Grenze International Journal of Engineering and Technology, June Issue

Heart Disease Prediction: Optimizing Accuracy through Hyperparameter Tuning

Nawed Alam[1], Farhan Khan[2], Moon Chatterjee[3], Sagarika Chowdhury[4] and Tanmoy Ghosh[5]
[1-5]Narula Institute of Technology/Department of CSE, Kolkata, India
Email: nawedalam890@gmail.com, {farhan.khanjsr657, moonchatterjee456, sagsaha2004, tanmoy.g.331}@gmail.com

Abstract*—Machine learning (ML) and Artificial intelligence (AI) play pivotal roles in diverse domains, particularly in handling the vast influx of data in recent times. These technologies offer potential for more accurate and expedited decision-making, particularly in predicting diseases. A significant health concern is cardiovascular disease, a prominent factor contributing to mortality in recent times. Detecting heart disease in its early stages can significantly impact survival rates and quality of life. Prompt diagnosis is crucial, given the complexity and varied causes associated with this ailment. The inquiry primarily focuses on improving the accuracy of disease prognosis through the application of Machine Learning approaches like Support Vector Machine, Random Forest and XGBoost. A dedicated effort was made to refine the accuracy of SVM by fine-tuning hyperparameters through GridSearchCV. The comprehensive evaluation of these algorithms revealed that SVM achieved an accuracy of 80.22%, XGBoost exhibited 75.82% accuracy, while Random Forest notably outperformed both, boasting an accuracy of 85.71%. These results underscore the strong predictive abilities of Random Forest in detecting heart disease, outperforming the efficacy of the other models. The study's results highlight the capability of machine learning, particularly Random Forest, in forecasting heart disease. Collaborative efforts with medical professionals and ongoing research aim to further refine these models, pushing accuracy rates closer to 100%. This advancement could mark a significant breakthrough in machine learning-driven disease prediction algorithms. In summary, this study showcases the prognostic capacities of SVM, Random Forest, and XGBoost in heart disease prognosis. Emphasizing the superior performance of Random Forest, the findings advocate for its adoption in improving disease diagnostics and intervention strategies.

***Index Terms*— cardiovascular disease, hyperparameter tuning, machine learning, random forest, support vector machine, and xg-boost.**

I. INTRODUCTION

In the continuously growing field of healthcare, the fusion of machine learning and medical science has prompted a substantial change in how we predict and proactively tackle heart disease. The world is grappling with the staggering statistics of 18.6 million global deaths attributed to cardiovascular diseases (CVDs) [1], Underscoring the imperative to harness the capabilities of sophisticated machine models such as Support Vector Machine, Random Forest, and XGBoost, to reshape the precision and effectiveness of Heart Disease Prediction Systems (HDPS). ML is described as the manipulation and retrieval of implicit, previously unexplored or known, and potentially valuable information from data [2], stands as a versatile tool in our quest for improved predictive models. The incorporation of classifiers, including Supervised and Unsupervised, plays a pivotal role in forecasting and assessing the precision of datasets. Our emphasis on HDPS emerges

Grenze ID: 01.GIJET.10.2.541

from its capacity to influence the well-being of numerous individuals by facilitating the timely identification of issues and intervention in cardiovascular diseases, the predominant cause of adult mortality globally.

Cardiac disorders involve a range of conditions that impact the heart, demanding comprehensive and accurate predictive models for effective mitigation. Our project addresses this urgency by utilizing medical history data to Forecast individuals susceptible to heart disease, streamlining diagnosis, minimizing the need for extensive medical tests, and facilitating timely and effective treatments. This endeavor is not merely a technological exploration but a significant stride towards the precision healthcare that can positively impact public health outcomes.

A. Algorithms used

Our research specifically focuses on three robust data mining techniques: Support Vector Machine, XGBoost, and Random Forest Classifier. The comprehensive integration of these algorithms, operating under the umbrella of supervised learning, endeavours to enhance the accuracy and efficiency of HDPS. The significance of our project is underscored by an impressive accuracy rate of 80.2%, signifying a substantial improvement over previous systems reliant on a single data mining technique. The synergy achieved through the integration of SVM, XGBoost, and Random Forest Classifier has positioned our HDPS as a cutting-edge tool in the realm of cardiovascular health. The predictive capabilities of our system are refined using a dataset acquired from the UCI repository, a widely acknowledged and reliable resource within the machine learning community. This dataset encapsulates diverse factors like age, gender, chest pain, fasting sugar levels, and additional factors, providing a comprehensive foundation for our predictive models. The intricate dance between these attributes is decoded through the lenses of SVM, XGBoost, and Random Forest Classifier, collectively analysing 14 medical characteristics to predict the likelihood of a patient being diagnosed with cardiovascular heart conditions. The motivation behind employing multiple algorithms lies in the acknowledgment of their unique strengths and complementary attributes. Support Vector Machine, known for its proficiency in high-dimensional spaces, collaborates with the ensemble learning capabilities of Random Forest and the gradient boosting prowess of XGBoost. This strategic integration aims not only to forecast the onset of cardiovascular conditions but to extend the limits of predictive accuracy, offering a detailed understanding of the complex interplay of factors influencing cardiovascular health.

