International Research Journal of Education and Technology

Peer Reviewed Journal
ISSN 2581-7795

HEART DISEASE PREDICTION USING MACHINE LEARNING

Harini.M, Ishwarya.A, Kaviya.P, Dr.A.Baskar M.E,Ph.D

Department of Computer Science and Engineering

E.G.S.Pillay Engineering College, Nagapattinam, Tamil Nadu,India.

Abstract: Cardiovascular disease is still a leading cause of death throughout the world with the imperative need for effective early diagnosis in mitigating significant health risks. It is argued herein that an apparatus of a machine learning framework utilising the Gradient Boosting algorithm would make efficient heart disease predictions.. The system overcomes noisy data, excessive false positives, and computational intensity by utilizing strong pre processing of data, feature extraction, and iterative model fitting. The outcomes prove enhanced diagnostic precision, presenting a valid decision-support aid for clinicians. The project shows the potential of machine learning to revolutionize the diagnosis and treatment planning of heart disease.

Keywords: Age, cholesterol, and blood pressure. It also have algorithms such as SVM, Decision Trees, and Random Forests are commonly used. Gender, heart rate, logistic regression, random forest, precision, recall, and F1 score measure model performance. Assist doctors in diagnosis and better decision-making.

INTRODUCTION:

Heart disease is a severe global health issue, responsible for a significant percentage of deaths and chronic disability. Early prediction is challenging due to the complexity of medical data and shortcomings in existing diagnostic systems, such as high false positives and computational inefficiencies. This project aims to develop an intelligent, automated system using the Gradient Boosting algorithm to enhance prediction accuracy and enable timely medical decisions. Develop an accurate heart disease predictive system using Gradient Boosting. Make data pre processing and classification better to reduce false positives. Assist medical professionals in diagnosis and treatment planning. Develop a scalable machine learning system for heart disease prediction employ data mining techniques for effective classification and analysis. Provide an automated early diagnosis system with customized treatment.

II. Literature review:

1. A systematic literature review on heart disease predicition using block chain and machine learning techniques(2002):

The 2022 systematic literature review by Nouman and Aleeza speaks of the implementation of blockchain technology within heart disease prediction systems. The study identifies the importance of secure, decentralized management of data in healthcare, especially when dealing with sensitive patient data used in AI-based diagnostic models.In a review of broad scholarship literature available from 2015 to 2022, the authors discuss how blockchain increases data privacy, security, and interoperability. The review explains how decentralization in blockchain is preventing fraud and tampering with data and how valuable medical records can be accessed only by authorized users.The review also points toward the smart contract future to facilitate automating healthcare procedures, with real-time information sharing and faster decision-making for diagnostics. The union of blockchain and artificial intelligence in the guise of machine learning algorithms is shown to improve the accuracy and reliability of heart disease prediction by leaps and bounds. The authors also refer to current challenges like integration with existing healthcare systems, scalability, and implementation costs. Regardless of these challenges, the review identifies the potential of blockchain technology to facilitate patient-centered care where the patient has control over who accesses their information.

2.Heart disease prediction using machine learning algorithms(2020):

In 2020, Thilagavathi authored a paper on heart disease forecasting based on machine learning algorithms. The study was conducted in order to enable early diagnosis through the analysis of patient information using sophisticated computational methods. Some of the algorithms like Decision Trees, Support Vector Machines (SVM), and Random Forests were analyzed with respect to accuracy and efficiency.
The study concluded that machine learning algorithms were able to discern such crucial indicators of risk as cholesterol level, blood pressure, and age. Among all the algorithms that were tested, Random Forests were found to be the best for predicting heart disease. Data preprocessing, especially normalization and feature selection, was also found to be important in enhancing the efficiency of the models. Thilagavathi emphasized the opportunity for such models to inform health professionals in the making of evidence-based decisions. The study leveraged publicly available datasets, for example, the Cleveland Heart Disease dataset, in validating and training the models. The findings were that integrating machine learning with typical health screening might make diagnosis quicker and more reliable. The effort, as a whole, expands on the rising role of AI in preventive services and clinical decision support.

