e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

DEVELOPING A HYPER PARAMETER TUNING AND OUTLIERS BASED ON MACHINE LEARNING APPROACH OF HEART DISEASE PREDICTION

Tejinder Kaur*1, Er. Harjasdeep Singh*2, Er. Rishamjot Kaur Dhaliwal*3

*1Baba Farid College Of Engineering & Technology, Bathinda, Punjab, India.

*2Assistant Professor, Computer Science & Engineering, MIMIT Malout, Punjab, India.

*3Assistant Professor, Computer Science & Engineering, Baba Farid College Of Engineering & Technology, Bathinda, Punjab, India.

DOI : https://www.doi.org/10.56726/IRJMETS51820

ABSTRACT

The body's principal organ is the heart, providing blood to every bodily component. Global death rates are rapidly increasing due to heart problems. In this research work, various ML algorithms such as RF, SVM, MLP, and NB are used for the forecast of HD. These four algorithms are evaluated by using four metrics of performance. The SVM model achieves the best accuracy of 92% from the other algorithms.

Keywords: Machine Learning, Support Vector Machine, Multilayer Perceptron, Random Forest, Naïve Bayes.

I. INTRODUCTION

The heart, a vital organ responsible for pumping and filtering blood throughout the body, is centrally located in the chest between the lungs. A heart is formed up of two right and left atria and two ventricles. The blood passes from the atria through the body and into the ventricles of the heart. Heart disease, as identified by the World Health Organization (WHO) [4], is a leading cause of death globally, with fatalities possible within minutes if the heart malfunctions. Cardiovascular disease (CVD) ranks highest among fatal conditions, with heart disease causing significant damage. WHO reports indicate that 10 million deaths were attributed to heart disease, highlighting the substantial real-world impact despite the vast, yet uncertain, nature of medical records? Effective disease forecasting remains a challenge, especially considering limitations in early patient access to operators. Coronary disease, particularly, exerts a significant global burden, contributing to rising mortality rates as indicated by investigations in January 2017. Cardiovascular malady (CVM) emerges as the foremost global killer, underscored by its consistent ranking among the top ten causes of death over the past fifteen years. Timely diagnosis holds potential for saving lives, mitigating the challenges associated with CVD procedures. Researchers leverage artificial intelligence software to analyze cardiovascular disease trends [5], revealing a staggering death toll attributable to heart disease. WHO data from 2016 further underscores the prevalence of heart disease among the top ten causes of mortality, with ischemia (coronary heart disease) contributing to a significant portion of fatalities, highlighting the sustained impact of these illnesses over the past fifteen years. One of the most vulnerable cities in the whole nation to HD is Bangladesh. The World Health Organization (WHO) reports that, in the nation of Bangladesh within 2016, non-communicable conditions (NCDs) represented 67% of all deaths, with cardiovascular diseases (CVDs) causing 30% of predicted deaths in the NCD category in Bangladesh's cities. However, the last study was conducted in South Asian nations; Bangladesh is the most speed-able city in terms of CVD [7]. It is said that Bangladesh is the city that is missing in action that has been taken in opposition to CVD. An increasing trend for CVD due to death and sickness has become one of the essential courses of clinical data analysis. The data produced by HD is humongous in the field of healthcare. ML is a compelling way with the vast majority of data produced by the medical services industry. It can be used for choices as well as predictions with a satisfactory level of accuracy [6]. This paper's main goal is to highlight various data processing techniques used in heart condition prediction. The principal component of the human body is the heart. It controls the body's overall blood flow. Any type of disruption within the circulatory system's supply of blood could lead to suffering in various other parts of the body. The classification of heart condition has often been made in the case of disturbance in the normal functioning of the guts. These days, heart problems are among the leading causes of mortality. A poor lifestyle can lead to heart disease. Hypertension could be caused by consuming smoking, alcohol, and an excess amount of fat. WHO reports that

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

every year 10 million patients die due to HD.? The only two ways can tackle this problem are a balanced lifestyle and early detection [7].