II. RELATED WORK

In this segment, we explore diverse methodologies and strategies utilized in the forecasting predicting heart conditions through machine learning models. This overview is derived from the research conducted by individuals in the field, offering a comprehensive insight: Numerous studies have investigated data mining's role in forecasting heart diseases, with researchers like Arora examining the emergence of significant patterns and insights from extensive datasets [3]. Arora's accuracy assessments across diverse machine learning and data mining approaches favoured support vector machines (SVM) as the optimal algorithm. Another study utilized the University of California Irvine (UCI) ML dataset, consisting of 303 data entries and 14 input features, to perform an exhaustive assessment of diverse machine learning algorithms [4]. The findings supported SVM as the superior choice, contrasting it with k-Nearest Neighbours and Decision trees. Aditi focused on a multi-layer perceptron model integrated with CAD technology for heart disease prediction, emphasizing the potential of predictive systems to increase disease awareness and reduce mortality rates [5]. Avinash Goland introduced a study on cardiovascular disease prediction utilizing efficient machine learning methods. The research incorporates K-nearest Neighbour, Decision Tree, Naive Bayes, and various other data mining procedures to distinguish different forms of cardiovascular diseases [6]. Shah attempted to forecast heart disease employing a ML algorithm., using the same dataset comprising 303 instances and 17 attributes from the University of California Irvine (UCI) machine learning repository [7]. Gupta has addressed various machine learning classifiers and Pearson correlation coefficients for addressing the absence of data in the UCI dataset [8]. Additionally, in the quest to identify the optimal combination of predictors for cardiovascular disease, Gazeloglu [9] examined eighteen machine learning models, with the incorporation of three feature selection techniques, applied to the UCI dataset consisting of 303 data entries and 13 attributes. In contrast to the aforementioned study, the amalgamated classifier is applied, employing six machine learning models on both the UCI dataset [10] and the IEEE Cardiovascular datasets [11]. This research consisted of six machine learning algorithms which were employed: random forest, k nearest neighbor, linear regression, naïve bayes, gradient boosting, and adaptive boosting. A strategy involving tuning of hyperparameters and a five-fold cross-validation approach has been employed to attain optimal accuracy results before deploying the models. Siddhartha assembled a novel

dataset by amalgamating five renowned cardiovascular disease datasets— Hungary, Switzerland, Statlog, Cleveland, and Long Beach VA. This consolidated dataset encompasses every common feature found in the five individual datasets. [12]. Moreover, recent advancements emphasize the interpretability and explain ability of machine learning models in the medical domain. Ensuring the transparency of predictive models assists medical professionals in comprehending the rationale behind predictions, enabling more informed decisions in clinical settings.

