Proceedings of the Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC-2023)
IEEE Xplore Part Number: CFP23V66-ART; ISBN: 979-8-3503-0009-3

2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) | 979-8-3503-0009-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICESC57686.2023.10193657

Performance Comparison Analysis of Predicting the Heart Diseases using Machine Learning Algorithms

Dhruva R. Rinku
Department of ECE
CVR College of Engineering
Hyderabad, India
dhruva.rinku@cvr.ac.in

Sandhya Devi Gogula
Dept. of CSE
Gitam School of Technology,
GITAM, Visakhapatnam,
Andhrapardesh, India.
sgogula@gitam.edu

Swarnalatha Prathipati
Dept. of CSE
Gitam School of Technology,
GITAM, Visakhapatnam, Andhra
pradesh, India
sprathip2@gitam.edu

P V Ramana Murthy
Dept. of CSE
Malla Reddy Engineering College,
Hyderabad, India
ramanamurthy19@gmail.com

Rokesh Kumar Yarava
Dept. of CSE
Chalapathi Institute of Engineering
Technology (A), Guntur, India
rokeshy12@gmail.com

Uma devi Kosuri
Department of H & S,
Gokraju Rangaraju Inst., of Engg., &
Tech., Hyderabad, India..,
umadevi.kosuri@gmail.com

Abstract—**In the medical field, the process of Heart Disease (HD) prediction process is a challenging task even in the modern digital world. Even though the data generated by the healthcare industries are huge, the data scientists are working tremendously to determine the correlation between the various parameters that causes the HDs. Therefore, there exists a need to predict the HDs to safeguard the human kind. The proposed method uses the Machine Learning (ML) models to predict the HD based on the existing symptoms of the patients. The dataset from the UPI repository is used to evaluate the performance of the proposed models. The various parameters namely precision (p), recall (r), and accuracy (a) are used evaluate the performance measures of the ML models. Observing the results concluded that, the Random Forest model outperformed the other models such as XGBoost, Decision Tree and traditional Neural Network model regarding the prediction accuracy with respect to UCI dataset.**

Keywords—Heart Diseases, Random Forest, XGBoost, Decision Tree, Neural Networks, Predicting performance

I. Introduction to the Heart Diseases

Heart is an organ in human body that pumps and circulates blood to the lungs and other parts of the human body through arteries and veins. The oxygen enriched (pure) blood from the lungs is pushed to the brain and the low oxygen (impure) blood from the brain is pushed to the lungs for purification. The flow of blood is controlled by several organs such as valves, capillaries and many more. The contraction and relaxation of the myocardium determine the heartbeat. An interruption in the function of any of the organs is termed as cardiovascular disease [1].

Some Cardiovascular diseases might arise due to: Irregular heartbeats, Thinning of Blood vessels thus interrupting transport of blood, Flaw in heart valves' functions, Genetically affected blood vessels, Aging, and Rheumatic disease. Risk factors of cardiovascular disease include: high blood pressure, high cholesterol, diabetes- 2, lack of physical exercise, obesity, chronic autoimmune inflammatory or kidney disease, toxaemia, alcohol consumption, and by bloodline. Figure 1 depicts the structure of the human heart compiled form [2].

The aim of this research is to investigate the application of machine learning algorithms for predicting heart disease based on a dataset containing various factors associated with the disease. The research focuses on identifying the most effective machine learning models for accurate prediction and evaluating their performance using precision, recall, and accuracy metrics.

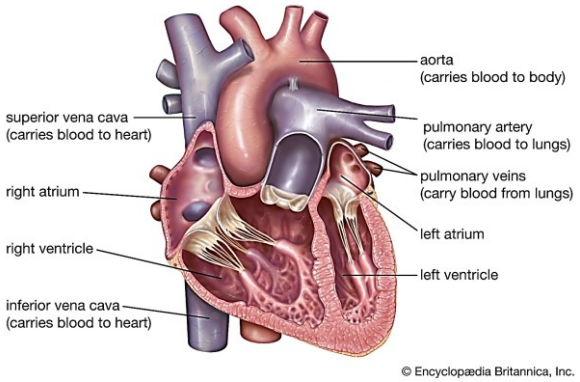

Figure 1. Human heart structure (Courtesy: Source [15])

II. Summary of Existing Approaches

The following section illustrates the summary of the existing approaches. Table 1 summarizes the significant features of the existing approaches.

