International Journal of Scientific Research and Engineering Development-–Volume 8 Issue 5, Sep - Oct 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Development of Predictive Modeling of Heart Disease Outcomes Using XGBoost Machine Learning and Assessing the Impact of XGBoost on Heart Disease Prediction Accuracy

Dr. D. Janaki Sathya
Assistant Professor (Sl.G), Department of EEE, PSG College of Technology, Coimbatore, Tamil Nadu, India,
djs.eee@psgtech.ac.in

Abstract:

Now a day the increase in global death is due to heart disease or heart related issues. So, in order to reduce this global death, count early detection of heart disease is very much important. This work helps to mainly predict the possibility of heart disease. So, early detection of such disease helps to reduce the death rate. The dataset used here is Statlog dataset and it has been downloaded from Goggle. It is a publicly available dataset that has been used to build the model and XG-Boost has been used for prediction. The platform used for performing this work is Python Jupyter notebook which is an open-source web application that helps to perform data visualization, machine learning and much more. So, this work mainly focused on DBSCAN, SMOTEEN and XG-Boost. DBSCAN for detection of outliers, SMOTEEN for balancing of data and XG-Boost for prediction of heart disease.

Keyword: Heart disease, Machine Learning, Classification Techniques, Python Jupyter Notebook, DBSCAN, SMOTEEN, XG-Boost.

1. INTRODUCTION

Heart disease, as one of the most life-threatening diseases, has become one of the most frequent diseases that affect both middle and older individuals, but younger people are also at risk these days. Blood pressure, cholesterol levels, changes in ECG, blood sugar levels, heart rate, and other parameters are all important in monitoring this heart condition. Obesity, smoking, drinking, lack of physical activity, eating choices, and other factors are among the leading causes of heart disease. To address these difficulties, academics have discovered or are working to develop new technologies such as data mining, machine learning, and deep learning, among others. Heart disease is on the rise these days, owing to poor eating habits, lack of exercise, and other factors. The main cause of heart disease is a blockage of blood flow to the heart, which causes heart attacks, breathing problems, and shortness of breath, among other symptoms. This blockage is primarily caused by the growth of new substances inside the arteries, such as plague, which also obstructs blood flow throughout the body. As a result, it's critical to avoid such diseases and raise awareness about them. As a result, early identification of cardiac disease is critical. The heart is one of our body's most expensive and critical organs, so it requires

special attention. Most natural parameters are taken into account in this work, such as age, sex, blood sugar level, and accuracy is compared for several classification algorithms, such as Linear Regression, KNN, Navies Bayes, and XG-Boost. Based on the computed accuracy, a judgement is formed as to which algorithm is optimal for heart disease prediction. The following is how the paper is structured: The second half of the paper discusses related research on several categorization algorithms for predicting heart disease. Section 3 has a full description of the requested study, whereas Section 4 contains the results, and Section 5 contains the conclusion and future work.

2. RELATED WORKS

This section lists the works that are related to this one. In [1]a logistic regression model was employed to predict cardiac disease and its accuracy was tracked. In [2],K-nearest neighbour, decision tree, linear regression, and support vector machine (SVM) were used to predict heart disease, and the repository dataset was used for training and testing, with python jupyter notebook. Age, sex, obesity, blood pressure, cholesterol, and other variables have been utilised to accurately predict heart disease using Machine Learning methods such as Gaussian Nave Bayes, Random Forest, K-Nearest Neighbour, and Support Vector Machine in [3]. [4] employed the Decision Tree Algorithm and the Nave Bayes Algorithm to accurately forecast cardiac disease.

[5]employed the Random Forest method, which is one of the most powerful algorithms among many classification techniques, and the file was read in csv format. Nave Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor, and logistic regression are all examples of decision trees. This study paper [6] used regression methods for effective cardiac disease prediction, and the performance was evaluated using parameters such as accuracy, ROC curve, precision, and so on.