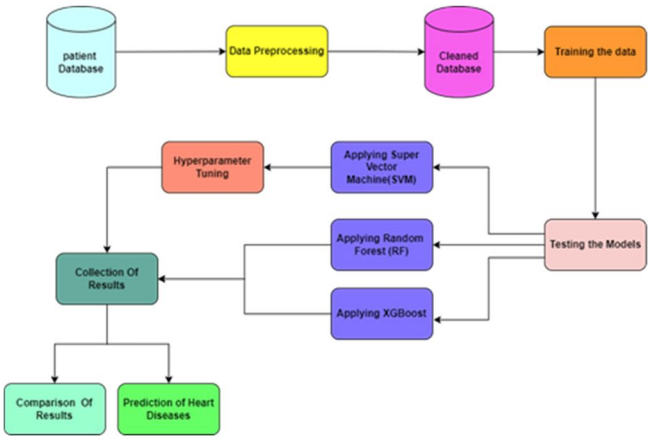

Figure 1. Proposed Model

III. THEORY

This paper offers an in-depth analysis of various machine learning models, with specific emphasis on Support Vector Machine, XGBoost, and Random Forest. The objective is to offer practical support to medical professionals and analysts in enhancing the accuracy of heart disease diagnosis. To accomplish this, the study examines journals, published papers, and recent data pertaining to cardiovascular disease. This approach entails a sequential process, commencing with data collection and subsequently extracting significant values, and then the pre-processing phase. The pre-processing phase is crucial, focusing on tasks like handling missing values, normalization, and data cleansing all tailored to the selected models for the analysis. After the thorough pre-processing of the data, the subsequent step involves utilizing classifiers for categorizing the processed data. Fig. 1 displays the proposed model. The classifiers utilized in this proposed model include Support Vector Machine, XGBoost, and Random Forest. Elaborating on the effectiveness analysis, the study employs diverse ML models like SVM, Random Forest, and XGBoost. The selection of features in the dataset is a key aspect of the analysis, aiming to optimize the correctness of heart disease forecasting. The study particularly underscores the utilization of feature selection to pinpoint the most efficient predictive model. This investigation relies on the empirical foundation provided by the Cleveland heart disease dataset, sourced from the Kaggle Machine Learning database. Besides the technical facets of the investigation, the paper recognizes the potential benefits of integrating clinical decision support with computer based patient records. This integration is seen as a strategy to reduce medical errors, enhance precision, and ultimately improve patient safety. The study envisions a method that considers risk factors associated with cardiac diseases, emphasizing the acquisition of patients' historical information as a vital component of the forecasting procedure. The research utilizes an online open-source platform to promote the development of recommended machine learning models, underscoring the significance of accessibility and collaboration in advancing predictive models for heart diseases. The culmination of the study involves a meticulous comparison of accuracy and performance across different classification models, offering valuable perspectives on the practical applications of machine learning in forecasting and managing cardiac conditions.

IV. EXPERIMENTAL METHOD

This paper aims to predict heart disease probability through computerized methods, benefiting medical practitioners and patients. This paper utilizes three distinct machine learning models to predict heart disease. These models are as follows:

A. Support Vector Machine(SVM)

Support Vector Machines is known for being a strong tool in supervised learning, especially crafted for tasks involving classification. Its primary objective is to ascertain the optimal hyperplane, ensuring a maximum separation between data points of different classes, the positioning of this hyperplane is strategic, aiming to optimize the margin [13]. The margin, denoted as the gap between the hyperplane and the nearest data points (commonly known as support vectors), plays a crucial role in this approach. SVM proves effective in managing datasets featuring both linear and non-linear separations. This adaptability is achieved by incorporating diverse kernel functions, including linear, polynomial, and radial basis function. These kernels enable the transformation of data points into higher dimensional spaces, improving the algorithm's effectiveness in data separation.

B. XGBoost (Extreme Gradient Boosting)

XGBoost, short for Extreme Gradient Boosting, stands out as a highly optimized gradient boosting method known for its exceptional performance and speed. This boosting algorithm assembles a collection of less powerful learners, typically decision trees, in a sequential manner, learning from the errors of previous models. It skilfully mitigates overfitting by gradually adding trees, minimizing a specific loss function. Renowned for its efficiency, rapidity, and accurate predictions, XGBoost has gained popularity in diverse machine learning competitions. However, achieving its optimal performance necessitates meticulous attention when fine-tuning hyperparameters to avert overfitting [14].

C. Random Forest

Random Forest is a machine learning algorithm utilizing numerous decision tree to build a robust and accurate predictive model. Each decision tree is built using a segment of the training set along with a randomly selected collection of features. The Random Forest approach involves creating multiple decision trees, and their outcomes are aggregated through a majority voting process (in the case of classification) or averaging (for regression) [15]. This approach effectively addresses overfitting, improving prediction accuracy by reducing variance. The appeal of Random Forest lies in its adeptness at handling voluminous datasets [16] and its provision of feature importance rankings.

V. RESULTS AND DISCUSSION

In this study, the dataset utilized is the Cleveland Heart Disease Dataset, and pre-processing has been implemented to ensure efficient data processing.

A. Dataset

The patient database comprises datasets obtained from the Cleveland Heart Disease Dataset, accessible on the UCI Repository [17]. The database encompasses a total of 303 patient records. Initially, the dataset included 76 attributes, but in the majority of research, only a subset of 14 attributes is utilized. This subset consists of 13 input attributes and 1 output attribute. The specifics of these attributes are presented in Table 1.