3.Multi label active learning based machine learnng model for heart disease prediction(2022):

In 2022, El Hasnony suggested a new machine learning approach for heart disease prediction using a multi-label active learning model. The study addresses the problem of the complexity of heart disease diagnosis in that patients usually have a combination of overlapped symptoms and risk factors. These instances require more than one-label models, and therefore the model attempts to predict multiple conditions or outcomes simultaneously. The active learning component allows the model to learn sufficiently from fewer labeled samples, and hence it works well on medical datasets where labeling is expensive or in short supply. El Hasnony validated the model with benchmark datasets and proved better predictive accuracy and robustness over traditional methods. The model also cut down training time by actively choosing the most informative instances to label. Feature selection methods were employed to determine the most important clinical parameters that are responsible for heart disease.

4. A web based heart disease prediction system using machine learning algorithms(2022):

Rahman and Mahbubur developed a web-based heart disease prediction system in 2022 that employs machine learning techniques to aid the detection of diseases at an early stage. The system is easy to operate and can be accessed online, where patients or physicians provide health data and receive live predictions.It combines different algorithms like Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) to examine patient data. The authors trained the system on clinical data, including blood pressure, cholesterol, and heart rate. Among the models tried, SVM made the most accurate prediction of heart disease probability. The web interface was created to offer ease of use for technical and non-technical users. Rahman and Mahbubur emphasized the importance of the tool in facilitating early diagnosis and relieving the burden on healthcare systems. The model also has data preprocessing operations to ensure maximum accuracy in prediction. Overall, the study shows how web-based AI technology can provide scalable and cost-effective solutions for heart disease risk prediction.

5. Enhanced heart disease prediction based on machine learning and x 2 statistical optimal feature selection:

Sarra and Raniya R introduced an enhanced heart disease prediction model in 2022 that combines machine learning with Chi-square (X^2) statistical feature selection to maximize accuracy. The study addresses the most impactful medical characteristics that affect heart disease, utilizing X^2 to reduce unnecessary or redundant features. Best feature selection eliminates clutter from data, enhancing the performance of some machine learning models. The authors used models such as Naive Bayes, Random Forest, and SVM with an optimized dataset and witnessed improved efficiency and prediction accuracy. They showed that their models, when trained on selected features, performed better than the whole dataset.The process also shortened training time and the cost of computations without a decline in precision. The most statistically relevant clinical features included were blood pressure, the nature of the chest pain, and cholesterol. The analysis stresses the value of feature extraction and data cleaning in building competent prediction systems. The combination of statistical techniques and ML is viewed by Sarra and Raniya as yielding predictions that are more interpretable and accurate. Their method presents a better mechanism in building clever, resource-constrained yet clinically valid predictors for heart diseases.

III.PROPOSED DESIGN

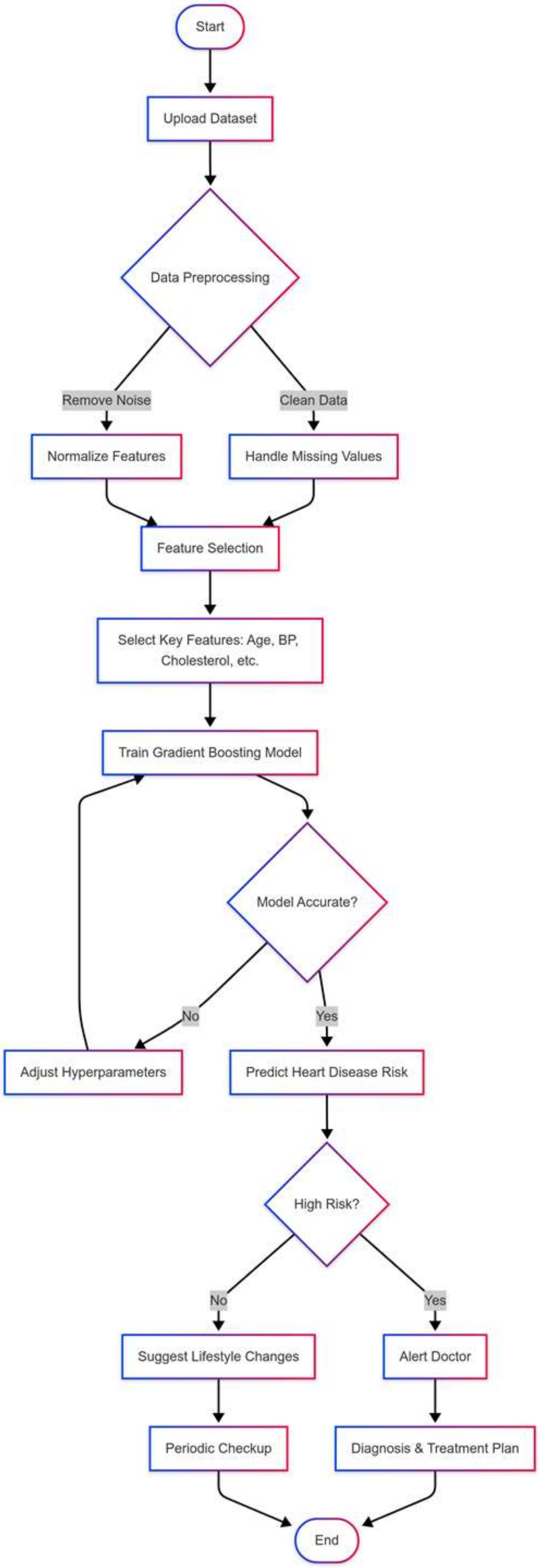
Dataset Acquisition: Retrieve datasets with characteristics like age, blood pressure, cholesterol level, and lifestyle habits (smoking, exercise).

