IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Prediction of Heart Disease Using Machine Learning Techniques

[1]Moffo Eddie Phiri, [2]Dr. Arul Valan, [3]Dr. Glorindal Selvam

[1] Master of computer science student, [2]Supervisor, [3]Post graduate coordinator
[1]Computer Science,
[1]DMI ST Eugene University, Lusaka, Zambia
[2]NIT, India
[3]DMI ST John the Baptist University, Malawi

Abstract: *According to WHO, heart disease is a leading cause of death globally with an estimated 17.9 million deaths each year. Timely and efficient disease diagnosis plays a very important role in health systems to avoid preventable deaths among people who eventually suffer from acute heart disease. In this paper, we propose an efficient and accurate system to predict heart disease based on supervised machine learning techniques. The system is based on classification algorithms including random forest, then naive bayes and K-nearest neighbor (KNN) to improve its strength with a bagging method, while standard feature selection algorithms will be used in order to obtain the best accuracy and quickly run the model*

Index Terms - heart disease diagnosis, machine learning, feature selection, random forest, K-nearest neighbor, model, CA, ML, SVM, RNN, NB, RF, AUC

I. INTRODUCTION

Heart disease refers to several types of heart conditions. The most common type of heart disease globally is coronary artery disease (CAD), which affects the blood flow to the heart. Sometimes heart disease may be silent and not diagnosed until a person experiences signs or symptoms of a heart attack, heart failure, or an arrhythmia and these eventually may lead to death.

In 2019, according to WHO [1], the world lost an estimated 17.9 million people to heart disease, representing 32% of total global deaths. These are very high numbers. In developing and middle-income countries, this may even be worse because of lack of trained doctors and most cases or deaths are hardly accurately classified as heart disease due to inadequate diagnostic protocols.

In many cases, it takes a lot of months to eventually diagnose someone as having heart disease and at this time a patient may already be suffering from an acute heart disease condition.

Much as this may be the case, health care institutions are collecting and keeping important patient data including for heart disease. We can therefore leverage the availability of this huge data to diagnose heart diseases in early stages where it can be treated. With the correct lifestyle education, most heart conditions can be prevented from developing into heart disease and consequently help nations save millions of dollars through treatment schemes. For example, between 2017 and 2018, the United States spent $229 billion in health care services, medicines and lost productivity due to deaths, according to the Centers of Disease Control and Prevention (CDC) [2]

The field of computer science provides for tools that can be used to analyse historical data and provide future insights. The medical field thus has not been left behind to benefit from data mining, artificial intelligence, machine learning and deep learning. The world needs expert systems or decision support system to help doctors make quicker and timely decisions on patients.

Machine learning (ML) is a computer science field which integrates statistics and artificial intelligence, and the resultant software models learns data patterns and provide prediction of an expected output when presented with similar kind of data in future. Machine learning is widely used in the fields of speech processing, image processing, fraud detection. It is also used in the field of medicine to predict ailments like diabetes, heart disease, and skin cancer. ML is classified into two classes, supervised and unsupervised learning. In supervised learning we have techniques like Random Forest which is a classifier and unsupervised learning uses techniques such as K-means etcetera.

The follow up sections of this paper discuss existing work by other scientists working in the field of machine learning and this work has provided a lot of input in the building of this model. It also discusses different machine learning techniques including Random Forests, feature selection, methodology, results, and recommendations.

II. EXISTING WORK

In literature various machine learning based diagnosis techniques have been proposed by researchers to diagnosis heart disease. This paper presents some existing machine learning based diagnosis techniques in order to explain the important

