International Journal of Advancement in Electronics and Computer Engineering (IJAECE)
Volume 14, Issue 02, February. 2025, pp. 108-115 ISSN 2278 -1412

Intelligent Heart Disease Diagnosis: Analyzing Predictive Accuracy of Machine Learning Models

Manoli Balasaheb Charmal[1] , Dr. Tripti Arjariya[2]
[1]Mtech Scholar, charmalmanoli98@gmail.com, Bhabha University, Bhopal, India
[2]Prof., Bhabha University , Bhopal, India

Abstract *- Heart disease remains a leading cause of mortality worldwide, necessitating early diagnosis and preventive healthcare strategies. With the increasing integration of machine learning (ML) into medical analytics, predictive models have become instrumental in enhancing the accuracy and efficiency of heart disease diagnosis. This review investigates recent advancements in heart disease prediction using ML algorithms,Emphasis is placed on widely adopted models such as Decision Trees, Random Forest, Naive Bayes, Logistic Regression, K-Nearest Neighbors, and XGBoost.These algorithms have demonstrated significant potential in identifying complex patterns in clinical datasets, surpassing traditional statistical methods in adaptability and predictive power. Various studies have highlighted the strengths of each algorithm: Decision Trees and Logistic Regression offer high interpretability, while ensemble techniques like Random Forest and XGBoost deliver superior accuracy and robustness. Naive Bayes proves effective with limited data, and KNN is noted for its performance in normalized, noise-free environments.The review also discusses the relevance of datasets like the UCI Cleveland Heart Disease Dataset and the critical role of preprocessing techniques such as normalization, imputation, and feature selection. Evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC are analyzed for their effectiveness in assessing diagnostic performance. Furthermore, hybrid and ensemble methods have shown promise in boosting predictive outcomes through model integration and optimization.This paper concludes by emphasizing the importance of algorithm selection, data quality, and preprocessing in developing reliable ML-based heart disease prediction systems. The insights presented aim to guide future research and support clinical decision-making through intelligent, data-driven solutions.*

Keywords: *Heart Disease, Coronary Artery Disease (CAD), Cardiovascular Risk, Global Health, WHO, Mortality, Public Health Burden, Modifiable Risk Factors, Preventive Healthcare, Atherosclerosis*

I. Introduction

Heart disease encompasses a spectrum of disorders that affect the structure and function of the heart. These range from coronary artery disease (CAD) and arrhythmias to heart failure and structural heart defects. Among these, CAD—characterized by the narrowing or blockage of the coronary arteries due to atherosclerotic plaque buildup—is the most common and deadliest form. The condition impairs blood flow to the heart, potentially leading to angina, myocardial infarction (heart attack), or sudden cardiac death.The World Health Organization (WHO) reports that cardiovascular diseases, including all forms of heart disease, are responsible for an estimated 17.9 million deaths globally each year, which accounts for 32% of all global deaths. Despite medical advancements and awareness programs, the incidence of heart disease remains high, highlighting the need for ongoing research, prevention strategies, and healthcare innovation.

II. Types of Heart Disease

Heart disease is a broad term that encompasses a range of cardiovascular conditions affecting the heart and blood vessels. Each type varies in The most common types include:

II. 1. Coronary Artery Disease (CAD)

Coronary Artery Disease is the most prevalent form of heart disease. It occurs when the

coronary arteries that supply blood to the heart muscle become narrowed or blocked due to plaque buildup (atherosclerosis). This reduces oxygen-rich blood flow, leading to chest pain (angina), heart attacks (myocardial infarction), or heart failure.