Data pre processing: Replace missing values with actual values using imputation methods. Remove outliers and normalize data to make it uniform.

Feature Selection:

Choose significant predictors (e.g., cholesterol, blood pressure) through statistical analysis.

Eliminate duplicate features to prevent model complexity.
Model Training (Gradient Boosting):
Train the model step by step, reducing errors in prediction at each step.
Optimize hyper parameters for better accuracy. Diagnosis and Prediction Use the trained model to predict heart disease risks.
Provide actionable insights for healthcare professionals.


Start
Upload Dataset
Data Preprocessing
Remove Noise
Clean Data
Normalize Features
Handle Missing Values
Feature Selection
Select Key Features: Age, BP, Cholesterol, etc.
Train Gradient Boosting Model
Model Accurate?
No
Yes
Adjust Hyperparameters
Predict Heart Disease Risk
High Risk?
No
Yes
Suggest Lifestyle Changes
Alert Doctor
Periodic Checkup
Diagnosis & Treatment Plan
End

IV System Requirements

HARDWARE REQUIREMENTS

Processor – Dual core processor 2.6.0 GHZ

Ram - 4 GB

Hard Disk - 320 GB

Compact Disk - 650 Mb

Keyboard - Standard keyboard

SOFTWARE REQUIREMENT S

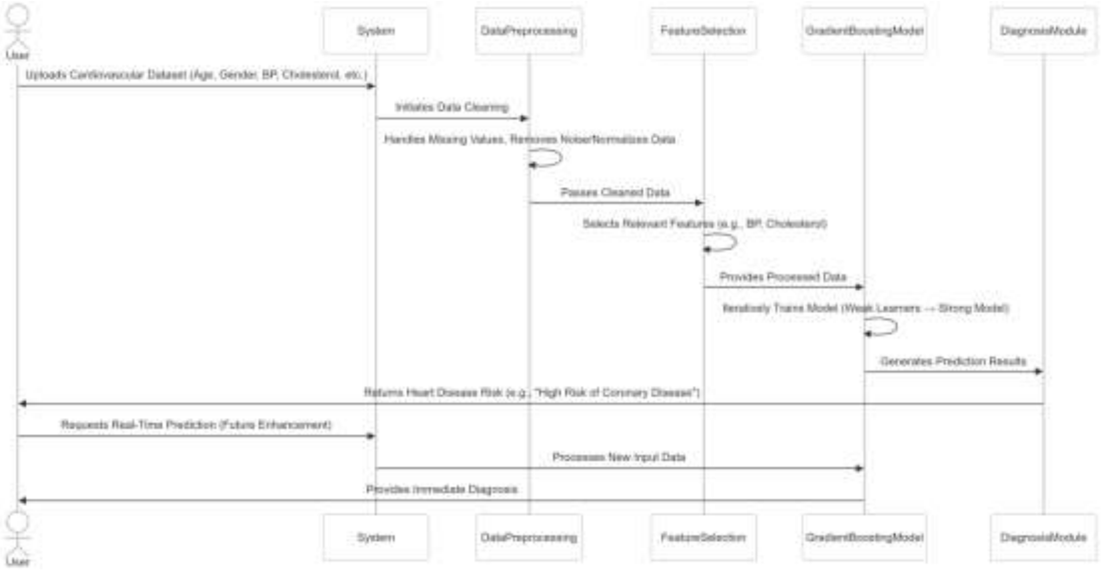
Operating System – Windows OS

Frontend: Python

Backend: MYSQL

IDE - PYCHARM

ACTIVITY DIAGRAM:


User
System
DataPreprocessing
FeatureSelection
GradientBoostingModel
DiagnosisModule
Uploads Cardiovascular Dataset (Age, Gender, BP, Cholesterol, etc.)
Initiates Data Cleaning
Handles Missing Values, Removes Noise/Normalizes Data
Passes Cleaned Data
Selects Relevant Features (e.g., BP, Cholesterol)
Provides Processed Data
Iteratively Trains Model (Weak Learners → Strong Model)
Generates Prediction Results
Returns Heart Disease Risk (e.g., "High Risk of Coronary Disease")
Requests Real-Time Prediction (Future Enhancement)
Processes New Input Data
Provides Immediate Diagnosis
User
System
DataPreprocessing
FeatureSelection
GradientBoostingModel
DiagnosisModule

ADDITIONAL DEPENDENCIES AND CONSTRAINTS

Dependencies

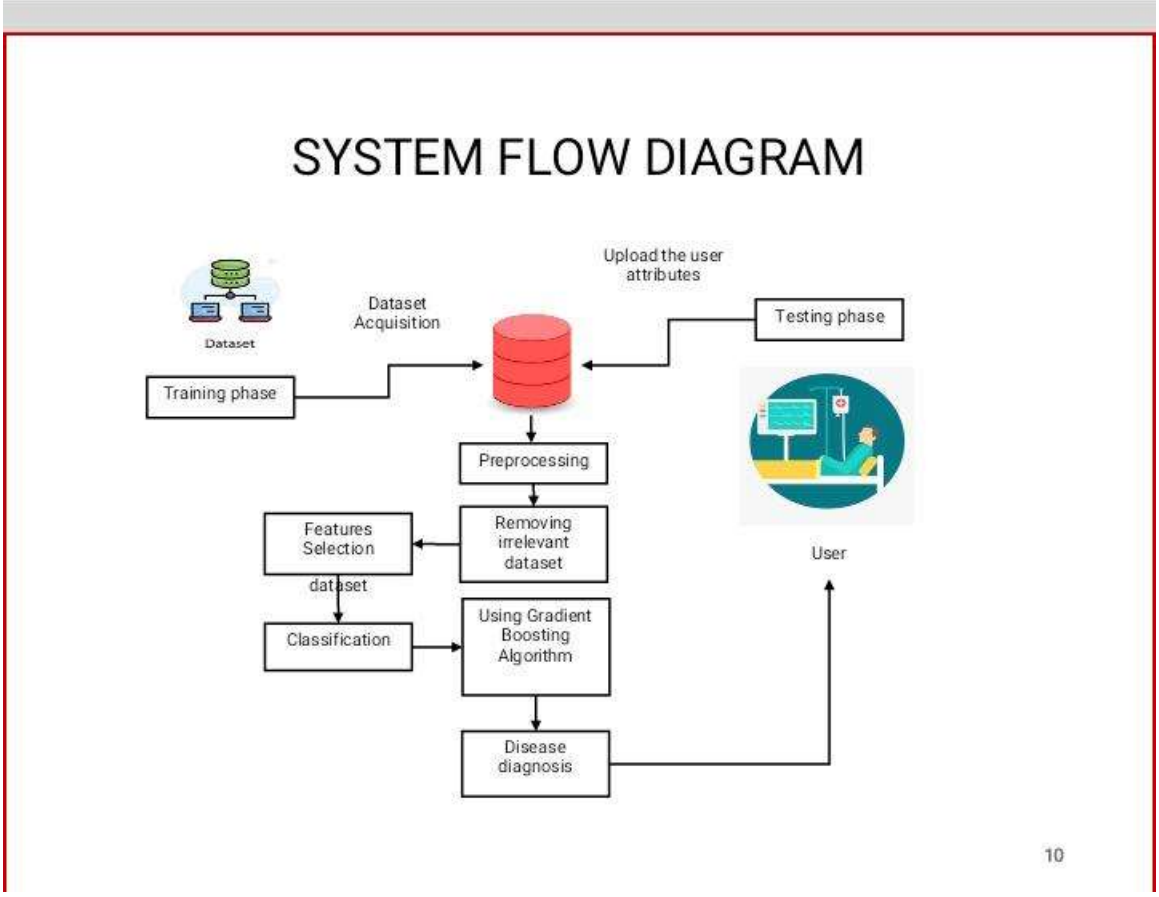
Gradient Boosting Algorithm: Iteratively adds decision trees, learning from mistakes, to build accuracy.
Data Preprocessing: Imputation and removal of outliers ensure data quality.
Feature Engineering: Produces useful features (e.g., BMI = weight/height) to make better predictions.
Missing/Noisy Data: handled through imputation
Missing/Noisy Data: Resolved through imputation and noise elimination.
High False Positives: Reduced through optimization techniques applied to the Gradient Boosting algorithm.
Computational Load: Reduced through effective data handling and parallelization.
Data Acquisition: Acquired varied cardiovascular datasets.
Pre processing: Normalized and cleaned data for model training.
Feature Selection: Determined key predictors (e.g., cholesterol, blood pressure).
Classification: Demonstrated high accuracy through the use of Gradient Boosting.
Diagnosis: Expressed conditions such as coronary disease and cardiac arrest.