of the proposed work. Shashikant et al [3] compared the performance of three models used to predict cardiac arrest in smokers. The models were evaluated based on accuracy, precision, sensitivity, specificity, F1 score and Area under the curve (AUC). Logistic regression achieved an accuracy score of 88.50%, a precision of 83.11%, a sensitivity of 91.79%, a specificity of 86.03%, an F1 score of 0.87, and an AUC of 0.88. Decision tree model achieved an accuracy score of 92.59%, a precision of 97.29%, a sensitivity of 90.11%, a specificity of 97.38%, an F1 score of 0.93, and an AUC of 0.94. Random forest achieved an accuracy score of 93.61%, a precision of 94.59%, a sensitivity of 92.11%, a specificity of 95.03%, an F1 score of 0.93 and an AUC of 0.95. The random forest model achieved the best accuracy classification, followed by the decision tree, and logistic regression shows the lowest classification accuracy. Madhumita et al [4] developed a heart disease prediction model using Random Forest algorithm with a heart disease dataset of 303 records containing 14 features. The model was evaluated based on accuracy, sensitivity, and specificity. The model achieved 86.9%, 90.6% and 82.7% of accuracy, sensitivity, and specificity respectively. Younas et al [5] conducted a systematic review of 35 articles published in the use of machine learning techniques to detect heart disease and support vector machines (SVM), Neural Networks and ensemble classifiers came out as most popular techniques. Misra et al [6] implemented a model that predicts the probabilities of heart condition and classifies patient's risk level using naïve bayes, decision tree, logistic regression and random forest algorithms in which random forest algorithm achieved the highest accuracy of 90.16%. Ping LI et al [7] developed a system based on Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbor, Naïve bays, and Decision tree while standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features and proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. Shrestha et al [8] developed a web based system to predict heart disease using Naïve Bayes algorithms and achieved an accuracy score of 88.163%. Kishore et al [9] developed a heart disease prediction system using deep learning techniques, specifically Recurrent Neural Networks (RNN). Shu [10] of University of California compared the use of logistic regression, random forest, extreme gradient boosting and neural network to evaluate the best technique for the most robust model and Random Forest achieved the best accuracy score of 88.5%. Metsker et al [11] developed a prediction system based on Gaussian Naive Bayes, Gradient Boosting Classifier, Random Forest Classifier, K-Neighbors Classifier and Logistic Regression. Karthick et al [12] developed heart disease risk prediction model based on Support vector machine (SVM), Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and random forest algorithm and achieved an accuracy score of 80.32%, 78.68%, 80.32%, 77.04%, 73.77%, and 88.5%, respectively. Vishal et al [13] implemented a heart disease prediction model using Random Forest, Naive Bayes algorithm, Decision Tree, and Support Vector Machine and the model achieved an accuracy score of 79.47%. Improvement in model accuracy and efficiency remains an issue and consequently a research gap.

III. METHODOLOGY

We downloaded a heart disease dataset from Kaggle [14] which was in a comma delimited (CSV) format and stored on a working laptop computer on which anaconda version 2021.11 was installed. This data set is comprised of data from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 1024 tuples and 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field indicates to the presence of heart disease 0 representing no heart disease and 1 representing presence of heart disease. We performed pre-processing using python with Jupiter notebook 6.4.5 on the data set. A variety of pandas libraries were used in the model and included, Sklearn, seaborn, matplotlib as well as NumPy. The preprocessing also involved, verifying data types and convert mismatching types for the model to run smoothly. We also checked the data for any outliers in the dataset, **Fig.3**. Outliers prevent the model from learning the data well, so they must be removed. We also checked to see if the dataset had any features that are highly correlated so that they can be removed to improve the model accuracy using a correlation matrix along with Pearson correlation by creating a seaborn heatmap as shown in **Fig.1**.

We used all the 1024 samples, and 13 features form the dataset including 1 output label. During data preprocessing, we renamed the feature codes to represent a much meaningful code related with medical coding. The data was split into training and testing data and was processed using Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN) algorithms. Both Random Forest, Naïve Bayes, and K-Nearest Neighbors (KNN) algorithms are supervised learning techniques. In this project we predominantly used Random Forest algorithm as it is an ensemble classifier and can convert high variance which is usually present in Decision Trees to low variance.

The output label has two classes to describe the absence of heart disease and the presence of heart disease. The dataset matrix details are given in **Table.1**.