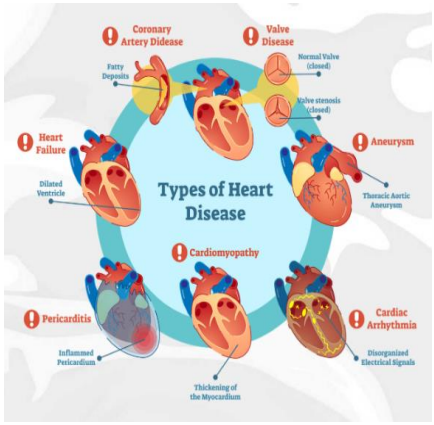


Figure.1 Types of Heart Disease

II. 2. Heart Arrhythmias

Arrhythmias refer to abnormal heart rhythms caused by improper electrical impulses regulating the heartbeat. Types include:

Tachycardia (fast heartbeat), Bradycardia (slow heartbeat)Atrial Fibrillation (AFib): Irregular and often rapid heartbeat, Ventricular Fibrillation: Life-threatening rhythm requiring immediate medical attention

II.3. Heart Failure

Heart failure, also known as congestive heart failure (CHF), occurs when the heart is unable to pump blood efficiently to meet the body's needs. It may result from weakened heart muscles (systolic failure) or stiffened heart chambers (diastolic failure), often due to prolonged CAD, hypertension, or previous myocardial infarction.

II.4. Cardiomyopathy

This disease affects the heart muscle, making it harder for the heart to pump blood. Types include:

Dilated Cardiomyopathy: Enlarged and weakened heart chambers, Hypertrophic Cardiomyopathy: Abnormal thickening of the heart muscle, Restrictive Cardiomyopathy: Stiffness of the heart muscle restricting blood filling

II.5. Congenital Heart Disease

These are structural heart defects present at birth. They may involve abnormalities in the heart walls, valves, or blood vessels. Common types include:

Atrial Septal Defect (ASD),Ventricular Septal Defect (VSD),Tetralogy of Fallot Some may require surgical correction in infancy or early childhood.

II. 6. Valvular Heart Disease

This occurs when one or more of the heart's valves do not function properly, affecting blood flow direction. Conditions include:

Aortic Stenosis: Narrowing of the aortic valve, Mitral Valve Prolapse ,Regurgitation or Insufficiency: Leaking of blood backward due to valve dysfunction

II. 7. Pericardial Disease

Pericardial diseases involve inflammation or infection of the pericardium, the sac surrounding the heart. Examples:

Pericarditis: Inflammation of the pericardium, Pericardial Effusion: Fluid buildup around the heart, Constrictive Pericarditis

II. 8. Rheumatic Heart Disease

A complication of rheumatic fever, often following untreated streptococcal throat infections. It results in permanent damage to the heart valves due to inflammation.

II.9. Ischemic Heart Disease (IHD)

This refers broadly to conditions caused by reduced blood flow to the heart, often used interchangeably with CAD. It can lead to chest pain and increased risk of heart attack.

II. 10. Inflammatory Heart Disease

This includes conditions where the heart tissues become inflamed due to infections or autoimmune reactions, such as:Myocarditis (inflammation of the heart muscle), Endocarditis (inflammation of the inner heart lining)

III. Method

To address the challenge of early detection of heart disease, this study proposes a methodological framework that integrates data preprocessing, feature engineering, algorithm selection, and model validation. The objective is to develop robust machine learning models capable of accurately identifying early signs and risk factors associated with heart disease onset.

Data Collection

Acquire a comprehensive dataset containing diverse patient information, including demographic details, medical history, lifestyle factors (e.g., smoking, exercise), and clinical test results (e.g., blood pressure, cholesterol levels, electrocardiogram).

Data Preprocessing

Handling Missing Data: Impute missing values using appropriate techniques such as mean/mode imputation or advanced imputation methods like K-nearest neighbors (KNN) or predictive modeling. Normalization and Scaling: Normalize numerical features to a standard scale (e.g., using Min-Max scaling or Z-score normalization) to ensure uniformity and mitigate the impact of different feature scales. Adjust the range of numerical features to a common scale, typically [0, 1]. For example, Min-Max scaling where each feature value is rescaled as follows:

$$x_{norm} = x - x_{min} / x_{max} - x_{min}$$

Standardization: Transform features to have a mean of 0 and a standard deviation of 1, which is particularly useful for algorithms like SVM and KNN. This can be done using

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

Encoding Categorical Variables

To ensure compatibility with machine learning algorithms, categorical variables must be converted into numerical representations:

One-Hot Encoding: Converts each category into a binary feature. For instance, smoking status categories ("current", "former", "never") become three distinct binary features.

Label Encoding: Assigns unique integers to categories, ideal for ordinal features like physical activity levels ("low", "medium", "high" → 0, 1, 2).

