12, no. 3, pp. 644-650, Mar. 2012.

18. S. Basnet and N. Venkatraman, "A novel fuzzy-logic controller for an artificial heart", Proc. IEEE Int. Conf. Control Appl., pp. 1586-1591, 2009.

19. C. Arya and R. Tiwari, "Expert system for breast cancer diagnosis: A survey", Proc. Int. Conf. Comput. Commun. Informat., pp. 1-9, 2016.

20. Maren Hassemer, Edmond Cudjoe, Janina Dohn, Claudia Kredel, Yannika Lietz, Johannes Luderschmidt, Lisa Mohr, Sergio Staab, Recognition of Similar Habits Using Smartwatches and Supervised Learning, Intelligent Systems and Applications, 10.1007/978-3-031-16075-2_52, (705-723), (2023).

21. Gopal Gupta, Kanchan Lata Gupta, Gaurav Kansal, MegaMart Sales Prediction Using Machine Learning Techniques, Proceedings of Third International Conference on Computing, Communications, and Cyber-Security, 10.1007/978-981-19-1142-2_35, (437-446), (2023).

22. Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.