B. Data preprocessing

In this stage, information is extracted from the UCI Dataset in a standardized manner, involving the transformation of data by addressing missing fields, normalizing values, and removing outliers. All categorical values in the dataset have been converted to integer values for processing by diverse algorithms. Subsequently, a check for null values is performed, and if identified, appropriate measures are taken for their removal.

	precision	recall	f1-score	support
0	0.62	0.82	0.71	49
1	0.67	0.43	0.52	42
accuracy			0.64	91
macro avg	0.65	0.62	0.61	91
weighted avg	0.64	0.64	0.62	91

Figure 2. Support Vector Machine

	precision	recall	f1-score	support
0	0.76	0.80	0.78	49
1	0.75	0.71	0.73	42
accuracy			0.76	91
macro avg	0.76	0.76	0.76	91
weighted avg	0.76	0.76	0.76	91

Figure 3. XGBoost Algorithm

TABLE I. ATTRIBUTES DETAILS USED FROM CLEVELAND DATASET

Variable Name	Characteristics		
	Description	Role	Type
age	Patient's age in years	Feature	Integer
sex	Indicates the gender of the patient	Feature	Categorical
cp	Reflects the intensity of chest pain experienced by the patient	Feature	Categorical
chol	Cholesterol levels in milligrams per deciliter (mg/dL)	Feature	Integer
testbps	Blood pressure at rest upon admission to the hospital	Feature	Integer
restecg	Finding from the resting electrocardiogram	Feature	Categorical
fbs	Elevated fasting blood sugar: >120 mg/dL	Feature	Categorical
thalach	Indicates the maximum heart rate of the patient	Feature	Integer
exang	Utilized to detect the presence of exercise-induced angina	Feature	Categorical
oldpeak	ST depression provoked by exercise in comparison to the resting state	Feature	Integer
slope	Characterizes the patient's condition during maximum physical exertion	Feature	Categorical
ca	Count of major vessels (0-3) displaying coloration under fluoroscopy	Feature	Integer
thal	Diagnostic tests recommended for patients experiencing difficulty in breathing	Feature	Categorical
target	Heart disease diagnosis	Target	Integer

C. Training and Testing of Models

The attributes outlined in Table 1 serve as inputs for the aforementioned ML algorithms. The dataset is split, with 70% used for training and the remaining 30% for testing. The training set teaches the model, and the testing set checks how well it learned. The performance of each algorithm is then calculated and examined using diverse metrics, encompassing precision, recall, accuracy, and F1 scores, as further elucidated. The Train Test Split function from the scikit-learn library, an essential tool in machine learning, is used to partition datasets into training and testing subsets. Scikit-learn, which is a machine learning library in python, was instrumental in this process, it provides a diverse set of tools for various tasks, covering classification, regression, dimensionality reduction, and clustering. Its optional parameter, random state, guarantees result reproducibility by fixing the random seed during data splitting. This ensures consistency when assessing and comparing machine learning models, enhancing the reliability of the evaluation process. After training the models we have tried to Anticipating the onset of cardiovascular disease using the UCI dataset, the classification reports for all three models—Support Vector Machine, XGBoost algorithm, and Random Forest algorithm—are illustrated in Fig. 2, Fig. 3, and Fig. 4 respectively. The performance of each of the three models in the experiment conducted till now resulted in Support vector machine 63.74%, Random Forest 85.71% and XGBoost algorithm 75.82%.

D. Hyperparameter tuning in SVM

Tuning hyperparameters in SVM involves optimizing parameters like the choice of regularization parameter (C), kernel, and parameters specific to the kernel (degree for polynomial, gamma for Radial Basis Function), which markedly affect the performance of SVM models. Techniques such as grid search or random search are often utilized to systematically explore a range of hyperparameter values. This process aids in mitigating overfitting or underfitting, amplifying the SVM's capacity to discern intricate patterns within the data, while also ensuring robustness across diverse datasets.

Kernels: A 'kernel' is like a tool that transforms input data so we can draw better boundaries between different groups. These boundaries can be curvy instead of straight. There are different types of tools (kernels) like linear, polynomial, radial basis function (RBF), and sigmoid. Each is good for different kinds of data and complexities.

C(Regularization): The hyperparameter 'C' governs the balance between establishing a simpler decision boundary and precisely categorizing training points. A lower C value promotes a more generalized boundary, whereas a higher C value strives for the accurate classification of all training examples, potentially leading to overfitting if set excessively high.