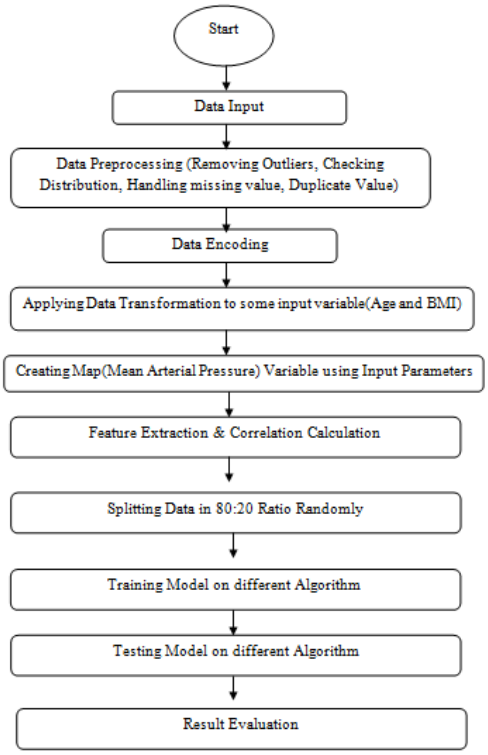


Figure .2 Proposed Flow Chart

Data Transformation (Age and BMI)

Normalization is applied to scale continuous variables like Age and BMI to a [0, 1] range. This prevents features with larger scales from dominating the model's learning process and enhances algorithm performance.

Mean Arterial Pressure (MAP) Calculation

MAP, a critical cardiovascular indicator, is calculated using:

$$MAP = (SBP + 2 \times DBP) / 3$$

This derived feature enhances the dataset's clinical relevance.

Feature Extraction

Feature extraction involves selecting and transforming relevant variables from raw data. This includes:

Demographic (Age, Sex) ,Lifestyle (Smoking, Diet, Physical Activity)

Clinical (BP, Cholesterol, Blood Sugar, BMI) Derived features such as Pulse Pressure (SBP − DBP) and BMI categories are also used. Data cleaning (handling missing values and outliers), normalization, and standardization (Z-score) are essential preprocessing steps.

Dataset Splitting

Data is randomly split into 80% training and 20% testing subsets using train_test_split from scikit-learn, with random_state set for reproducibility.

Model Training

Seven machine learning models are evaluated:

Logistic Regression: Binary classification using a sigmoid function.K-Nearest Neighbors (KNN): Classifies based on majority vote of k closest data points using Euclidean distance.

Naive Bayes: Probabilistic classifier assuming feature independence.

Decision Tree: Splits data based on feature values to build a classification tree.

Extreme Gradient Boost: Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting, designed for speed and performance. It builds an ensemble of decision trees sequentially, where each new tree corrects errors made by the previous ones.

XGBoost minimizes a regularized objective function, which combines a loss function measuring the model's predictive accuracy and a regularization term controlling model complexity. This helps prevent overfitting.

IV. Result

The performance of various machine learning algorithms for heart disease prediction was evaluated using key metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). The dataset was split into training (80%) and testing (20%) sets, ensuring balanced class distribution. Each model was trained on the preprocessed data and tested on the unseen portion of the dataset.

Table 1 Data Distribution

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

It's important to note that personal identifiers such as names and social security numbers of the patients have been removed from the database and replaced with dummy values to ensure privacy and confidentiality. This dataset serves as a valuable resource for analyzing and predicting heart disease, providing a wide range of clinical and demographic data points for robust analysis.

Exploratory Data Analysis

In this section, we perform an exploratory data analysis (EDA) to understand the structure and distribution of the dataset. We begin by visualizing key features to observe patterns, trends, and potential outliers

Age Distribution

The 'Age' attribute in this dataset provides detailed information on the distribution of patients' ages. The minimum recorded age is 29, while the maximum is 77, resulting in a range of 48 years. The quantile statistics offer further insights into the age distribution. The 5th percentile is at 39.1 years, the first quartile

(Q1) is 47.5 years, the median age is 55 years, the third quartile (Q3) is 61 years, and the 95th percentile is 68 years. This indicates that the middle 50% of the ages fall within an interquartile range (IQR) of 13.5 years, specifically between 47.5 and 61 years.

