IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Aws Sagemaker

[1]Miss. Sanika Sanjay Bhosale, [2]Miss.Madhura Vilas Patil, [3]Miss.Manasi Prashant Jadhav,

[4]Miss. Samiksha Sunil Abdar, [5]Prof. Suraj K. Patil

[1, 2, 3, 4] Student, [5] Assistant Professor

Department of CSE (Data Science),

D. Y. Patil College of Engineering & Technology, Kolhapur, India

Abstract: Heart disease is a leading global cause of death, making early detection essential. This project uses AWS SageMaker to build a predictive model based on key health indicators like age, cholesterol, and blood pressure. A robust pipeline handles preprocessing, feature selection with Random Forest, and model training using XGBoost. Hyperparameter tuning improves performance across accuracy, precision, and AUC. The model is deployed to a real-time endpoint for fast predictions, and a Streamlit app offers a simple interface for healthcare professionals and patients to assess heart disease risk.

***Keywords*- AWS SageMaker, XGBoost, Random Forest, Feature Selection, Cardiovascular Risk Assessment**

I. INTRODUCTION

Heart disease remains one of the most significant global health challenges, causing millions of deaths each year. The ability to predict heart disease at an early stage is crucial because it enables timely medical intervention, reducing the risk of severe complications such as heart attacks or strokes. Traditional diagnostic methods often rely on physical examinations, medical history, and basic tests, but these approaches may not always detect risks in their early stages. Machine learning and artificial intelligence (AI) have revolutionized medical research by offering more precise and data-driven predictions, helping doctors identify at-risk patients before symptoms become critical. Early detection allows for preventive measures such as lifestyle changes, medication, and continuous monitoring, improving patient survival rates and quality of life. Additionally, it helps reduce the burden on healthcare systems by preventing severe cases that require extensive medical treatment.

With advancements in data science, predictive models can analyze vast amounts of medical data, including patient history, cholesterol levels, blood pressure, and lifestyle habits, to identify potential risks. Machine learning algorithms, particularly Random Forest and deep learning models, can recognize complex patterns in patient data that might be overlooked in traditional medical assessments. By integrating predictive analytics into healthcare systems, hospitals and clinics can enhance early diagnosis, streamline treatment plans, and improve patient management. Furthermore, predictive models contribute to medical research by identifying new risk factors and improving personalized treatment strategies. In the long run, AI-powered heart disease prediction can lead to better healthcare outcomes, cost savings, and advancements in cardiovascular medicine, making it an essential area of research and implementation.

II. LITERACTURE SURVEY

Kulkarni, Thakur, Nalbalwar, Shah, and Chordia [1] proposed a model which conducts a detailed comparative analysis of two prominent cloud platforms, Amazon SageMaker and Heroku, focusing on their efficiency and scalability in deploying machine learning models. The study evaluates the platforms based on their features, performance, community support, and extensibility, emphasizing the importance of developer ecosystems in fostering innovation. Special attention is given to healthcare, particularly heart disease prediction, where globally accessible tools like these platforms can significantly impact health outcomes. The paper serves as a practical guide for researchers and practitioners, providing insights into the deployment of machine learning models, particularly for heart disease prediction, using Amazon SageMaker and Heroku. This comprehensive analysis helps inform decisions regarding platform selection for scalable and efficient model deployment in real-world applications.

Nayab Akhtar [2] created a model to predict heart disease, which is a major cause of death worldwide. This model uses machine learning and data analysis techniques to help doctors make better diagnoses. The research looks at five different algorithms: Naive Bayes, k-Nearest Neighbour (KNN), Decision Tree, Artificial Neural Network (ANN), and Random Forest. These algorithms use health data like age, gender, cerebral palsy (CP), blood pressure, and blood sugar levels to predict heart disease. The study tests and compares these methods to see which one gives the most accurate results.

Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel [3] proposed a model that diagnosis a heart disease, numerous studies have employed data mining techniques to enhance prediction accuracy and reduce the necessity of extensive medical testing. Decision tree algorithms, particularly J48, Logistic Model Tree, and Random Forest, have gained attention for their effectiveness in classifying and predicting heart disease. Utilizing datasets such as the Cleveland database from the UCI repository, researchers have explored the performance of these algorithms on a sample of 303 instances with 76 attributes. The goal is to uncover hidden patterns and trends within the data to help healthcare professionals make faster and more efficient diagnoses. The comparative analysis of different decision tree algorithms aims to determine the most effective method for heart disease prediction, particularly in larger datasets, thereby potentially reducing heart disease-related fatalities.

Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, and Harshali Rambade [4] developed a model to predict heart disease using data mining techniques. Their aim is to improve healthcare efficiency and cut costs by creating an automated system for diagnosing heart disease. They used data from sources like Kaggle and the Cleveland Foundation to find hidden patterns in large datasets, which are key for accurate predictions. The researchers point out that handling the huge and complex data for heart disease prediction is difficult, and traditional methods aren't enough. Their goal is to find a machine learning method that is both fast and accurate, using statistical analysis, machine learning, and database tools to find important patterns in the data.

Chintan M. Bhatt, Parth Patel, Tarang Ghetia and Pier Luigi Mazzeo [5] proposed a machine learning-based framework for the diagnosis and prognosis of cardiovascular disease, aiming to enhance the accuracy of classification models used in clinical decision-making. Their study emphasizes the potential of machine learning in reducing misdiagnosis by recognizing patterns in medical data. The authors introduce a methodology incorporating k-modes clustering with Huang initialization to improve classification accuracy, followed by the application of models including decision tree (DT), random forest (RF), multilayer perceptron (MP), and XGBoost (XGB). Utilizing a real-world dataset of 70,000 instances from Kaggle, the models were trained with an 80:20 train-test split and optimized using GridSearchCV for hyperparameter tuning. The study reports that the multilayer perceptron with cross-validation achieved the highest accuracy of 87.28%, surpassing the other models. Additionally, all models demonstrated strong performance with AUC values of 0.94 to 0.95. This research highlights the effectiveness of neural networks in cardiovascular disease prediction and provides practical insights for implementing optimized, data-driven diagnostic tools in healthcare.

Ahmad, A.A.; Polat, H [6] presented a heart disease prediction framework leveraging machine learning to enable early detection and improve patient outcomes. Recognizing the critical need for timely diagnosis, the study focuses on the Cleveland heart disease dataset and explores the impact of feature selection and model choice on predictive performance. To address the challenge of high dimensionality and prevent overfitting, the authors employed the Jellyfish optimization algorithm, known for its fast convergence and adaptability in selecting the most relevant features. Various machine learning models were trained on the optimized dataset, and their performance metrics were compared. Among the models tested, the Support Vector Machine (SVM) classifier demonstrated superior results, achieving a sensitivity of 98.56%, specificity of 98.37%, accuracy of 98.47%, and AUC of 94.48%. The research highlights the

effectiveness of combining advanced feature selection with robust classification techniques and underscores the promise of the Jellyfish-SVM approach in building accurate, data-driven diagnostic systems for heart disease.

Tianqi Chen, Carlos Guestrin [7] developed XGBoost, a scalable and efficient tree boosting system that delivers state-of-the-art performance on machine learning tasks. The system introduces sparsity-aware learning, a weighted quantile sketch for tree construction, and optimizations in memory access and data sharding, allowing it to handle billions of data points with high speed and low resource use.

III. SYSTEM DESIGN

There are five key stages in the system architecture: data input, preprocessing, feature selection, risk classification, and result generation. As soon as a user inputs patient data through the Streamlit interface, the system preprocesses it by cleaning, encoding, and normalizing features. Next, feature selection is performed using a Random Forest model to retain only the most relevant health indicators. The refined data is then passed to a trained XGBoost classifier hosted on AWS SageMaker. The model predicts whether the individual is at risk of heart disease. Finally, the system provides a prediction.