Further Constraints

Use bigger datasets to generalize better.

Compare Gradient Boosting with other algorithms (e.g., Neural Networks).Create a real-time prediction interface for clinical use.
Incorporate genetic information for better risk assessment. Future studies involve incorporating large, diverse datasets, tuning hyper parameters, and merging Gradient Boosting with deep learning methods. Improving model interpretability and implementation in real-time clinical environments can enhance prediction accuracy and real-world application in heart disease diagnosis. Future studies can involve feature selection methods, class imbalance handling, and integration of wearable sensor data for on going monitoring. Investigating federated learning can guarantee data privacy while enhancing model performance. Comparative analysis with other ensemble methods can also verify the efficacy of Gradient Boosting in heart disease prediction.

Methodology:



SYSTEM FLOW DIAGRAM

The methodology of research to the heart disease project with machine learning is very systematic to present safe and reliable results. This begins with the attainment of datasets, pertinent datasets with characteristics such as age, gender, blood pressure, cholesterol, and lifestyle. These are obtained from known sources. These data are subsequently preprocessed with the removal of missing values, noisy data, and feature normalization for consistency. Feature selection is then used to determine the most influential features that contribute to the prediction of heart disease. Correlation analysis and dimension reduction methods are used in removing duplicate features and optimizing model efficiency. The project nature is creating the model using the Gradient Boosting algorithm, meaning recursively creating weak predictors (decision trees) which are to be combined to create a strong prediction model. Hyper parameters are tuned to provide the highest possible performance, and the model is tested against accuracy, precision, and recall measures in a bid to realize high reliability. Trained, the model can then be applied to diagnose disease and predict outcomes by classifying patients into risk categories and providing actionable recommendations to health care providers. Missing data, computational complexity, and generalization problems of models are resolved with preprocessing, optimization, and crossvalidation. Improving datasets, other algorithms, and creating predictive tools in real-time to maximize the system utility for clinical applications are areas that will be improved on in further research. Having a universal strategy guarantees the achievement of the project in the early diagnosis of cardiovascular disease and enhancing patient outcomes.

Conclusion:

Ethical issues are always the top priority in such medical usage We've made arrangements to ensure patient data privacy and model explainability. The system will augment, not replace, clinical judgment, and maintain the human touch that is necessary in healthcare decision-making.

As the age of more dynamic healthcare in the digital era unfolds, these kinds of machine learning systems will be increasingly important tools for disease prevention, early detection, and customized treatment protocols. This research adds to the mounting pool of evidence witnessing the revolutionary role of AI in medicine, the benefit of possessing a reproducible model that potentially can be transferred to other applications in healthcare innovation.

This success highlights the value of interdisciplinarity between computer science and medicine. Together with their combination, we can realize clinically relevant and technically advanced technologies - ultimately propelling the shared objective of enhanced patient health around the world. nformation record systems would allow for real-time clinical use. Other studies can investigate hybrid models that blend Gradient Boosting and deep learning structures to achieve even more precise prediction. There is also room to add new biomarkers and genetic data as these become increasingly prevalent in the clinical environment.