II. LITERATURE SURVEY

- Chaimaa Boukhatem et al. [2022] utilized many learning techniques for e.g. multilayer perceptron, naïve bayes, support vector machine, random forest to predict heart disease using a dataset from Kaggle. Performance was evaluated based on criteria including F1-score, recall, accuracy, and precision, with SVM achieving the best accuracy of 91.67%. Future research aims to apply additional techniques and in-depth data analytics to improve accuracy.
- Jaspreet Singh et al. [2022] enhanced the predictive system's performance by using the NFS (Novel Feature Selection) method to select important features. Various techniques such as MLP, Bagging, RF, Random committee, Kstar, Bayes Net, and SMO were analyzed, with Bayes Net demonstrating the highest accuracy. The study suggests employing Egar and lazy ML algorithms to further improve accuracy.
- Abdullah Alqahtani et al. [2022] employed ensemble techniques for predicting CVD, combining DL and ML approaches. RF algorithm was used to extract essential features of CVD, and an ML-based ensemble model achieved the best accuracy of 88.70%. The study recommends exploring RL techniques and using additional datasets to enhance CVD prediction efficiency.
- M. Snehith Raja et al. [2021] developed an HD prediction system using the RF technique, processing patient data from CSV files in Python. The system demonstrated high accuracy and flexibility, with additional features including informing family members and doctors about the patient's condition and facilitating online meetings with doctors.
- Rubini PE et al. [2021] compared machine learning methods for cardiovascular disease categorization, including NB, LR, SVM, and RF. RF showed the highest accuracy and reliability, highlighting its effectiveness in predicting cardiac disease and its correlation with diabetes. The study suggests exploring new parameters and algorithms to further enhance system accuracy.
- Mohammed Nowshad Ruhani Chowdhury et al. [2021] introduced a predictive model for heart disease (HD) utilizing a dataset collected from Sylhet city, Bangladesh, through physical visits to health centers and hospitals. Various machine learning methods, including KNN, decision tree, support vector machine (SVM), logistic regression, and naïve bayes, were employed for HD prediction. SVM achieved the highest accuracy of 91%, prompting the authors to consider using artificial neural network (ANN) for HD forecasting due to its superior accuracy. Additionally, supervised learning regression analysis was proposed to calculate the sensitivity rank of individuals regarding HD as a percentage, providing more nuanced insights than binary categorization.
- Rachit Misra et al. [2021] employed several data processing techniques such as logistic regression, random forest, decision tress, and naïve bayes to estimate the probability of cardiac problems and categorize sick person threat. Through comparative analysis, random forest demonstrated the highest accuracy of 90.16% among various machine learning algorithms. The authors suggest building an online application utilizing random forest technique and utilizing larger datasets to enhance their predictive model, aiming to assist medical professionals in accurately predicting cardiovascular diseases more efficiently.
- Archana Singh et al. [2020] utilized techniques of learning likely linear regression, decision tree, support vector machine, KNN, to predict cardiac illness using data from the UCI repository. Implementation was done using Anaconda (Jupyter notebook) for Python, ensuring precision and accuracy. KNN exhibited the highest accuracy at 87% based on the confusion matrix for heart disease prediction. The authors plan to explore various ML techniques further to achieve the most comprehensive analysis for heart disease.
- Khalid Amen et al. [2020] explored various models of potential supervision trained through ML techniques and achieved one of the best accuracies. Utilizing hyper features from the UCI dataset (accuracy, precision, recall, and F measure), they investigated ML approaches including Logistic regression, Random forest, Support vector machine, Gradient tree boosting, Extra RF. Logistic regression yielded the maximum accuracy of 82% compared to other techniques. Moreover, they suggest potential improvements to enhance the system's performance.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

- Due to time constraints, they recommend focusing future research on: Training larger datasets. Comparing alternative ML techniques. Developing an automated HD detection system. Comparing various techniques for the best accuracy using real data from healthcare centers. Utilizing deep learning (DL) techniques such as Convolutional Neural Networks (CNN) or Artificial Neural Networks (ANN) to design intelligent decision-making layers.

- Tsehay Admasssu Assegie et al. [2019] presented a prediction model for HD classification using the SVM algorithm, assessing metrics like accuracy, recall, precision, and confusion matrix on a Kaggle dataset. The study achieved a moderate accuracy of 72.41% using the support vector machine algorithm.