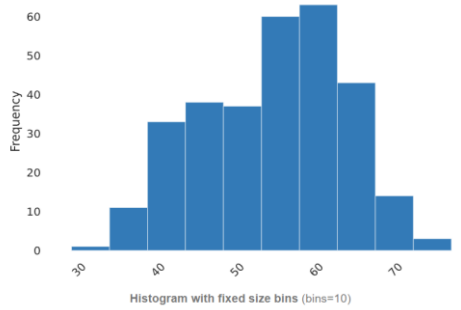


Figure 3 Age Distribution

From the descriptive statistics, we observe that the mean age of the patients is 54.36 years, closely aligning with the median age of 55 years. This suggests a roughly symmetrical age distribution, further supported by the skewness value of -0.202, indicating a slight negative skew. T

Resting blood pressure

The attribute 'resting blood pressure' in this dataset features a distinct count of 49, indicating that there are 49 unique values present within the dataset. This constitutes 16.2% of the data, highlighting a relatively low uniqueness for this attribute.

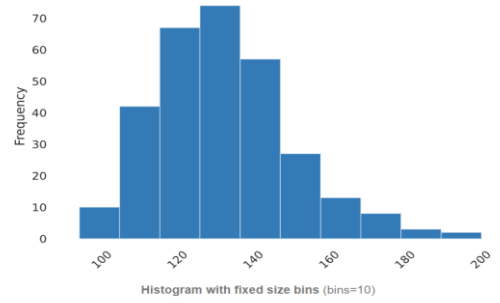


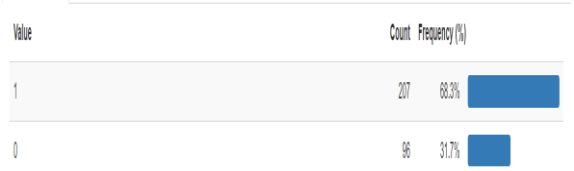
Figure 4 Blood Pressure

The mean resting blood pressure in the dataset is 131.62, providing a central tendency measure that suggests a moderately high average blood pressure among the patients. The minimum recorded value for resting blood pressure is 94, while the maximum value is 200, indicating a wide range of blood pressure levels among the dataset's subjects.

Notably, there are no zero values in the dataset for this attribute, further indicating that all entries are valid resting blood pressure measurements. The zeros percentage is thus 0.0%, affirming the absence of non-informative data points.

Gender Distribution

Table:2 Gender Distribution



Data Preprocessing

After initial data exploration, several key steps are taken to prepare the dataset for further analysis and modeling. Each step serves a specific purpose, from data cleaning and encoding to balancing and scaling the dataset.

First, the categorical columns in the dataset are identified. Categorical columns contain non-numeric data, which cannot be directly used in most machine-learning algorithms. Therefore, these columns need to be encoded into numeric formats. The column "Gender" is encoded by replacing "Female" with 0 and "Male" with 1, making it a binary numeric variable.

After encoding, the original categorical columns are dropped from the dataset, as they have been replaced by their encoded counterparts. This ensures that all features in the dataset are numeric, which is a requirement for most machine-learning algorithms.

A correlation matrix is then computed to identify the relationships between different features. A heatmap is created to visualize these correlations, highlighting features that are highly correlated with each other. Features

with a correlation coefficient greater than or equal to 0.75 are considered highly correlated.

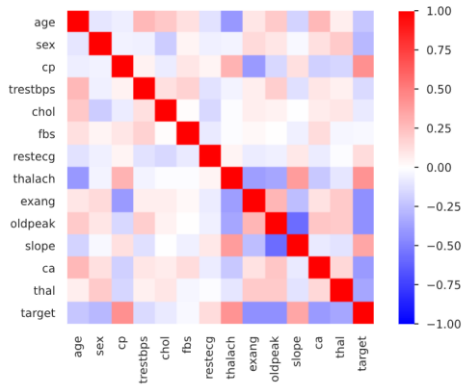


Figure 5 Data preprocess output

These results indicate that the SVC model performed well, with high precision, recall, and F1-scores for both classes. The balanced performance across all metrics highlights the model's capability to accurately classify both positive and negative instances, making it effective for the Heart Disease Dataset.