Overall, this project marks a milestone in the use of artificial intelligence in cardiovascular medicine. By the convergence of technical excellence in data science with clinical utility, we have created a system that not only represents technical excellence but also has tangible potential to have a positive effect on patient care. In a time of rapidly evolving healhcare in the digital era, these machine learning systems will become increasingly priceless assets in disease prevention, early diagnosis, and personalized therapy regimens. The study adds to the expanding evidence base for the therapeutic use of AI, with the added advantage of having a reproducible model that can be transferred to other domains of healthcare innovation.

This achievement brings out the merit of computer science-medicine interdisciplinary collaboration. Through unifying these two disciplines, we can create technology that is technologically sophisticated and clinically meaningful - ultimately for the common purpose of improved worldwide patient health.

The foundation of our system is its end-to-end data processing pipeline. Beginning from high-quality datasets with the principal clinical parameters - demographic information, vital signs, blood chemistry markers, and lifestyle information - we used stringent preprocessing techniques.

References:

[1] (2023). World Health Organization. Cardiovascular Diseases (CVDs). Accessed: May 5, 2023. [Online]. Available: https://www.afro.who.int/ health-topics/cardiovascular-diseases

[2] Z. Alom, M. A. Azim, Z. Aung, M. Khushi, J. Car, and M. A. Moni, ''Early stage detection of heart failure using machine learning techniques,'' in Proc. Int. Conf. Big Data, IoT, Mach. Learn., Cox's Bazar, Bangladesh, 2021, pp. 23–25.

[3] S. Gour, P. Panwar, D. Dwivedi, and C. A. Mali, ''Machine learning approach for heart attack prediction,'' in Intelligent Sustainable Systems. Singapore: Springer, 2022, pp. 741–747.

[4] C. Gupta, A. Saha, N. S. Reddy, and U. D. Acharya, ''Cardiac disease prediction using supervised machine learning techniques,'' J. Phys., Conf. Ser., vol. 2161, no. 1, 2022, Art. no. 012013.

[5] K. Shameer, ''Machine learning predictions of cardiovascular disease risk in a multi-ethnic population using electronic health record data,'' Int. J. Med. Inform., vol. 146, Feb. 2021, Art. no. 104335.

[6] M. Liu, X. Sun, Y. Liu, X. Yang, Y. Xu, and X. Sun, ''Deep learning- based prediction of coronary artery disease with CT angiography,'' Jpn.

J. Radiol., vol. 38, no. 4, pp. 366–374, 2020.

[7] N. Zakria, A. Raza, F. Liaquat, and S. G. Khawaja, ''Machine learning based analysis of cardiovascular disease prediction,'' J. Med. Syst., vol. 41, no. 12, p. 207, 2017.

[8] M. Yang, X. Wang, F. Li, and J. Wu, ''A machine learning approach to identify risk factors for coronary heart disease: A big data anal- ysis,'' Comput. Methods Programs Biomed., vol. 127, pp. 262–270, Apr. 2016.

[9] C. Ngufor, A. Hossain, S. Ali, and A. Alqudah, ''Machine learning algo- rithms for heart disease prediction: A survey,'' Int. J. Comput. Sci. Inf. Secur., vol. 14, no. 2, pp. 7–29, 2016.

[10] A. Shoukat, S. Arshad, N. Ali, and G. Murtaza, ''Prediction of cardiovas- cular diseases using machine learning: A systematic review,'' J. Med. Syst., vol. 44, no. 8, p. 162, Aug. 2020.

[11] G. R. Shankar, K. Chandrasekaran, and K. S. Babu, ''An Analysis of the Potential Use of Machine Learning in Cardiovascular Disease Prediction,''

J. Med. Syst., vol. 43, no. 12, p. 345, Mar. 2019.

[12] N. Khandadash, E. Ababneh, and M. Al-Qudah, ‘‘Predicting the risk of coronary artery disease in women using machine learning techniques,’’ J. Med. Syst., vol. 45, p. 62, Apr. 2021.

[13] S. Moon, W. Lee, and J. Hwang, ‘‘Applying machine learning to pre- dict cardiovascular diseases,’’ Healthcare Inform. Res., vol. 25, no. 2,

pp. 79–86, Jun. 2019.