This suggested system makes use of the Cleveland heart disease dataset, as well as data mining techniques such as regression and classification. [7] employed three machine learning algorithms: random forest, decision tree, and hybrid model (a hybrid of random forest and decision tree). In [8] Decision Tree classification complements techniques based on Naive Bayes, Logistic Regression, Random Forest, SVM, and KNN. In [9], three strategies were employed to choose candidate feature subsets: mean Fisher score-based feature selection algorithm, forward feature selection algorithm, and reverse feature selection technique.

To identify cardiac illness, [10] uses a rough sets-based attribute reduction and a 23-interval type-2 fuzzy logic technique (IT2FLS). Six machine learning algorithms were used to assess heart disease prediction in [11]. KNN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network, and Vote are the seven methodologies used in this study report [12] to evaluate heart sickness (a hybrid technique combining Nave Bayes and Logistic Regression).

The Cleveland dataset [13] , which is a dataset from UCI heart disease, was used in this study, and the prediction was performed. The prediction model in [14] is built utilizing a variety of feature combinations and various well-known classification algorithms. In comparison to other

proposed models or systems, this proposed work [15] performs better.

The proposed technology can quickly distinguish between those with heart disease and those who are healthy. Receiver optimistic curves and area under the curves were also computed for each classifier. [16] employed all of the classifiers, as well as feature selection techniques, pre-processing methods, validation methods, and classifier performance evaluation measures, to make it easier to identify and classify persons with heart disease from healthy people. This study [17] uses a backpropagation neural network (RS-BPNN) and is divided into two parts. In the first stage, missing values are identified, the data set is smoothed, and feature selection is performed. Backpropagation neural networks are used in the second stage of classification. In [18], the Nave Bayes and KNN algorithms were used.

Three steps were completed in [19], including data gathering, user registration, and classification technique selection. [20] employed a hybrid model, which included Nave Bayes, ANN, SVM, and Hybrid Nave Bayes, SVM, and ANN, to predict cardiac disease. The comparison was based on classification techniques' accuracy, specificity, and sensitivity.

3. PROPOSED WORK

Researchers employed a variety of categorization algorithms to determine the best algorithm for predicting heart disease. The proposed classification system is illustrated as a block diagram in Figure 3.1.

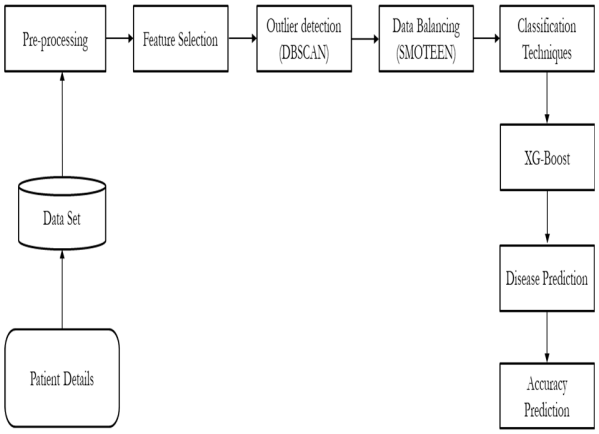

Figure 3.1 Block diagram of the proposed system

3.1 Dataset

The Statlog heart disease dataset from the UCI machine learning lab is used in this work (Figure 3.2). This collection contains 270 samples, 150 of which are free of heart disease or do not have heart disease, and 120 of which do. In order to predict the existence or absence of heart illness, a total of 13 features were examined. A list of 13 different parameters that were studied for prediction is shown in Figure 3.2.

No.	Symbol	Description
1	*age*	Subject age in years
2	*sex*	Subject gender
3	*cp*	Chest pain type
4	*trestbps*	Resting blood pressure in mmHg
5	*chol*	Serum cholesterol in mg/dl
6	*fbs*	Fasting blood sugar with value > 120 mg/dl
7	*restecg*	Resting electrocardiographic result
8	*thalach*	Maximum heart rate
9	*exang*	Exercise induced angine
10	*oldpeak*	ST depression induced by exercise relative to rest
11	*slope*	Slope of the peak exercise ST segment
12	*ca*	Number of major vessels (0–3) colored by flourosopy
13	*thal*	Defect type

Figure 3.2 Statlog Dataset

3.2 Data pre-processing

The first part of our procedure, which involves using a dataset obtained from UCI as input,

consists of 270 records. Data is frequently insufficient and inconsistent. Machine learning algorithms are substantially influenced by inconsistent and incomplete datasets. The missing values and null values in the heart disease data set were examined since null values have a significant impact on the conclusions obtained from the data. These inputs will be subjected to a pre-processing step. XG-Boost is the methods that are taken into consideration. Checking for nullity or missing terms in the dataset we've evaluated is one of the primary pre-processing strategies considered in the suggested system, and it's critical for improving the proposed system's accuracy.