Table 1: The UCI heart disease dataset matrix

No.	Feature	Description
1	Age1	Age is continuous
2	Gender 1	1=male 0=female
3	Cp1	Chest pain
4	Trestbps	Resting blood pressure results during hospitalised: continuous(mmHg)
5	chol	cholesterol level in mg/dl
6	Fbs1	Fasting blood sugar 0:<=120mg/dl,1:>120mg/dl
7	restecg	electrocardiographic results during resting 1=true 0=false
8	thalach	Maximum heart rate achieved: continuous
9	exang	Exercise induced angina
10	oldpeak	ST depression
11	slope	ST segment slope
12	ca	Number of major vessels coloured by fluoroscopy: discrete (0,1,2,3)
13	thal	3: normal 6: fixed defect 7: reversible defect

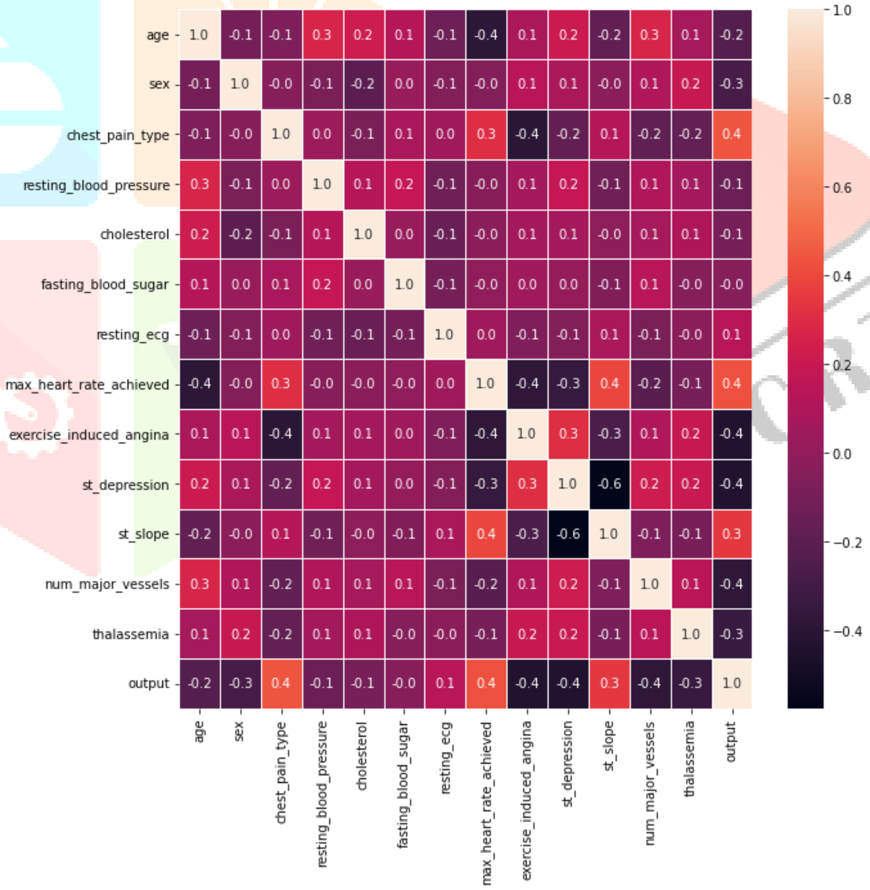

Figure 1: Correlation matrix of the dataset

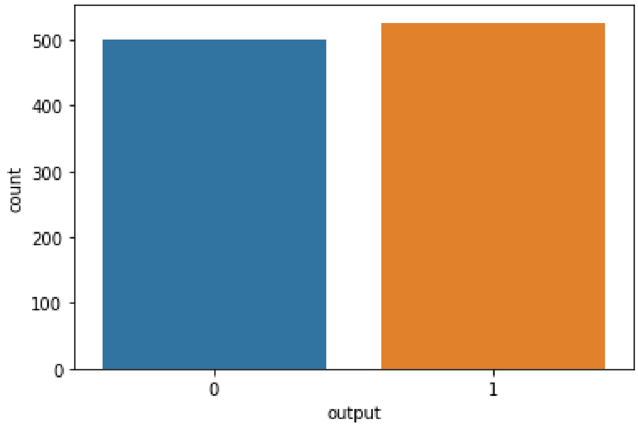

Figure 2: Data points balance check by KNN

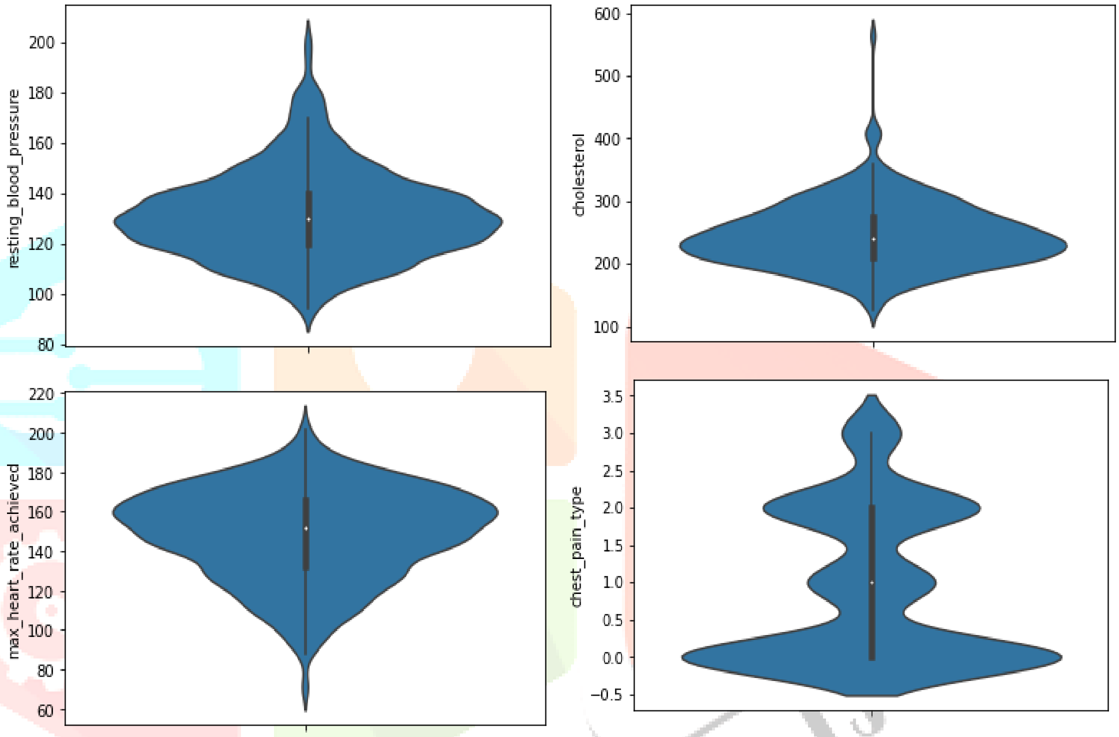

Figure 3: Violin plots to check for outliers for selected features

IV. THE MODEL, RESULTS AND DISCUSSION

This model performs heart disease prediction using Random Forest (RF), Naïve Bayes (NB), and K-Nearest Neighbors (KNN) algorithms. The model is run on the dataset comprising of 1024 tuples with 14 clinical features as shown in **Table.1**, used for determining clinical heart conditions of patients. The sample of the dataset with the 14 features is shown in **Table.2**.

The dataset was split into training and testing data frames with test size at 40% and random state of 42. The 40% test size value was chosen to give the model enough data to obtain better accuracy. The model also creates a confusion matrix to check for true negatives, true positives, false negatives, and false positives figure 3. The confusion matrix is used to compute important evaluation metrics such as Sensitivity, Specificity and Accuracy. The model correctly classified 48.29% true positives, 2.44% false positives, 43.17% true negatives, 6.1% false negatives using RF respectively. The model achieved an accuracy of 91.463%, an AUC of 98.3%, a sensitivity of 95.1%, a precision of 88.78%, and specificity of 87.62% using RF, which was highest among the three algorithms.