GAMMA: The 'gamma' parameter, specifically applied in RBF kernels, dictates the impact of individual training samples, where low values imply a broader influence and high values indicate a more localized influence. A

reduced gamma value implies a larger similarity radius, influencing smoothness, while an elevated gamma value results in a more intricate decision boundary around individual data points, potentially causing overfitting.
After careful adjustments to the hyperparameters, Support Vector Machines demonstrated significant improvement in performance, a focal point in the subsequent results section. The fine-tuned parameters were instrumental in boosting the model's effectiveness, underscoring the importance of thoughtful parameter refinement for enhanced machine learning outcomes.

	precision	recall	f1-score	support
0	0.83	0.92	0.87	49
1	0.89	0.79	0.84	42
accuracy			0.86	91
macro avg	0.86	0.85	0.85	91
weighted avg	0.86	0.86	0.86	91

Figure 4. Random Forest Algorithm

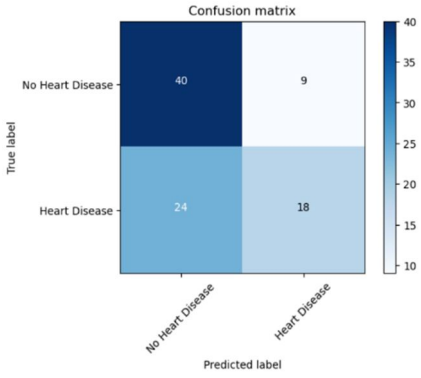

Figure 5. Before Tuning Hyperparameters

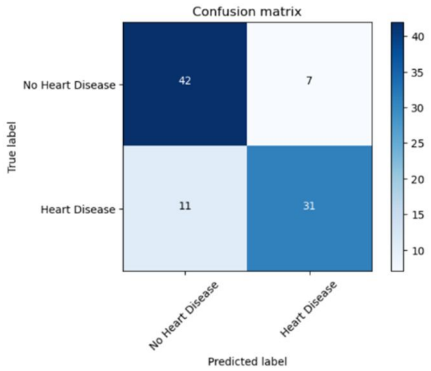

Figure 6. After Tuning Hyperparameters

E. Results

In the Jupyter Notebook, we employed three supervised classification methods. To assess the model's performance, 70% of the data was assigned for training, while the rest of 30% allocated for testing. Initially, we chose the 13 attributes discussed and pre-processed them by removing null values and irrelevant data. Subsequently, the SVM algorithm was trained using a RBF as the kernel, resulting in an initial accuracy of 63.74%. Following this, we refined the SVM hyperparameters, specifically tuning C and Gamma using grid search, With the radial basis function as the kernel., the SVM accuracy continued to evolve. improved to 80.22%. The confusion matrices for predictions before and after hyperparameter tuning are presented above in Fig. 5 and Fig. 6 respectively. Optimizing hyperparameters markedly enhanced the precision of the Support Vector Machine (SVM) when utilizing the Radial Basis Function (RBF) as a kernel. In this comparative analysis shown in Fig. 7, it becomes apparent that the Random Forest approach consistently surpasses the other two algorithms. SVM attained the highest accuracy at 80.22%, with XGBoost achieving 75.82%, and Random Forest reaching an impressive 85.71%. This outcome emphasizes the strength and effectiveness of the Random Forest approach compared to other classification models.

VI. Conclusions

A model for detecting cardiovascular disease has been created utilizing various machine learning classification techniques. This investigation anticipates individuals with cardiovascular disease by analysing patient medical histories associated with fatal heart conditions from a dataset containing comprehensive medical records. The employed algorithms in constructing this model encompass Support Vector Machine, Random Forest Classifier, and XGBoost. The peak accuracy achieved by our models reached 85.71%. The utilization of a larger training dataset, facilitated by dataset balancing, enhances the likelihood of the model accurately predicting whether an individual has a heart disease. Through these methodologies, we can achieve faster and more accurate patient

predictions, contributing to significant cost reductions. There exists a multitude of medical databases where these machine learning techniques prove superior, surpassing human predictions and benefiting both patients and healthcare providers. In summary, this study aids us in predicting patients diagnosed with heart diseases by meticulously cleaning the dataset and implementing hyperparameter tuning.