To check for nullity data1.isnull().sum() command is used in python jupyter notebook and its shown in Figure 3.3. From this Figure 3.3 its clear that the data set considered has no missing terms and its ready to undergo the next step that is feature selection.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 270 entries, 0 to 269
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             270 non-null    int64
 1   sex             270 non-null    int64
 2   chest_pain      270 non-null    int64
 3   blood_press     270 non-null    int64
 4   serum_chol      270 non-null    int64
 5   blood_sugar     270 non-null    int64
 6   electrocard     270 non-null    int64
 7   max_heart_rate  270 non-null    int64
 8   induced_ang     270 non-null    int64
 9   oldpeak         270 non-null    float64
 10  peak_st_seg     270 non-null    int64
 11  major_ves       270 non-null    int64
 12  thal            270 non-null    int64
 13  presence        270 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 29.7 KB
None
```

Figure 3.3 Pre-processing

3.3 Feature Selection

The main core notion of machine learning that aids in the improvement of a system's performance is feature selection. The main benefit of feature selection is that it improves accuracy and helps the system or classifier work better. It also aids in the reduction of over fitting and training time. In this work, a heatmap was created as shown in Figure 3.4, and histogram representations for various parameters in the selected dataset were created as shown in Figure. 3.5.

The connection between attributes may have an impact on the machine learning model's performance. To calculate data correlation and evaluate the relationship between attributes, utilise Pearson's Correlation Coefficient (PCC). A heatmap is a graphical representation of data that uses colour coding to indicate different values. Heatmaps can be used in a variety of analytics platforms, although they're most commonly used to show user behaviour on certain webpages or designs.

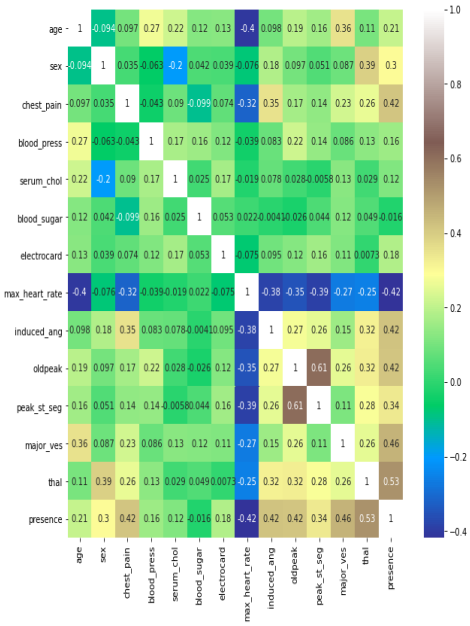

Figure 3.4 Heatmap

PCC ranges from 1 to -1, with a positive or negative number suggesting a strong positive or

strong negative correlation between the variables, and a value near to zero indicating a weak correlation. Figures 3.4 show the heatmap correlation between attributes for dataset considered. The green colour denotes a correlation near to 0, but the white and blue colours denote a correlation close to 1 and -1, respectively.

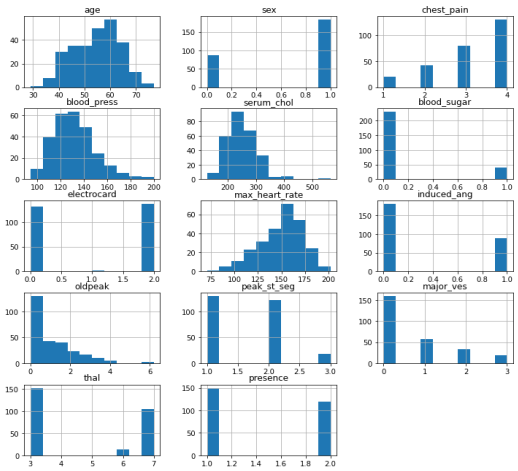

Figure 3.5Histogram

There are number of techniques for customising histograms. Hist (figsize=(12,12),layout=(5,3)) function has numerous attributes that can be used to modify a histogram. In Figure3.5 histogram representations of each parameter is shown which helps to classify each parameter into continues or categorical event.

3.4 DBSCAN

Density Based Spatial Clustering of Applications with Noise is what DBSCAN stands for. This is a clustering algorithm that primarily aids in the separation of high- and low-density clusters.

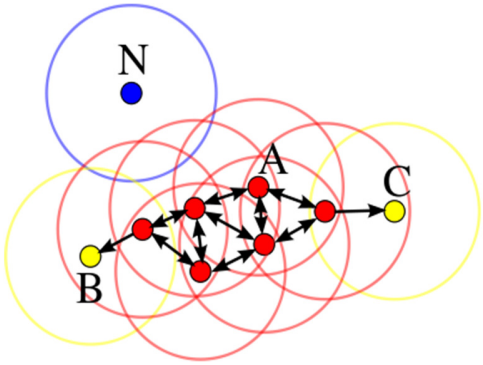

Figure 3.6 DBSCAN

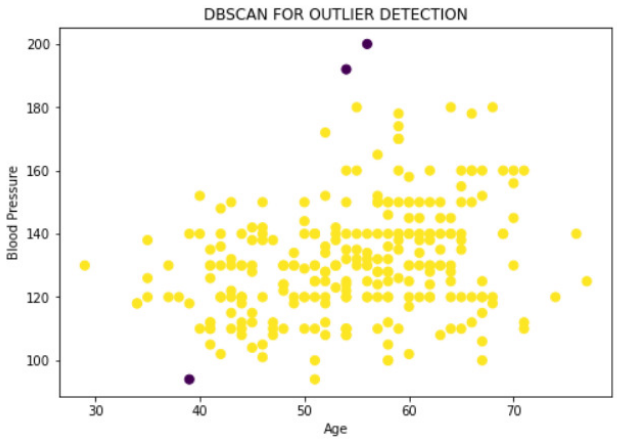

Figure 3.7 Outlier detection

Two parameters must be given when using DBSCAN: minpts (minimum point) and eps (equivalent to radius). DBSCAN considers three different types of points: core points, border points, and outliers. In Figure 3.6, these points are denoted by the letters A, B, and N, respectively. Three outliers were discovered in this study, as indicated by the black colour in Figure 3.7, and the minpts and eps values used to obtain these outliers were 5 and 8, respectively.