Kwakye and Dadzie [3] devised an algorithm to predict Cardio vascular disease based on the Kaggle data of the patients. The classification algorithms were applied on unbalanced data and balanced data. In order to increase mean accuracy and balance the data, the data obtained in pre-processing phase was transformed Synthetic Minority Oversampling Technique (SMOTE). The Cross validation accuracy and Hold-out prediction values using ROC-AUC for both unbalanced and balanced data were determined for several classification algorithms and compared. The results depicted

accuracy values: Logistic regression (0.728592) and Naïve bias (0.707762) outperformed in unbalanced data while Random Forest (0.946337) and KNN (0.886542) performed in balanced data.

Padmaja and team members [4] predicted heart diseases using Machine Learning Classification Models on Cleveland dataset. Successive to pre-processing, the features were selected by Chi Square Test to increase the performance and decrease execution time of the classification algorithms. Among other classification algorithms, Random Forest classifier yielded 93.44% accuracy.

Ahmed and his team members [5] predicted Heart Failures through Machine Learning algorithms categorised as Linear, Ensemble and Boosting. The score such as Accuracy, Recall, Precision and F1were computed for all the classifiers like KNN, SVM, RF, GNB, CatBoost, GBC, ABC, DT, XGB, LGBM, HGBC, and MNB. The results exhibited RF, GBC and CataBoost outperformed other classifiers with the accuracy of 87.93%.

Karthick and his team members [6] predicted heart diseases using Machine learning models SVM, LR, XGB, LGBM, RF and GNB. The Cleveland dataset consists of several features including missing attributes. The key features were identified by Chi square test to overcome the issue and increase efficiency. The comparative results depicted RF classifier outperformed other with the accuracy of 88.5%.

Sarra, Dinar, Mohammed and Abdulkareem [7] enhanced heart disease prediction model through statistical feature selection model (Chi square) before applying the SVM classifier. The Cleveland and Statlog dataset were used for the prediction model. The correlation heatmap predited for two datasets showed the degree to which the attribute was correlated with target class. The metric values Accuracy, F1, Specificity and Sensitivity were evaluated for both the datasets and compared. The Statlog dataset performed superior with accuracy values of 89.7% (with Chi square) and 85.29%(without Chi square) respectively.

Xiao and his research team [8] proposed Deep Residual Neural Network for Heart Disease prediction. Experiments using machine learning classifiers such as NB, LR, DT, KNN and RF on the UCI dataset proved Logistic Regression was superior with the accuracy of 87% followed by Random Forest with 83%.

Chandrika and Madhavi [9] devised Hybrid Random Forest with Linear Model (HRFLM) to predict heart diseases and compared the results of metrics such as Accuracy, Specificity, Sensitivity, F measure, Precision and classification error with other machine learning algorithms – NB, GLM, LR, DL, DT, RF, GBT, SVM, and VOTE respectively. The results exhibited the proposed HRFLM showed accuracy of 88.4%.

Nayeem, Rana, and Islam [10] predicted Heart diseases using Machine Learning algorithms by imparting some techniques in the dataset such as imputing mean value technique for handling null values; info-gain feature technique for selecting the features from the Kaggle dataset. The classification algorithms such as KNN, NB and Random forest were experimented on the dataset by computing the metrics. The results depict Random Forest attained classification accuracy of 95.63%.

Malavika, Rajathi, Vanitha and Parameswari [11] projected heart diseases using Machine learning algorithms NB, LR, RF, SVM, DT and KNN. The pre-processing phase of UCI dataset proved men were more exposed to heart disease than women. The performance of classification algorithms through Confusion matrix depicted Random Forest to be superior by 92.59%. Moreover, among the classifiers, Random Forest showed greater accuracy by 91.80%.

Manjula and her research team [12] devised a machine learning model to forecast heart attack. The model involve several phases such as Data acquisition, Pre-processing, Model Stacking which includes experimenting the data on classifiers – LR, KNN, NB, DT, SVM, XGBoost, and RF. Upon evaluating the performance based on accuracy, the possibility of heart attack is predicted. The Random Forest Classifier ascertained its lead with 90.16% accuracy.