III. METHODOLOGY

COLLECTION OF DATA: Kaggle is the source of the dataset utilized in this study.

PREPROCESSING DATA: A crucial component of machine learning is preprocessing data. Data processing is used to take out corrupted or missing data points and outliers. Besides, data transformation, resampling, and applying feature selection are also done by it.

Data Visualization and cleaning: Nothing is found after checking for missing values. After that the outliers also investigate. To exclude extreme outliers, the moderate outliers help with the final diagnosis. Using (1) and (2), where IQR stands for interquartile range, severe outliers have been detected. With Q1 and Q3 representing the bottom and higher quartiles, it provides a measurement for data dispersion. IQR= Quartile3- Quartile1

Q1 - 1.5*IQR (Interquartile range) (1)

Q3 + 1.5*IQR (Interquartile range) (2)

Data Splitting: The data set has been separated into two sections: testing and training data for machine learning. The model is trained in the training part and the testing part is used to test the trained data as well as forecast output. Collecting 80% of the data for training and 20% for testing is the goal of the hold-out method.

Feature selection: It is the most important part of machine learning that extremely affects the performance of the system. The feature is the characteristic that can affect the issue helpful for the issue as well and selecting the essential features for the system is called FS (feature selection). The process of ML is based on the FE (feature engineering) that holds the two tasks first one is FS (feature selection), second one is FE (feature extraction). FS defines the subset that is selected from the original feature set, and FE generates the new attributes. FS decreases the IV (input variables) for the system by applying individual significant data to decrease the over fitting in the system.

IV. APPLIED ALGORITHMS

Random Forest (RF): It is an additional SL (Supervised Learning) approach that may be applied for binary, regression. It uses the EL method which unites various algorithms to make accurate forecasts in complicated situations. It builds the forecast based on outcomes of various DT (Decision Trees) over bootstrap aggregation or bagging. Figure 1 shows the process of working of RF.

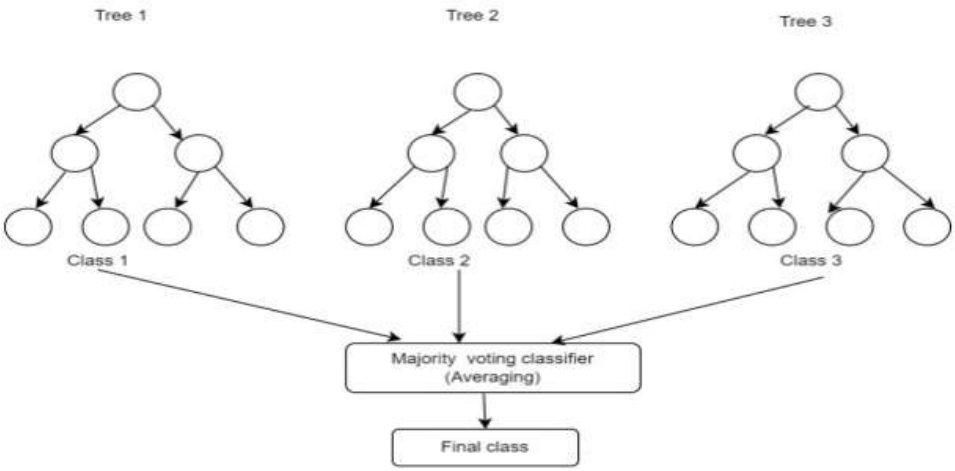

FIGURE 1: Random forest

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

SVM: This type of supervised learning is applied to classification, regression analysis, as well as identifying outliers. More than 2 featured vectors are used for the labeling of prediction which tries to build decision boundaries between different classes. This decision boundary also called hyper plane which is designed away from the near data points. These nearest points are called support vectors. Figure 2 represents the working of SVM.