3.5 SMOTEEN

SMOTEEN is a hybrid strategy that aids in the removal of additional data points in order to improve accuracy. It's simply a mix of SMOTE and ENN methods. SMOTE means for Synthetic Minority Over-sampling Technique, which is an

oversampling technique, and ENN stands for Edited Nearest Neighbor, which is an under-sampling technique that primarily uses KNN to locate the nearest neighbour. Prior to the use of SMOTEEN, the absence and presence of disease in this study were 134 and 109, respectively. The data was balanced after experiencing SMOTEEN methods.

3.6 Classification Techniques

There are many Classification models in machine learning which acts as a predictive model. Feature extraction is done based on the model considered. A classifier or a classification model is used for predicting categorical classes. As discussed earlier there are many machine learning classification techniques available among them in this work XG-Boost has been considered.Accuracy is calculated for the classification technique considered.

3.6.1 XG-Boost

XGBoost is an acronym for extreme gradient boosting. A gradient boosting framework is used by the XG-Boost classification algorithm. In contrast to regular gradients, optimal gradients help to reduce error for each iteration. It has a low computational cost, a rapid running speed, and good accuracy. The boosting is intended at transforming a base classifier into a classifier in order to attain good accuracy, as it is an ineffective ensemble learning approach. Furthermore, gradient boosting aims to improve resilience by lowering the algorithm's loss function along its gradient direction during iteration. Furthermore, as an implementation of the gradient boosting algorithm, XGBoost can take full advantage of multi-core CPUs for parallel processing and improve accuracy, significantly reducing computational loads and improving accuracy when compared to other widely used algorithms like decision trees and random forests.

Ensemble learning is a method for systematically combining the prediction skills of several learners. The end result is a single model that combines the outputs of numerous models. The foundation learners, or models that make up the ensemble, could be from the same learning algorithm or from distinct learning algorithms. Bagging and boosting are two types of ensemble learners that are commonly utilized. Though these two strategies can be applied to a variety of statistical models, decision trees have been the most popular.

XGBoost is a sort of supervised machine learning that is used for regression modelling and classification. XGBoost is a more advanced technique that uses gradient boosting DTs with numerous modifications in terms of regularization, loss function, and column sampling. Gradient boosting is a prediction approach in which new models are built and used to predict the error or residuals, and then the scores are summed to produce the final prediction result. The gradient descent approach is used to lower the loss score when constructing new models. To evaluate the model's performance, the objective function, which is made up of two parts: training loss and regularization, must be used. The regularizations term penalizes the model's complexity, preventing overfitting. The objective function is depicted using the equation :

$$L(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_i \Omega(f_k) \qquad (1)$$