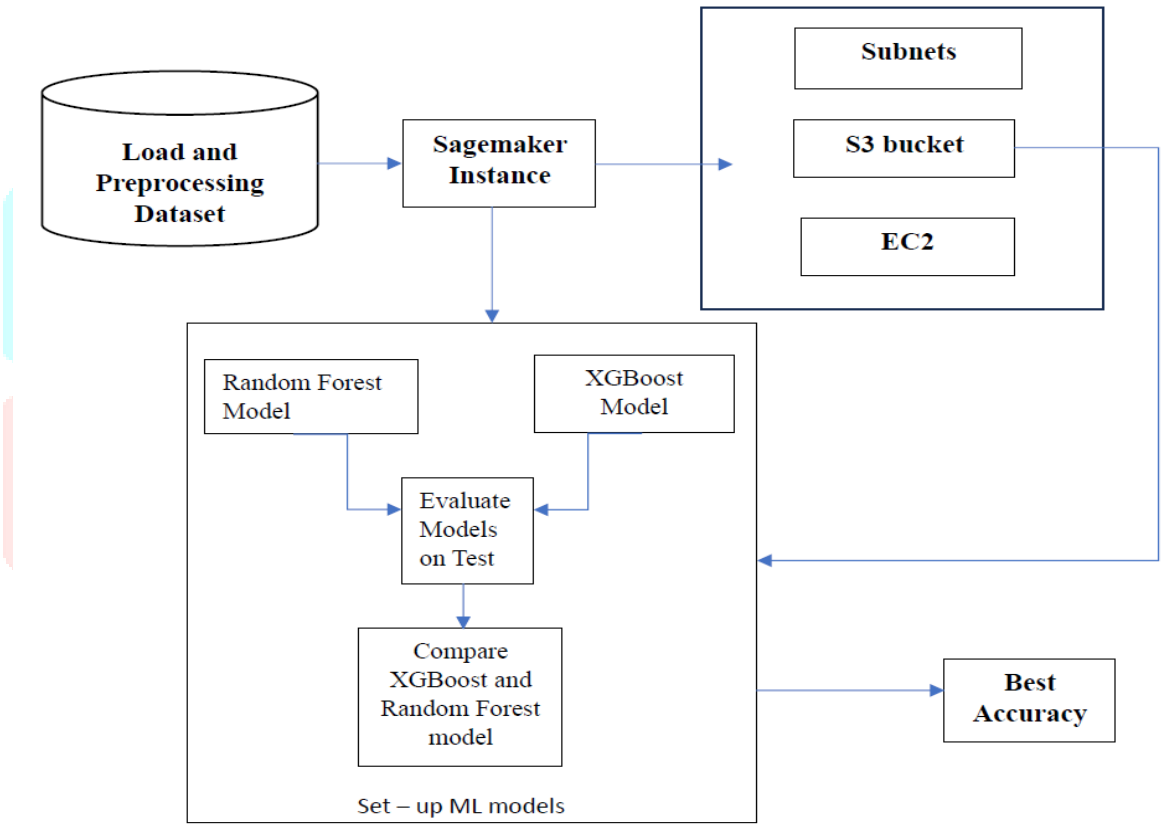

Fig.1. system architecture of heart disease prediction

- **Data Collection:** The dataset for this heart disease prediction project was collected from electronic health records (EHRs) of multiple hospitals between 2010 and 2015. It includes 70,000 patient records with 13 attributes, covering demographics, lifestyle factors, and medical test results. Data was recorded manually by healthcare professionals and through automated hospital databases, ensuring accuracy. Key features such as age, height, weight, blood pressure, cholesterol, and glucose levels were standardized for analysis. Missing values and inconsistencies were handled to improve data quality. The dataset was stored in CSV format for easy processing and used for predictive modeling to assess heart disease risk.

- **Load and Preprocess Dataset:** The first step involves loading the heart disease dataset and preparing it for model training. The dataset, stored in an S3 bucket or a local storage system, is retrieved for preprocessing. This process includes handling missing values (e.g., filling with mean/median or removing rows), standardizing numerical features, and encoding categorical variables if necessary. Additionally, the dataset is split into training and testing sets to evaluate model performance later. A data preprocessing pipeline is implemented using sklearn's Pipeline or AWS SageMaker Pipelines, incorporating feature

selection methods such as RandomForestRegressor to improve model efficiency. After transformation, the processed data is saved back into S3 for further use.

- **Launch AWS SageMaker Instance:** To train the models, an AWS SageMaker instance is launched. SageMaker is a fully managed machine learning service that allows scalable model training and deployment. The instance connects to S3 to fetch the preprocessed dataset and utilizes EC2 compute resources for model training. This setup ensures efficient training, leveraging AWS cloud infrastructure for optimized performance. The SageMaker instance also interacts with other AWS services, such as subnets for secure networking and IAM roles for managing access permissions.