Yewale, Vijayaragavan, and Munot [13] surveyed Decision Support System to predict Heart Disease using UCI dataset. After a detailed survey of heart disease and its risk factors, the classification models such as ANN, DNN, Mulitlayer perceptron, LR, NB, RF, KNN, DT were experimented on the data to predict heart disease. The results were compared with proposed hybrid classifier RF+Chi-PCA classifier; proposed hybrid technique accuracy improved to 98.7%.

Jindal, Agrawal, Khera, Jain and Nagrath [14] devised HDPS using the machine learning classifiers KNN, LR, and RF. The system was investigated on UCI dataset of 304 patients. The classifier algorithms efficiency on pre-processed data showed best accuracy of 88.5% attained by KNN and LR.

Desai and Mantri [15] ascertained a hybrid ML model for predicting CVDs based on MLs- LR, NB, RF, XGB, KNN, DT, SVM. On the basis of accuracy (91.8%, 88.5% and 88.5%) obtained, a hybrid stacking technique is built using XGB, KNN and SVM. In other words the highest accuracy MLs are fused to increase accuracy and efficiency. The proposed hybrid model yielded 93.4% accuracy.

TABLE I. SIGNIFICANT FEATURES OF EXISTING APPROACHES

Ref. No.	Description	Dataset	Accuracy
[1]	Preprocessing, Balancing data by SMOTE, Classification and Prediction of Cardiovascular disease by classification algorithms	Framingham dataset from Kaggle	Unbalanced data: LR - 0.728592, Balanced data: RF-0.946337
[2]	Pre-processing, Feature selection by Chi square test, prediction of disease by classification algorithms such as GNB, LR, RF, KNN, SVM, MNB, DT and Gradient Boosting	Cleveland dataset	Random Forest-93.44%
[3]	Evaluation of Accuracy, Precision, Recall and F1 score for Linear ML, Ensemble ML, Boosting ML. Comparison of results	Kaggle, UCI, Real world - Heart Failure (HF) dataset	Accuracy: RF, GBC, CatBoost: 87.93%
[4]	Pre-processing, Feature selection by Chi square, HD prediction through	Cleveland heart	Random Forest: 88.5%

	GNB, SVM, LGBM, XGB, LR, RF	disease dataset	
[5]	Pre-processing, Feature selection through Chi-square model, SVM classifier	Cleveland and Statlog dataset	Statlog Dataset With Chi2: 89.7% Without Chi2: 85.29%
[6]	Deep Residual Neural Network	UCI Repository	Accuracy of Logistic Regression - 87%
[7]	Hybrid Random Forest Linear Model	UCI dataset	Accuracy: 88.4% F Measure: 90%
[8]	Pre-processing: Imputing mean value technique, info-gain feature selection technique, classification and prediction using KNN, NB and Random forest	Kaggle	Random Forest Accuracy:95.63%
[9]	Pre-processing, ML algorithms NB, LR, RF, SVM, DT and KNN for disease prediction	UCI	Accuracy of RF: 91.8%
[10]	Acquisition, Pre-processing, Model stacking through LR, KNN, NB, DT, SVM, XGBoost, and RF	NA	RF: 90.16%
[11]	Prediction of heart disease through MLs: ANN, DNN, LR, NB, RF+Chi PCA, KNN, DT	UCI	RF+CHI-PCA: 98.7%
[12]	Pre-processing, HD prediction using ML classifiers: KNN LR, and RF	UCI dataset of 304 patients	KNN and LR: 88.5%
[13]	Pre-processing, MLs-LR, NB, RF, Extreme GB, KNN, DT, SVM, and Stacking classifier technique(KNN, XGBoost, SVM)	UCI dataset with 303 instances	Hybrid stacking technique - 93.4%