The difference between the prediction $\hat{y}_i$ and the target y_i is calculated using the differentiable convex loss function l. The regularized term Ω penalizes the model's complexity, while T is used to denote the number of leaves in the tree.

The XG - Boost models have the highest accuracy when it comes to predicting the presence or absence of heart disease. XG - Boost is a supervised machine learning technique for classification and regression. We chose XG-Boost over other classification algorithms because it offers the best model performance and has a high-speed prediction property. To increase accuracy, this study used the outlier identification tool DBSCAN and data balancing using SMOTEEN before conducting XG – Boost. DBSCAN revealed three outliers, or unwanted locations, as seen in Figure 3.6.The data was then balanced using SMOTEEN, and the process was completed.

Finally, this balanced data set is classified using the XG – Boost classifier. And it was determined that the accuracy was around 87.8%.

4. RESULTS AND ANALYSIS

h. Accuracy

Work could be concluded by performing or implementing confusion matrix and accuracy comparison of four different classification techniques to obtain the best technique to detect or predict heart disease.

4.1 Confusion matrix

The model's four potential outputs were measured using a confusion matrix (see Figure 4.1): true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The simulation has been performed using Python Jupyter. Simulation has been performed for each different classifier. Initially simulation is performed for pre- processing step, heat map extraction, histogram generator etc. before implementing classifier. The implementation steps carried out in this work are:

a. Dataset updation
b. Pre-processing
c. Feature selection
    i. Heatmap
    ii. Histogram representation
d. Splitting of dataset
e. Outlier Detection using DBSCAN
f. Data Balancing using SMOTEEN
g. Classifier implementation

i. XG- Boost

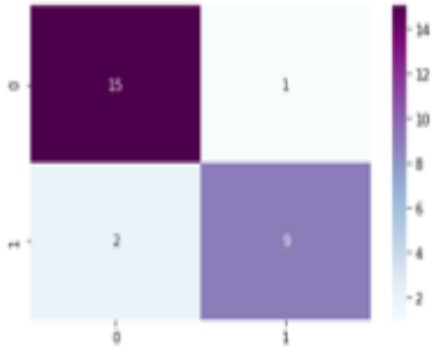

Figure 4.1 Confusion Matrixes of XG - Boost

The number of subjects correctly classified as "positive" (presence of heart disease) and "negative" (healthy/absence of heart disease) is defined as TP and TN outputs, respectively, and the number of subjects incorrectly classified as "positive" (presence of heart disease) when they are actually "negative" (healthy/absence of heart disease) is defined as FP and FN outputs.

4.2 Accuracy

Accuracy of the algorithms are depending on four values namely true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

$$\text{Accuracy} = \frac{(\text{FN}+\text{TP})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})} \qquad (2)$$

The numerical value of TP, FP, TN, FN defines as:

TP= Number of persons with heart diseases
TN= Number of persons with heart diseases and no heart diseases
FP= Number of persons with no heart diseases
FN= Number of persons with no heart diseases and with heart diseases

After performing the machine learning approach for testing and training it is found that accuracy of the XG - Boost is much efficient. Accuracy should be calculated with the support of confusion matrix of each algorithm as shown in Figure 4.1 and the number of counts of TP, TN, FP, FN are given and using the equation of accuracy given by equation 2, value has been calculated and it is concluded that XG - Boost gives a accuracy of about 88.88% for the dataset considered

5. CONCLUSION AND FUTURE WORK