$$\frac{||W||}{2} + C\sum_{i=1}^{n} zeta$$

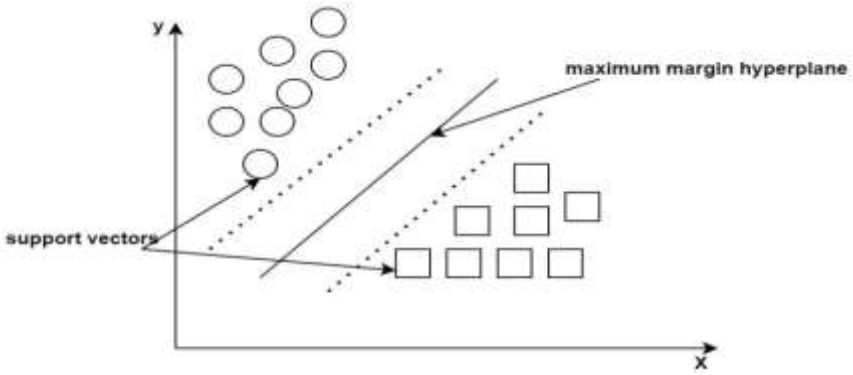

FIGURE 2: Support vector machine

MLP (Multilayer Perceptron): It features one or more hidden layers with many neurons placed on top of input and output layers. It is used for a supervised learning format. MLP also uses a back propagation algorithm. Back propagation is used for training the network. In the input layers, hidden layers and output layers are connected. The following figure shows the process of MLP.

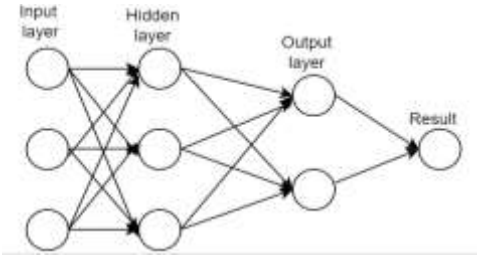

FIGURE 3: Multi-layer perceptron

Naïve Bayes: This kind of supervised learning method relies on the Bayes Theorem (BT). It takes into account that every attribute is independent, similar to the object of destination. The Bayes Theorem (BT) establishes the likelihood that an action A will occur.

Probability (A/B) = probability (B/A) * probability (A) / probability (B)

Evaluation Metrics: These are used to assess the learning algorithm's efficiency and standard. An N*N matrix, in which N is the number of is the number of classes being predicted, is a confusion matrix. Similar to a prediction problem with two alternative outcomes, the confusion matrix in this study is (2*2). The constituents of the matrix are the calculations of the right or wrong prediction. Figure 4 represents the process of the confusion matrix.

		Predicted
Acutal	TP (True positive)	FN (False negative)
	FP (False positive)	TN (True negative)

FIGURE 4: confusion matrix

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

Accuracy: To analyze the classification models, the accuracy metric is used. The following method is used to find out the accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP=FN}$$

Precision: The total number of FP (False Positive) and TP (True Positive) added together yields the number of TP.

$$\text{Precision} = \frac{True\ Positive}{TP+FP}$$

Recall: TPR is yet another term for recall. It computes how well the model can identify positive samples.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive+False\ Negative}$$

F1 score: It is also an ML calculation metric to determine the accuracy of the model. The F1 score merges the recall and precision outcome of an ML (Machine Learning) model.

$$\text{F1 score} = \frac{2TP}{2TP+FP+FN}$$

The dataset used in the research work was collected from the Kaggle. This dataset contains the seventy-six attributes, also together with the predicted attribute. The fourteen attributes are used for the published experiments. In this dataset, the target attribute indicates the heart disease of the patient. The '0' value represents the patient does not have heart disease and the '1' value represents that patient has heart disease.

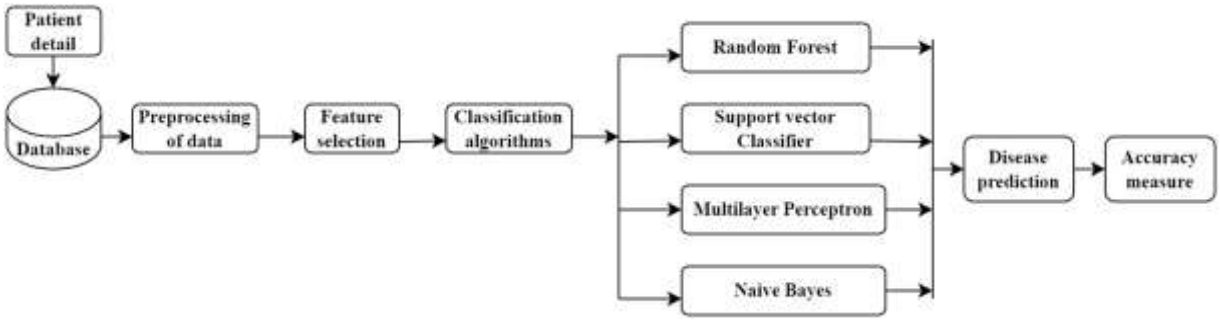

FIGURE 5: without removing outliers

The above Fig 5 represents the working process without removing the outliers.