[14] M. Lakshmi and A. Ayeshamariyam, ‘‘Machine learning techniques for prediction of cardiovascular risk,’’ Int. J. Adv. Sci. Technol., vol. 30, no. 3,

pp. 11913–11921, Mar. 2021.

[15] M. R. Hassan, S. Huda, M. M. Hassan, J. Abawajy, A. Alsanad, and

G. Fortino, ‘‘Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion,’’ Inf. Fusion, vol. 77, pp. 70–80, Jan. 2022.

[16] S. P. Mikles, H. Suh, J. A. Kientz, and A. M. Turner, ‘‘The use of model constructs to design collaborative health information technologies: A case study to support child development,’’ J. Biomed. Informat., vol. 86,

pp. 167–174, Dec. 2018.

[17] M. R. Delavar, M. Motwani, and M. Sarrafzadeh, ‘‘A comparative study on feature selection and classification methods for cardiovascular disease diagnosis,’’ J. Med. Syst., vol. 39, no. 9, p. 98, Sep. 2015.

[18] C. Puelz, S. Acosta, B. Rivière, D. J. Penny, K. M. Brady, and C. G. Rusin, ‘‘A computational study of the Fontan circulation with fenestration or hep- atic vein exclusion,’’ Comput. Biol. Med., vol. 89, pp. 405–418, Feb. 2017.

[19] Q. Z. Mirza, F. A. Siddiqui, and S. R. Naqvi, ‘‘The risk prediction of cardiac events using a decision tree algorithm,’’ Pakistan J. Med. Sci., vol. 36, no. 2,

pp. 85–89, Mar./Apr. 2020.

[20] A. Farag, A. Farag, and A. Sallam, ‘‘Improving heart disease prediction using boosting and bagging techniques,’’ in Proc. Int. Conf. Innov. Trends Comput. Eng. (ITCE), Mar. 2016, pp. 90–96.

[21] S. Jhajhria and R. Kumar, ‘‘Predicting the risk of cardiovascular dis- eases using ensemble learning approaches,’’ Soft Comput., vol. 24, no. 7,

pp. 4691–4705, Jul. 2020.

[22] N. Samadiani, A. M. E Moghadam, and C. Motamed, ‘‘SVM-based classification of cardiovascular diseases using feature selection: A high-dimensional dataset perspective,’’ J. Med. Syst., vol. 40, no. 11,

p. 244, Nov. 2016.

[23] X. Zhang, Y. Zhang, X. Du, and B. Li, ‘‘Application of XGBoost algorithm in clinical prediction of coronary heart disease,’’ Chin. J. Med. Instrum., vol. 43, no. 1, pp. 12–15, 2019.

[24] Y. Liu, X. Li, and J. Ren, ‘‘A comparative analysis of machine learn- ing algorithms for heart disease prediction,’’ Comput. Methods Programs Biomed., vol. 200, Nov. 2021, Art. no. 105965.

[25] N. S. Hussein, A. Mustapha, and Z. A. Othman, ‘‘Comparative study of machine learning techniques for heart disease diagnosis,’’ Comput. Sci. Inf. Syst., vol. 17, no. 4, pp. 773–785, 2020.

[26] S. Akbar, R. Tariq, and A. Basharat, ‘‘Heart disease prediction using dif- ferent machine learning approaches: A critical review,’’ J. Ambient Intell. Humanized Comput., vol. 11, no. 5, pp. 1973–1984, 2020.

[27] A. Zarshenas, M. Ghanbarzadeh, and A. Khosravi, ‘‘A comparative study of machine learning algorithms for predicting heart disease,’’ Artif. Intell. Med., vol. 98, pp. 44–54, Oct. 2019.

[28] I. Kaur G. Singh, ‘‘Comparative analysis of machine learning algorithms for heart disease prediction,’’ J. Biomed. Inform., vol. 95, Jul. 2019, Art. no. 103208.

[29] Y. Li, W. Jia, and J. Li, ‘‘Comparing different machine learning methods for predicting heart disease: A telemedicine case study,’’ Health Inf. Sci. Syst., vol. 6, p. 7, Dec. 2018.

[30] X. Zhang, Y. Zhou, and D. Xie, ‘‘Heart disease diagnosis using machine learning and expert system techniques: A survey paper,’’ J. Med. Syst., vol. 42, no. 7, p. 129, 2018.