III. PROPOSED METHOD

The objective of the proposed work is to predict the Heart Diseases (HDs) based on the existing symptoms of the concerned patient using ML algorithm. The dataset adopted to evaluate the present work is downloaded from the UCI repository [16]. Even though there are 76 attributes in the dataset, only four attributes are considered for predicting HDs. The attributes considered for a person are as follows: (1) age (2) gender (3) chest pain type (4) Blood pressure during resting period (5) serum cholesterol (6) blood sugar during fasting period (7) electro-cardiographic (ECG) results during the resting period (8) determined heart beat rate (9) induced angina during the exercising period (10) the old peak value of ST depression created due to exercise from rest (11) the peak slope value of the ST segment (12) major vessel(s) count (13) thalassemia represents the inherited features that affects haemoglobin level and finally (14) the target variable represents the presence of HD. Initially, the Exploratory Data Analysis (EDA) is done to understand the correlation among the various features with respect to the target field. Moreover, during the EDA phase, more insight about the dataset is attained and their exploration leads to identify the better Machine Learning model for predicting the HDs.

- Initially the shape of the dataset is determined. Its size is 303X14; that is 303 rows and 14 columns; the rows and columns specify the number of patients and their symptoms respectively.
- Later, the correlation between the columns is determined. For example, the percentage of persons with and without HD is visualized. Figure 2 illustrates the correlation between the person's with and without HD.
- Similarly the other parameters are being visualized with respect to target parameter to understand and explore the dependency among them. Figure 3 illustrates the correlation between the gender and target variable. The conclusion is that the female gender is affected more than male gender.
- Figure 4 depicts the correlation between the tupe of the chest pain and target variable. Observations conclude that the person's with chest pain type (angina) are affected less than all other types.
- Furthermore, the correlation between blood sugar during testing period doesn't have significant impact on the target variable.
- Figure 5 depicts the correlation between ECG during the resting period and target variable. Observation conclude that persons with '0' and '1' restecg are affected more than the persons with restecg=2.
- Figure 6 depicts the correlation between the exang and the target variable. Observation leads to the conclusion that the person(s) with exang=1, that is, angia induced due to exercise are less affected to HD.
- Figure 7 depicts the correlation between the slope parameter and the target variable. Observation leads to the conclusion that the persons with slope=2 are affected more than the persons with slope=1 and 2, respectively.
- Figure 8 depicts the correlation between the ca (the major vessel count) and the target variable. Observation leads to the conclusion that the persons with ca=4 are affected more than the rest of the persons.
- Figure 9 illustrates the workflow of the proposed method. The following segment depicts the process of dataset exploration using EDA.

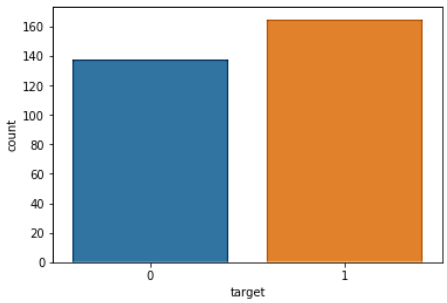

Figure 2. Correlation between the persons with and without the HD

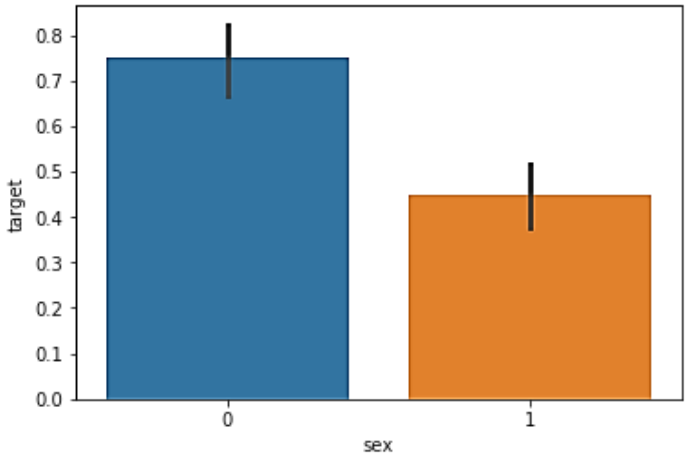

Figure 3. Correlation between the Gender and the Target variable.