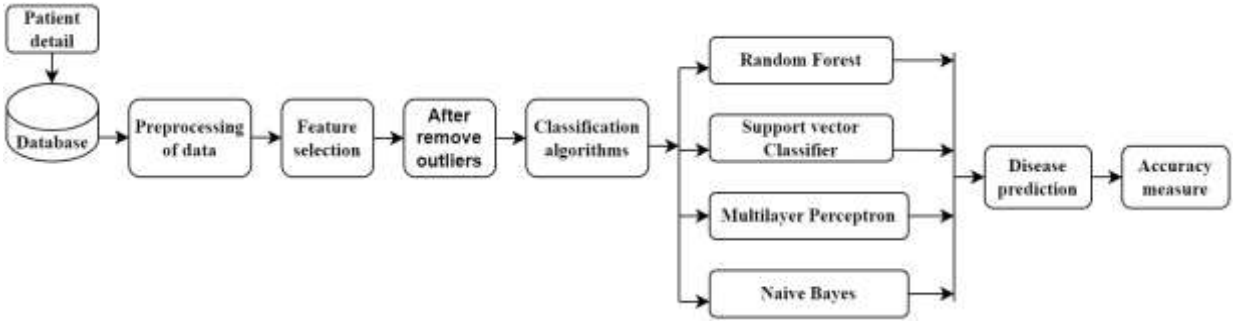

FIGURE 6: with removing outliers

Fig 6 shows the flow diagram of the dataset after removing the outliers.

V. RESULTS AND DISCUSSION

Model / Metric	RF	SVC	MLP	NB
Accuracy	84.61%	83.51%	83.51%	80%
F1 score	86%	85.14%	86.48%	81.25%
Recall	89.58%	89.58%	99.9%	81.25%
Precision	82.69%	81.11%	76.19%	81.25%

FIGURE 7: before removing outliers

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

In the above Fig 7, the results before removing the outliers are described. In this table, four classification learning approaches such as MLP, SVC, NB (Naïve Bayes) and RF (Random Forest are used. The accuracy of the RF algorithm is (84.61%), F1 score (86%), recall (89.58%), and precision (82.69%). The SVC algorithm has (83.51%) accuracy, (85.14%) F1 score, (89.58%) recall, and (81.11%) precision. The accuracy of MLP is (83.51%), F1 score (86.48%), recall (99.9%), and precision (81.25%) of MLP algorithm. The accuracy (80%), F1 score (81.25%), recall (81.25%), and precision (81.25%) of the NB algorithm. The following figures show the Confusion matrixes before removing outliers: The below figures show the accuracy performance of ML algorithms that are used in this research work. The fig 8 represents the accuracy performance of RF. The accuracy performance of SVC is given in the fig 9. Fig 10 indicates the accuracy performance of MLP. Fig 11 shows the accuracy performance of NB.

34	9
5	43

FIGURE 8: Random Forest

33	10
5	43

FIGURE 9: Support vector classifier

28	15
0	48

FIGURE 10: Multilayer Perceptron

34	9
9	39

FIGURE 11: Naïve Bayes

Model / Metric	RF	SVC	MLP	NB
Accuracy	81.6%	87%	82.75%	79.31%
F1 score	81.8%	84.78%	83.51%	79.06%
Recall	85.7%	92.85%	90.47%	81.25%
Precision	78.26%	78%	77.55%	72.27%