- **Train Machine Learning Models (Random Forest & XGBoost):** The preprocessed data is then used to train two different machine learning models: Random Forest and XGBoost. Random Forest is an ensemble learning technique that builds multiple decision trees to improve accuracy and reduce overfitting, whereas XGBoost (Extreme Gradient Boosting) is a powerful boosting algorithm that optimizes performance using gradient descent techniques. Both models are trained on the training dataset within the SageMaker instance, leveraging EC2 instances for computation. The trained models are stored in S3 for further evaluation.

- **Evaluate Models on Test Dataset:** Once training is complete, both models are evaluated on the test dataset to measure their performance. The evaluation metrics, such as accuracy, precision, and recall are computed to determine how well each model generalizes to unseen data. These evaluations help in understanding which model performs better for heart disease prediction**.**

- **Compare Model Performance and Select the Best Model:** After evaluation, the Random Forest and XGBoost models are compared based on their performance metrics. The model that achieves the highest accuracy and better generalization is selected for deployment. If both models have similar performance, hyperparameter tuning techniques may be applied to further optimize them.

- **Store The Best Model:** Once the best-performing model is, it needs to be stored and deployed for real-world usage. This step ensures that the trained model is accessible for future predictions without the need for retraining. After selecting the best-performing model (Random Forest or XGBoost), it is saved using Joblib as a pkl file for future use. This allows efficient model storage and retrieval without retraining. The model can be reloaded later using Joblib's load function and used for predictions.

- **Interface:** After selecting the best-performing model (Random Forest or XGBoost), it is saved using Joblib as a pkl file. The model is then integrated into a Streamlit application to create an interactive web interface for real-time predictions. Users can input relevant features through the Streamlit UI, and the app will load the saved model to generate predictions. This ensures a user-friendly and efficient deployment without needing complex backend infrastructure**.**

IV. METHODOLOGY

The system proposed in this project utilizes a machine learning-based approach for heart disease risk prediction. A structured tabular dataset containing 70,000 records with 13 key health features—such as age, cholesterol level, resting blood pressure, and maximum heart rate—was used for training and testing. Data preprocessing involves handling missing values, encoding categorical variables, and normalizing numerical features using a unified preprocessing pipeline. Feature selection is performed using a Random Forest model to identify the most relevant predictors of heart disease. The final predictive model is built using the XGBoost algorithm, selected for its high accuracy and efficiency on structured medical data. Hyperparameter tuning is conducted using SageMaker's built-in optimization tools to improve model performance across accuracy, precision, and AUC. The trained model is then deployed on an AWS SageMaker real-time endpoint for scalable inference. A user-friendly web interface is developed using Streamlit, enabling users to input patient data and receive instant predictions along with medical recommendations.

Implementation Details:

- **Input:** Feature values via a web interface
- **Output:** Binary Classification Result: Prediction: "High Risk" or "Low Risk" for heart disease
- **Algorithm:**

Step 1: Start

Step 2: Data Acquisition (User Input)

- Accept structured user input through a Streamlit web interface.
- Health metrics to be collected:
  - Age, Height, Weight, Systolic BP (ap_hi), Diastolic BP (ap_lo)
  - Cholesterol, Glucose, Gender
- Derive additional features:
  - BMI, Pulse Pressure, Mean Arterial Pressure (MAP), Sys/Diast Ratio, Age Box-Cox transform

Step 3: Data Preprocessing

- Send input data to AWS SageMaker endpoint where preprocessing is included in the model pipeline:
  - Replace infinity values
  - Impute missing values
  - Scale numerical features
  - One-hot encode categorical features
  - Apply feature selection

Step 4: Model Inference (Risk Classification)

- Load the XGBoost model from SageMaker.
- Predict the probability of heart disease.
- If prediction probability > 0.5 → Label as Risk, else No Risk.

Step 5: Result Display

- If No Risk: “Your current metrics indicate a low risk of heart disease.”
- If Risk: “You are predicted to be at risk of heart disease.”