Table:3 Different Model Evaluation

Model	**Accuracy**
Logistic Regression	85.245902
Naive Bayes	85.245902
Random Forest	86.885246
Extreme Gradient Boost	90.163934
K-Nearest Neighbour	88.524590
Decision Tree	81.967213
Support Vector Machine	88.524590

V. Conclusion

This research comprehensively explored the effectiveness of multiple machine learning algorithms for heart disease prediction using a structured and clinically relevant dataset. The primary objective was to assess the predictive performance of each model and their potential utility in supporting early diagnosis and clinical decision-making in cardiovascular care. Among the evaluated models, Extreme Gradient Boosting (XGBoost) achieved the highest predictive accuracy at 90.16%, highlighting its strength in capturing complex, non-linear patterns within medical data. Close contenders such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) also demonstrated robust performance with accuracies of 88.52%, showcasing their reliability in handling multi-dimensional, nonlinear classification tasks relevant to heart disease. The Random Forest model achieved 86.89% accuracy, benefiting from ensemble learning's ability to reduce overfitting and improve generalization. Logistic Regression and Naive Bayes, while simpler in architecture, delivered consistent baseline results with accuracies of 85.25%, making them suitable for scenarios requiring interpretability and rapid deployment. Meanwhile, the Decision Tree algorithm, although intuitive and easy to implement, recorded a comparatively lower accuracy of 81.97%, suggesting limitations in capturing the intricate relationships present in heart disease risk factors..

References

[1] Pooja Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques" Vol. 9 Issue 11, November-2020, International Journal of Engineering Research & Technology (IJERT).

[2] Rahul Katarya & Sunit Kumar Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis" Volume 11, pages 87–97, (2020).

[3] Ban Salman Shukur, Maad M. Mijwil, "Involving machine learning techniques in heart disease diagnosis: a performance analysis" Vol. 13, No. 2, April 2023, International Journal of Electrical and Computer Engineering (IJECE),DOI: 10.11591/ijece.v13i2.pp2177-2185

[4] Gufran Ahmad Ansari, Salliah Shafi Bhat, Mohd Dilshad Ansari, Sultan Ahmad, Jabeen Nazeer, A. E. M. Eljialy, "Performance Evaluation of Machine Learning Techniques (MLT) for Heart Disease Prediction" Volume 2023, DOI: https://onlinelibrary.wiley.com/doi/10.1155/2023/819126.

[5] M. Chandralekha∗ and N. Shenbagavadivu, "Performance Analysis Of Various Machine Learning

Techniques To Predict Cardiovascular Disease: An Emprical Study", Appl. Math. Inf. Sci. 12, No. 1, 217-226 (2018), Applied Mathematics & Information Sciences, DOI: http://dx.doi.org/10.18576/amis/120121.

[6] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" ,IEEE (2019), DOI : 10.1109/ACCESS.2019.2923707.

[7] Azam Mehmood Qadri, Ali Raza, Kashif Munir, Mubarak S. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning" IEEE, 2023, DOI: 10.1109/ACCESS.2023.3281484.

[8] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, "Heart Disease Prediction using Machine Learning Techniques" IEEE 2020,DOI: 10.1109/ICACCCN51052.2020.9362842

[9] Luis Rolando Guarneros-Nolasco, Nancy Aracely Cruz-Ramos, Giner Alor-Hernández, Lisbeth Rodríguez-Mazahua andJosé Luis Sánchez-Cervantes, "Identifying the Main Risk Factors for Cardiovascular Diseases Prediction Using Machine Learning Algorithms", MPDI 2021, DOI: https://doi.org/10.3390/math9202537

[10] Jian Ping Li; Amin Ul Haq; Salah Ud Din, Jalaluddin Khan, Asif Khan, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare" Volume 8 , IEEE Access 2020, DOI: 10.1109/ACCESS.2020.3001149

[11] Rachana R Sanni , H.S. Guruprasad, "Analysis of performance metrics of heart failured patients using Python and machine learning algorithms" Volume 2, Issue 2, November 2021, KeAi(Chinese Roots Global Impact), DOI: https://doi.org/10.1016/j.gltp.2021.08.028

[12] Mangesh Limbitote, Dnyaneshwari Mahajan, Kedar Damkondwar, Pushkar Patil, "A Survey on Prediction Techniques of Heart Disease using Machine Learning", ISSN: 2278-0181, Vol. 9 Issue 06, June-2020, International Journal of Engineering Research & Technology (IJERT).