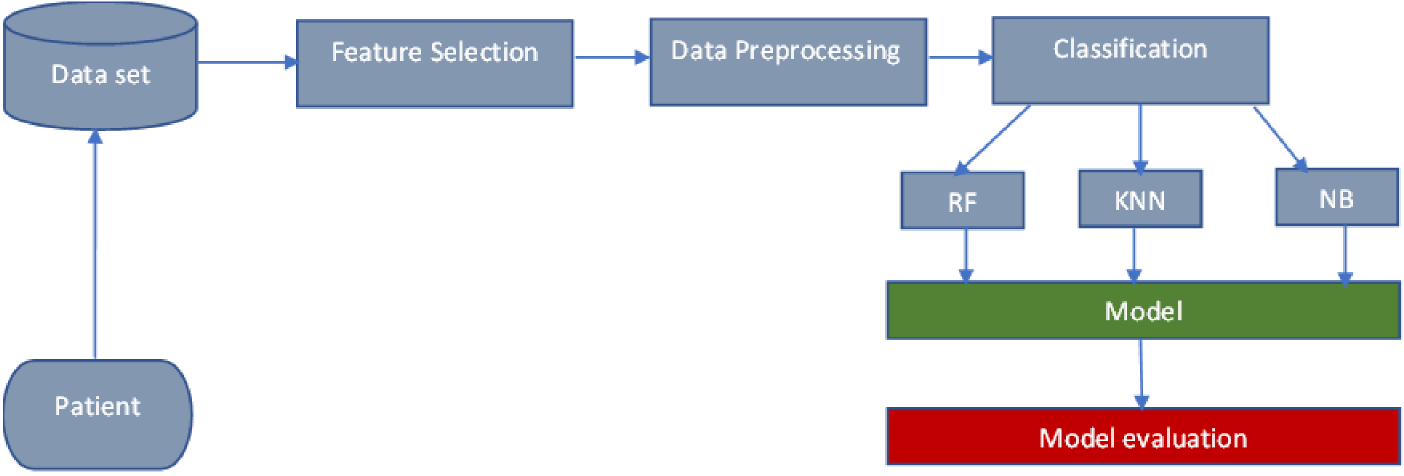

Figure 4: Framework of the model

Table 2: First 5 and last 5 samples of the dataset

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	resting_ecg	max_heart_rate_achieved	exercise_induced_angina	st_depression	st_slope	num_major_vessels	thalassemia	output
0	52	1	0	125	212	0	1	168	0	1	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	resting_ecg	max_heart_rate_achieved	exercise_induced_angina	st_depression	st_slope	num_major_vessels	thalassemia	output
1020	59	1	1	140	221	0	1	164	1	0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

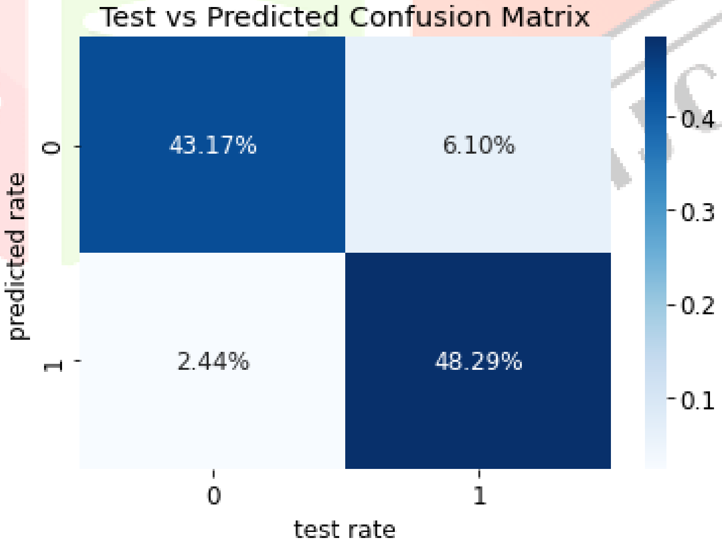

Figure 5: confusion matrix of the model using RF

The ROC curve **Fig.5**, as plotted, indicated that the model will 98.3% accurately predict a patient has a heart disease or not every time the model is executed.