V. RESULT ANALYSIS

Logistic Regression has a train accuracy score of 0.7447, meaning it correctly classifies approximately 74.5% of the instances in the training dataset. In contrast, both Random Forest and XGBoost models achieve a significantly higher train accuracy of 0.9988, correctly classifying over 99.8% of the training instances. Based on accuracy alone, Random Forest and XGBoost clearly outperform Logistic Regression, with nearly perfect classification performance on the training set. This suggests that they are the most effective models in terms of overall classification accuracy in this comparison. XGBoost stood out for heart disease prediction due to its strong performance, efficient handling of missing data, and advanced control features like regularization and early stopping. Though it matched Random Forest in accuracy, its sequential learning and scalability made it better suited for medical data with high predictive demands.

Table.1. model comparison

Model	Train Accuracy	Train Precision	Train Recall
Logistic Regression	0.744682	0.644275	0.392732
Random Forest	0.998844	0.998900	0.997366
XGBoost	0.998844	0.998900	0.997366

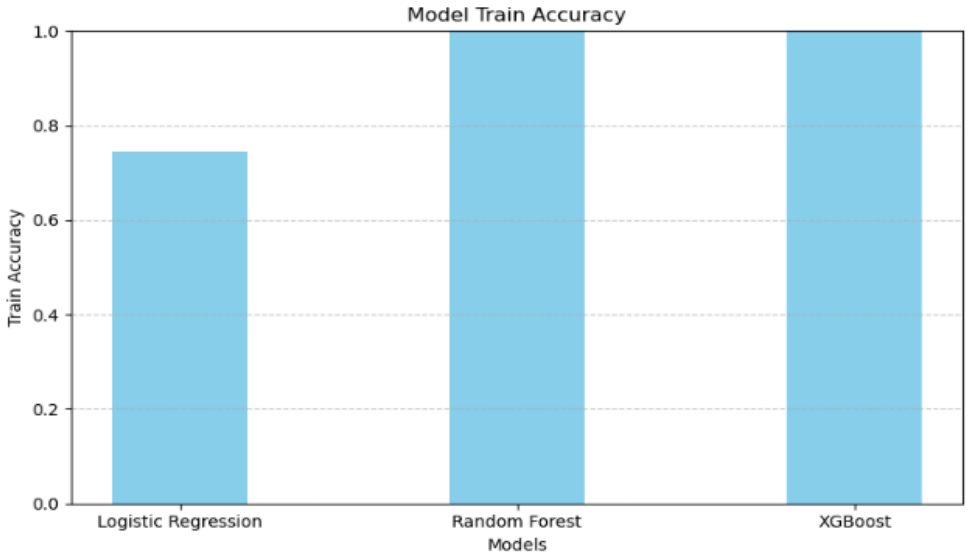

Fig.2. model train accuracy graph

- **Existing system vs Proposed system:**

The existing system for heart disease prediction relies on basic machine learning models such as Logistic Regression and Decision Trees, which offer limited accuracy (78.5%) due to minimal preprocessing, lack of feature selection, and absence of model tuning or deployment capabilities. In contrast, the proposed system employs an advanced XGBoost classifier, achieving a significantly higher accuracy of 99.8%. It integrates automated preprocessing pipelines using ColumnTransformer and Feature Union, applies feature selection through RandomForestRegressor, incorporates interaction features for improved learning, and uses hyperparameter tuning for performance optimization. Additionally, the system is deployed via a user-friendly Streamlit web application for real-time prediction, making it far more robust and practical for real-world healthcare use.

Table.2. Existing system vs Proposed system

Feature	Existing System	Proposed System
Model Accuracy	78.5%	99.8%
Preprocessing	Minimal or Manual	Pipeline
Feature Selection	None	RF – based Selection
Hyperparameter Tuning	Not Applied	Performed