[13] Md. Julker Nayeem Nayeem, Sohel Rana, Md. Rabiul Islam, "Prediction of Heart Disease Using Machine Learning Algorithms", Vol 1 Issue 3, European Journal of Artificial Intelligence and Machine Learning, 2022, ISSN:2796-0072, DOI: 10.24018/ejai.2022.1.3.13

[14] Ghulab Nabi Ahamad, Shafiullah, Hira Fatima, Imdadullah, S. M. Zakariya , Mohamed Abbas , Mohammed S. Alqahtani, and Mohammed Usman, "Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease" March 2023, Vol-11, Issue-3, MPDI, DOI: https://doi.org/10.3390/pr11030734

[15] Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam. "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System" Vol.6 No.4, November 2018, Scientific Research An Acadmin Publisher, DOI: 10.4236/wjet.2018.64057

[16] Ramya Perumal, Kaladevi AC, "Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques" Vol. 29, No. 6, ISSN: 2005-4238, (2020), International Journal of Advanced Science and Technology.

[17] Ghulab Nabi Ahmad; Hira Fatima; Shafi Ullah, Abdelaziz Salah Saidi, Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV" 2022, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2022.3165792

[18] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" Volume 3, Issue 3, 2018, International Journal of Scientific Research in Computer Science.

[19] Rehan Ahmed, Maria Bibi, Sibtain Syed, "Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms" Vol. 3 No. 1 (2023), International Journal of Computations, Information and Manufacturing (IJCIM), DOI: https://doi.org/10.54489/ijcim.v3i1.223

[20] Isreal Ufumaka, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction" International Journal of Scientific and Research Publications, Volume 11, Issue 1, January 2021, ISSN 2250-3153, DOI: http://dx.doi.org/10.29322/IJSRP.11.01.2021.p10936

[21] Ch. Anwar ul Hassan, Jawaid Iqbal, Rizwana Irfan, Saddam Hussain, Abeer D. Algarni, Syed Sabir Hussain Bukhari, Nazik Alturki and Syed Sajid Ullah, "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers" Volume 22, Issue 19, 2022, MDPI DOI: https://doi.org/10.3390/s22197227

[22] Khalid Amen, Mohamed Zohdy and Mohammed Mahmoud, "MACHINE LEARNING FOR MULTIPLE STAGE HEART DISEASE PREDICTION", 2020, CSEIT, DOI: 10.5121/csit.2020.101118

[23] Shakila Basheer, Rincy Merlin Mathew, M. Shyamala Devi, "Ensembling Coalesce of Logistic Regression Classifier for Heart Disease Prediction using Machine Learning" ISSN: 2278-3075, Volume-8 Issue-12, 2019, International Journal of Innovative Technology and Exploring Engineering (IJITEE), DOI: DOI: 10.35940/ijitee.L3473.1081219.

[24] Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan Soumyabrata Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction" Volume 2, November 2022, Healthcare Analytics, DOI: https://doi.org/10.1016/j.health.2022.100060

[25] M. Ganesan, N. Sivakumar, "IoT based heart disease prediction and diagnosis model for healthcare using machine learning models" 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), DOI: 10.1109/ICSCAN.2019.8878850

[26] Visvasam Devadoss, Ambeth Kumar, Chetan Swarup, Indhumathi Murugan, Abhishek Kumar, Kamred Udham Singh, Teekam Singh and Ramu Dubey, "Prediction of Cardiovascular Disease Using Machine Learning Technique—A Modern Approach" 2022, vol.71, no.1, DOI:10.32604/cmc.2022.021582.

[27] G. Parthiban and S.K.Srivatsa "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients" International Journal of Applied Information Systems (IJAIS), ISSN : 2249-0868, Volume 3– No.7, August 2012.