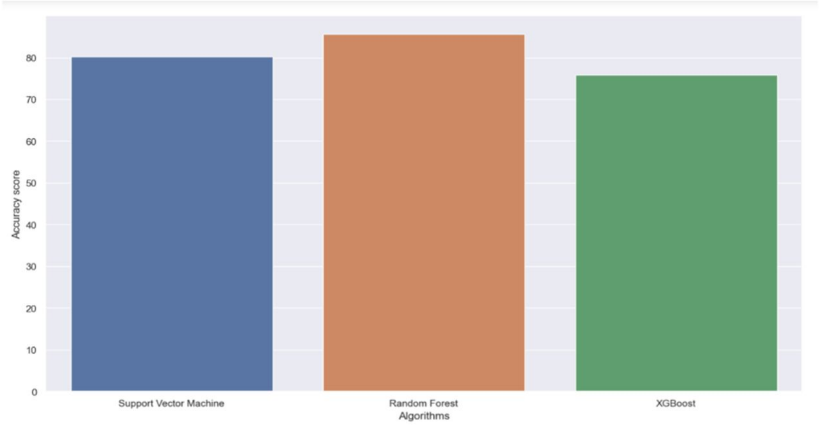

Figure 7. Comparison of all three algorithms

Future investigations could address the limitations by applying hyperparameter tuning to alternative algorithms, including SVM. Moreover, integrating grid search with algorithms like XGBoost and Random Forest has the capacity to enhance precision and offer a more comprehensive performance evaluation. A comprehensive analysis of how missing data and outliers affect model accuracy, coupled with the creation of methods to handle these situations, could be advantageous. Looking ahead, exploring more intricate datasets, incorporating images, and applying diverse deep learning techniques like CNN could be a promising avenue. Enhancements to our techniques may involve experimenting with additional layers and different activation functions to determine optimal compatibility with our dataset. Establishing a user-friendly system for individuals to input their medical records related to heart issues, with the system predicting the likelihood of heart disease diagnoses, is also a prospective development.

REFERENCES

[1] Roth, Gregory A., George A. Mensah, Catherine O. Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M. Baddour, Noël C. Barengo et al. "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study." *Journal of the American College of Cardiology* 76, no. 25 (2020): 2982-3021.

[2] Jindal, Harshit, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath. "Heart disease prediction using machine learning algorithms." In *IOP conference series: materials science and engineering*, vol. 1022, no. 1, p. 012072. IOP Publishing, 2021.

[3] Kaur, Amandeep, and Jyoti Arora. "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY." *International Journal of Advanced Research in Computer Science* 9, no. 2 (2018).

[4] Kumar, M. Nikhil, K. V. S. Koushik, and K. Deepak. "Prediction of heart diseases using data mining and machine learning algorithms and tools." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3, no. 3 (2018): 887-898.

[5] Gavhane, Aditi, Gouthami Kokkula, Isha Pandya, and Kailas Devadkar. "Prediction of heart disease using machine learning." In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pp. 1275-1278. IEEE, 2018.

[6] Golande, Avinash, and T. Pavan Kumar. "Heart disease prediction using effective machine learning techniques." *International Journal of Recent Technology and Engineering* 8, no. 1 (2019): 944-950.

[7] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." *SN Computer Science* 1 (2020): 1-6.

[8] A. Gupta, R. Kumar, H. Singh Arora and B. Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," in IEEE Access, vol. 8, pp. 14659-14674, 2020, doi: 10.1109/ACCESS.2019.2962755.

[9] Gazeloglu C. Prediction of heart disease by classifying with feature selection and machine learning methods . 2020 Jun. 12 [cited 2023 Dec. 31];22(2):660-7.

[10] UCI Machine Learning Repository Heart Disease Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/Heart+ Disease (accessed on 31 December 2023).

[11] IEEE Dataport Heart Disease Dataset. Available online: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive (accessed on 31 December 2023).
[12] Manu Siddhartha Heart Disease Dataset (Comprehensive). Available online: https://ieee-dataport.org/authors/manu-siddhartha (accessed on 31 December 2023).
[13] Pattnaik P P, Padhy S R, Mishra B S P, Mishra S and Mallick P K 2021 Heart disease prediction using machine learning technique Advances in Electronics, Communication and Computing 709 193-201.
[14] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized xgboost based diagnostic system for effective prediction of heart disease," Journal of King Saud University-Computer and Information Sciences, 2020.
[15] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of heart disease using random forest and feature subset selection," in Innovations in bio-inspired computing and applications. Springer, 2016, pp. 187–196.
[16] S. Asadi, S. Roshan, and M. W. Kattan, "Random forestswarm optimization-based for heart diseases diagnosis," Journal of Biomedical Informatics, vol. 115, p. 103690, 2021.
[17] Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. “Heart Disease UCI Machine Learning Repository”. Heart Disease Dataset. Accessed November 23,2023. https://archive.ics.uci.edu/dataset/45/heart+disease