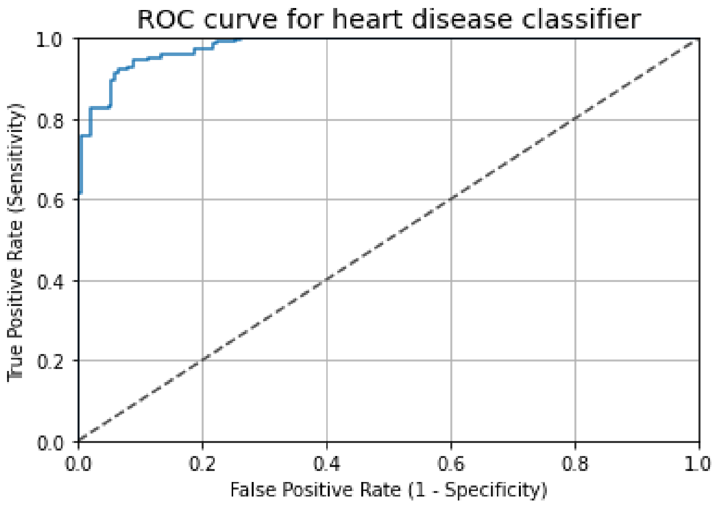

Figure 5: ROC curve for the heart disease prediction model

V. CONCLUSION

In this system random forest ensemble, Naïve Bayes, and K-Nearest Neighbors was used to implement a model for predicting heart disease conditions. Upon evaluation, RF achieved an accuracy of 91.463%, an AUC of 98.3%, a sensitivity of 95.1%, a precision of 88.78%, and specificity of 87.62%. The obtained AUC of 98.3% suggests the model will 98.3% accurately predict a patient has a heart disease or not every time the model is executed using RF compared to 45.9% by KNN and 89.8% by NB, **Table.3**.

The idea to build this model also takes into consideration that the model can be used for other algorithms because the framework is a hybrid ensemble that can take in several algorithms to come up with a strong model.

The model can also be used to build expert systems that just require a doctor to enter values and at a click of a button make important decisions to help patients timely.

Table 3: Model comparison for evaluation metrics

Model	Accuracy Score	Specificity Score	Sensitivity Score	Precision Score	AUC
Random Forest (RF)	91.71%	88.61%	94.71%	89.55%	98.19%
K-Nearest Neighbor (KNN)	85.86%	85.64%	86.06%	86.06%	45.88%
Naïve Bayes (NB)	81.50%	76.73%	86.06%	79.20%	89.83%

VI. ACKNOWLEDGMENT

The publication of this work owes to the unwavering support of Dr. Arul Valan, my supervisor and Dr. Glorindal Selvam, the postgraduate coordinator at DMI ST JOHN THE BAPTIST University, Malawi. Irrespective of their busy schedules, they provided guidance whenever I reached out.

REFERENCES

[1] "Cardiovascular diseases (CVDs)," 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Aug. 03, 2022).

[2] "Heart Disease Facts," 2022. https://www.cdc.gov/heartdisease/facts.htm (accessed Aug. 03, 2022).

[3] C. Engineering and E. Pune, "Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter," 2019, doi: 10.1016/j.aci.2019.06.002.

[4] A. Garg, B. Sharma, R. Khan, R. Morya, and S. Singh, "Prediction of Heart Diseases using Random Forest Prediction of Heart Diseases using Random Forest," 2021, doi: 10.1088/1742-6596/1817/1/012009.

[5] Y. Khan, U. Qamar, N. Yousaf, and A. Khan, "Machine Learning Techniques for Heart Disease Datasets : A Survey," pp. 27–35, 2019.

[6] R. Misra, P. Gupta, and P. Jain, "Prediction of Heart Disease Using Machine Learning Algorithms," vol. 8, no. 2, pp. 643–646, 2021.

[7] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, no. Ml, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[8] R. Shrestha and J. M. Chatterjee, "Heart Disease Prediction System Using Machine Learning," vol. 1, no. 2, 2020.

[9] A. Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir, "Heart Attack Prediction Using Deep Learning," pp. 4420–4423, 2018.

[10] P. Date, "UCLA UCLA Electronic Theses and Dissertations," 2020.

[11] O. Metsker, S. Sikorsky, A. Yakovlev, and S. Kovalchuk, "ScienceDirect ScienceDirect mortality prediction using machine

learning techniques for International Young Scientist Conference on Computational Science acute cardiovascular cases Dynamic mortality prediction using machine learning techniques for acute ca," *Procedia Comput. Sci.*, vol. 136, pp. 351–358, 2018, doi: 10.1016/j.procs.2018.08.279.

[12] K. Karthick, S. K. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, and A. R. Thelkar, "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," vol. 2022, 2022.

[13] I. No, "Available Online at www.ijarcs.info ANALYSIS OF MACHINE LEARNING ALGORITHM FOR PREDICTION OF," vol. 11, no. 3, pp. 42–46, 2020.

[14] "Heart Disease Dataset," 2022. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset (accessed Aug. 03, 2022).