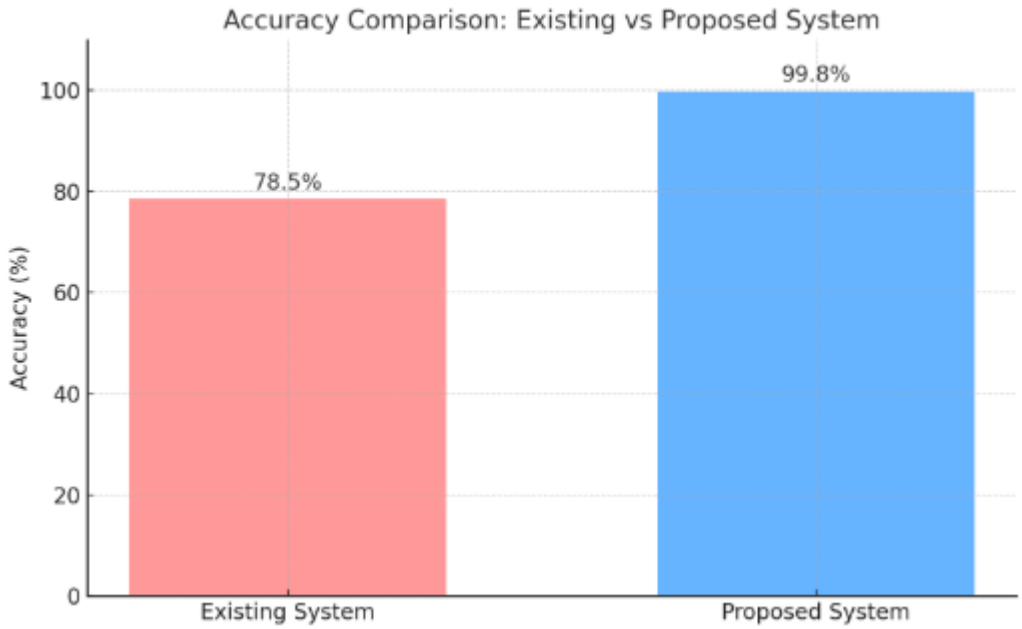

Fig.3. accuracy comparison: existing vs proposed system

VI. CONCLUSION

In conclusion, this project aimed to develop an efficient heart disease prediction system using machine learning. The dataset was first preprocessed through cleaning, standardization, and feature selection techniques to ensure quality input for model training. Two powerful models, Random Forest and XGBoost, were trained and evaluated based on their accuracy and performance metrics. The best-performing model was selected and saved using Joblib for future use. To make the model accessible and user-friendly, a Streamlit web application was developed, allowing users to input health data and receive real-time predictions. This deployment approach eliminates the need for complex backend systems and makes the model usable through a simple web interface. The entire workflow, from data preparation to interactive deployment, showcases how machine learning can be practically applied in the healthcare domain. The project not only demonstrates technical skills in model development but also highlights the importance of accessibility and usability in real-world applications. Ultimately, it provides a valuable tool that can assist in early diagnosis and support medical professionals in decision-making.

VII. FUTURE SCOPE

To enhance the current heart disease prediction model, future work could involve adding features like ECG readings, genetic data, and family history to improve accuracy. Integration with Electronic Health Record (EHR) systems would enable real-time clinical use. Deep learning models, such as CNNs or RNNs, may uncover more complex patterns within larger datasets. A mobile app version can increase accessibility and support proactive monitoring. Continuous retraining with new data will help maintain model relevance. Lastly, using explainable AI tools like SHAP or LIME can improve transparency, build trust, and aid medical professionals in understanding the model's decisions.

REFERENCES

[1] R. V. Kulkarni, A. Thakur, S. Nalbalwar, S. Shah and S. Chordia, "Exploring Scalable and Efficient Deployment of Machine Learning Models: A Comparative Analysis of Amazon SageMaker and Heroku," 2023 International Conference on Information Technology (ICIT), Amman, Jordan, 2023, pp. 746-751, doi: 10.1109/ICIT58056.2023.10225793

[2] Nayab Akhtar, "Heart Disease Prediction", Anatomy, February 2021.

[3] Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique", IJCSC, volume 7, number 1, Sept 2015 – March 2016.

[4] Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, Harshali Rambade, "Heart Disease Prediction using Machine Learning", International Journal of Research in Engineering, Science and Management, Volume-2, Issue-2, February-2019, www.ijresm.com | ISSN (Online): 2581-5792

[5] Chintan M. Bhatt, Parth Patel, Tarang Ghetia and Pier Luigi Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques", Algorithms 2023, 16, 88, 6 February 2023

[6] Ahmad, A.A.; Polat, H., "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm." Diagnostics 2023, 13, 2392. https://doi.org/10.3390/diagnostics13142392

[7] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", University of Washington.