Many ways to predict heart disease using multiple machine learning algorithms has been learned from our research. A dataset for cardiac disease was obtained from the UCI Machine Learning Repository, which comprised its characteristics, and further classification models were applied to the Statlog dataset. To anticipate cardiac illness in a short amount of time, retrieve the results, and lower people's expenditure, our proposed methodology uses XG-Boost machine learning models. In the future, alternative machine learning classification algorithms, in addition to existing classification approach like XG - Boost algorithms, may be employed to predict or diagnose different diseases. The project can be improved or expanded in the future.

REFERENCES

[1]M. Saw, T. Saxena, S. Kaithwas, R. Yadav, and N. Lal, “Estimation of prediction for getting heart disease using logistic regression model of machine learning,” 2020. doi: 10.1109/ICCCI48352.2020.9104210.

[2]A. Singh and R. Kumar, “Heart Disease Prediction Using Machine Learning Algorithms,”2020.doi:10.1109/ICE348803.2020. 9122958.

[3]S. Farzana and D. Veeraiah, “Dynamic heart disease prediction using multi-machine learning techniques,” 2020. doi: 10.1109/ICCCS49678.2020.9277165.

[4]N. Rajesh, T. Maneesha, S. Hafeez, and H. Krishna, “Prediction of heart disease using machine learning algorithms,” International Journal of Engineering and Technology (UAE), vol. 7, no. 2.32 Special Issue 32,2018,doi:10.47059/alinteri/v36i1/ajas21 039.

[5]M. S. Raja, M. Anurag, C. P. Reddy, and N. R. Sirisala, “Machine Learning Based Heart Disease Prediction System,” 2021. doi: 10.1109/ICCCI50826.2021.9402653.

[6]P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," 2020. doi: 10.1109/INOCON50539.2020.9298354.

[7]M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model,"2021.doi:10.1109/ICICT50816.2021.9358597.

[8]R. Jane Preetha Princy, S. Parthasarathy, P. Subha Hency Jose, A. R. Lakshminarayanan, and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," 2020.doi: 0.1109/ICICCS48265.2020.9121169.

[9]S. M. Saqlain et al., "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," Knowledge and Information Systems, vol. 58, no. 1, 2019, doi: 10.1007/s10115-018-1185-y.

[10]N. C. Long, P. Meesad, and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," Expert Systems with Applications, vol. 42, no. 21, 2015, doi: 10.1016/j.eswa.2015.06.024.

[11]A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Computing and Applications, vol. 29, no. 10, 2018, doi: 10.1007/s00521-016-2604-1.

[12]M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol. 36, 2019, doi: 10.1016/j.tele.2018.11.007.

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

[14]S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2923707.

[15]L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," IEEE Access, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2909969.

[16] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Information Systems, vol. 2018, 2018, doi: 10.1155/2018/3860146.

[17]K. B. Nahato, K. N. Harichandran, and K. Arputharaj, "Knowledge mining from clinical datasets using rough sets and backpropagation neural network," Computational and Mathematical Methods in Medicine, vol. 2015, 2015, doi: 10.1155/2015/460189.

[18]S. Ambekar and R. Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network,"2018.doi:10.1109/ICCUBEA.2018.8697423.

[19]A. N. Repaka, S. D. Ravikanti, and R. G.

Franklin, "Design and implementing heart disease prediction using naives Bayesian," in Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019,2019,vol.2019-April.doi:10.109/icoei.2019.8862604.

[20]M. J. A. Junaid and R. Kumar, "Data Science and Its Application in Heart Disease Prediction,"2020.doi:10.1109/ICIEM48762.2020.9160056.