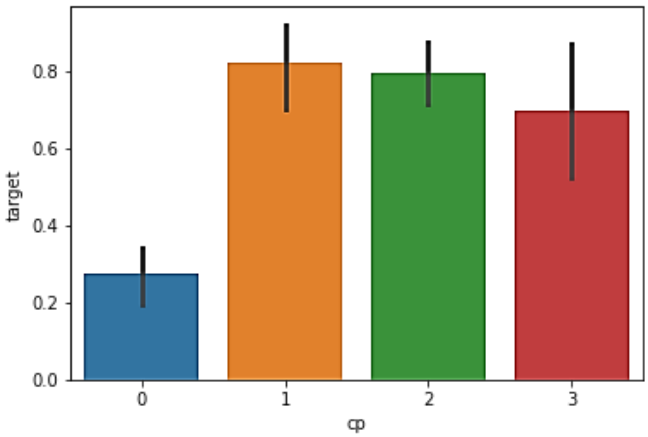

Figure 4. Correlation between the Type of the Chest Pain and the Target Variable

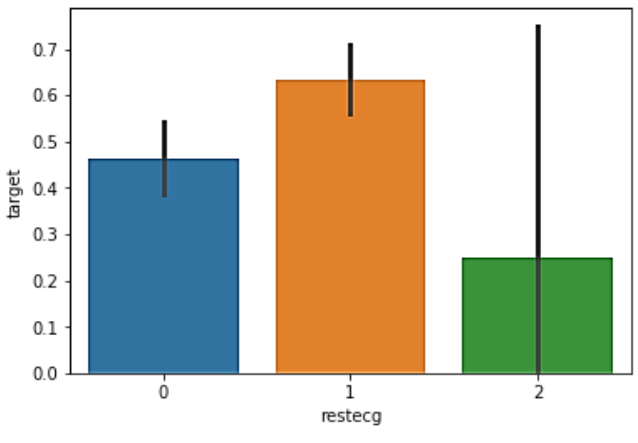

Figure 5. Correlation between the restecg parameter and the Target

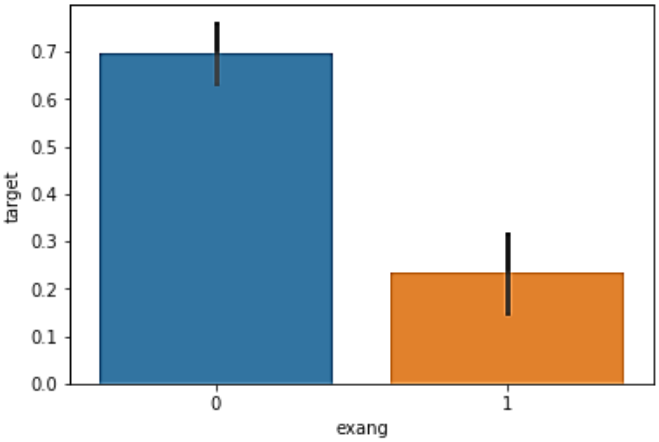

Figure 6 Correlation between the exang parameter and the Target

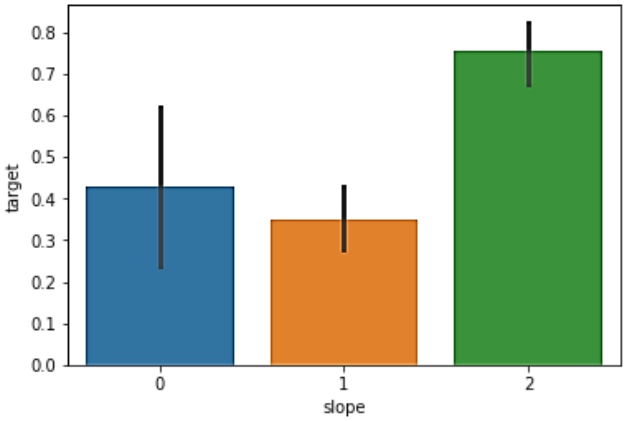

Figure 7. Correlation between the slope parameter and the Target

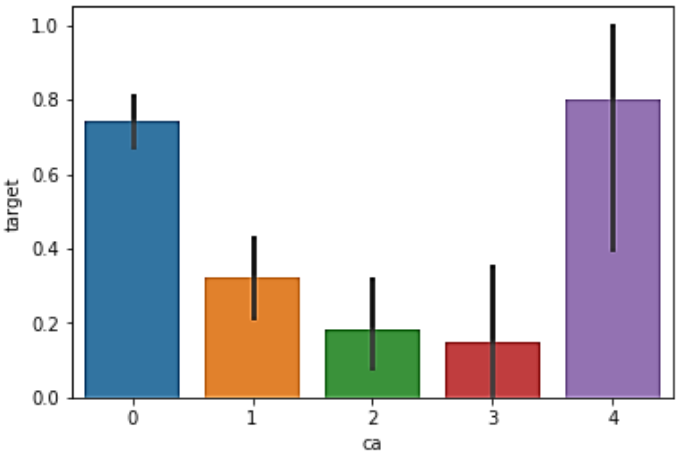

Figure 8. Correlation between the ‘ca’ parameter and the Target

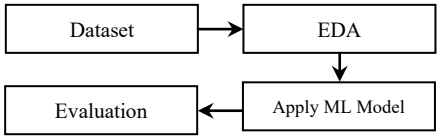

Figure 9. Proposed Method's Workflow

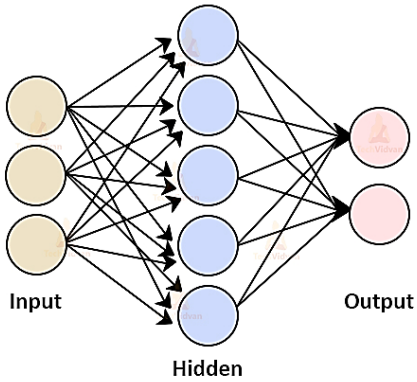

Figure 10. NN Architecture (Courtesy: Source [17])

IV. Experimental Results

Finally, the dataset is split into 80:20 ratios for the training and testing purpose, respectively. Now, the models are chosen for predicting the possibilities of HD with respect to various factors. The distinct models adopted in the proposed method to determine the HDs are (1) Decision Tree (DT), (2) Random Forest (RF), (3) neural network (NN), and (4) XGBoost (XGB). Figure 10 depicts the architecture of NN adopted in the proposed model, with one input, hidden and output layer, respectively. The parameters used to determine the efficiency of the proposed ML models are precision (p), recall (r) and the accuracy (a). Since F1 score will result in significant performance regarding the imbalanced datasets, in the proposed method, the parameters adopted are 'p', 'r' and 'a' for evaluating the ML models. Equations 1, 2, and 3 are used to determine the 'p', 'r', and 'a' respectively.

$$p = \frac{number\ of\ persons\ correctly\ identified\ with\ HD}{number\ of\ predicted\ to\ have\ HD} \quad (1)$$

$$a = \frac{number\ of\ persons\ correctly\ predicted}{dataset's\ size} \quad (2)$$