FIGURE 12: After removing outliers

In the above Figure 12, the results after removing the outliers are described. In this table, four classification learning algorithms like Naïve Bayes, Multilayer perceptron, Random Forest, and Support vector classifier are used. The accuracy of RF is (81.6%), F1 score (81.8%), recall (85.7%), and precision (78.26%). The SVC algorithm has accuracy (87%), F1 score (84.78%), recall (92.85%), and precision (78%). The accuracy of the MLP algorithm is (82.75%), F1 score (83.51%), recall (90.47%), and precision (77.55%). The accuracy (79.31%), F1 score (79.06%), recall (80.95%), and precision (72.27%) of the NB algorithm. The following figures indicate the Confusion matrixes after removing outliers: The below figures show the accuracy performance of ML algorithms that are used in this research work. RF's accuracy performance is illustrated in Fig. 13. The accuracy performance of SVC is displayed in Fig. 14. Fig. 15 indicates MLP's efficiency performance. The accuracy performance of NB is displayed in Fig. 16.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

35	10
6	36

FIGURE 13: Random Forest

34	11
3	39

FIGURE 14: Support vector machine

34	11
4	38

FIGURE 15: Multi layer Perceptron

35	10
8	34

FIGURE 16: Naïve bayes

Further, we experimented with a support vector algorithm to increase the accuracy of the SVM algorithm with a 'gamma' value of 0.1; kernel 'risk' gives 92% accuracy. The following Fig 17 shows the accuracy of the SVM algorithm.

Model / Metric	SVC
Accuracy	92%

FIGURE 17: Accuracy of SVC

VI. CONCLUSION

The purpose of this research work is to make possible predictions of heart disease among patients on behalf of special health measurements. Four types of classified mechanisms are used to make the prediction model. Collection of data as well as data cleaning is completed from any missing values and extreme outliers. The SVM algorithm shows more accuracy as compared to other algorithms after removing the outliers. Further, we used hyper parameter tuning to enhance the accuracy of the SVM algorithm. Applying additional machine learning algorithms and Deep Learning algorithms to achieve maximum accuracy will improve research efforts.

VII. REFERENCES

[1] Boukhatem, Chaimaa, Heba Yahia Youssef, and Ali Bou Nassif. "Heart disease prediction using machine learning." In 2022 Advances in Science and Engineering Technology International Conferences (ASET), pp. 1-6. IEEE, 2022.

[2] Singh, Jaspreet, Shruti Agarwal, Piyush Kumar, Divyansh Rana, and Rohit Bajaj. "Prominent features based chronic kidney disease prediction model using machine learning." In 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1193-1198. IEEE, 2022.

[3] Alqahtani, Abdullah, Shtwai Alsubai, Mohemmed Sha, Lucia Vilcekova, and Talha Javed. "Cardiovascular disease detection using ensemble learning." Computational Intelligence and Neuroscience 2022 (2022).

[4] Raja, M. Snehith, M. Anurag, Ch Prachetan Reddy, and Nageswara Rao Sirisala. "Machine learning based heart disease prediction system." In 2021 international conference on computer communication and informatics (ICCCI), pp. 1-5. IEEE, 2021.

[5] Rubini, P. E., C. A. Subasini, A. Vanitha Katharine, V. Kumaresan, S. Gowdham Kumar, and T. M. Nithya. "A cardiovascular disease prediction using machine learning algorithms." Annals of the Romanian Society for Cell Biology (2021): 904-912.

[6] Chowdhury, Mohammed Nowshad Ruhani, Ezaz Ahmed, Md Abu Dayan Siddik, and Akhlak Uz Zaman. "Heart disease prognosis using machine learning classification techniques." In 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1-6. IEEE, 2021.

[7] Misra, Rachit, Pulkit Gupta, and Prashuk Jain. "Prediction of Heart Disease Using Machine Learning Algorithms." IJIRT 152152 (2021).

[8] Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." In 2020 international conference on electrical and electronics engineering (ICE3), pp. 452-457. IEEE, 2020.

[9] Amen, Khalid, Mohamed Zohdy, and Mohammed Mahmoud. "Machine learning for multiple stage heart disease prediction." In Proceedings of the 7th International Conference on Computer Science, Engineering and Information Technology, pp. 205-223. 2020.

[10] Assegie, TSEHAY ADMASSSU. "A support vector machine based heart disease prediction." J Softw Eng Intell Syst 4 (2019): 111-116.