$$r = \frac{number\ of\ persons\ correctly\ predicted\ with\ HD}{number\ of\ persons\ with\ HD\ in\ the\ dataset} \quad (3)$$

Table 2 illustrate the performance measures of the various ML algorithms adopted in the present method regarding HD predictions.

Observing the performance measures from the Table 2, the RF has significant performance than the other models in terms of accuracy. Even though, the NN model has significant performance, there is room for improving the performance with the increased training ratios.

The selection of machine learning models for predicting heart disease depends on various factors such as the nature of the problem, the available data, the interpretability requirement, and the desired performance metrics. The following section illustrates the reason for choosing the proposed ML models.

- DTs provide intuitive decision rules based on feature values and are easy to interpret.
- RF performs well on high-dimensional datasets, and handle missing values and outliers effectively.
- NN can learn complex patterns and relationships from the `data, capturing intricate interactions between features.
- XGBoost is specialized for its speed, scalability, and ability to handle imbalanced datasets.

The estimation errors are determined using mean-square error parameter and its values are 0.24, 0.12, 0.29, and 0.19, for DT, RF, NN and XGB models, respectively.

TABLE II. Performance Measures

ML Models	p (%)	r (%)	a (%)
DT	0.82	0.84	83.18
RF	0.97	0.95	96.02
NN	0.80	0.79	81.15
XGB	0.89	0.89	88.54

V. Conclusion and Future Enhancements

In general, the individuals affected with HDs are increasing rapidly day-by-day due to various factors. To safeguard the human life from the dangerous heart disease, there is a need for a prediction system to identify the heart issues and suggesting subsequent remedial measures. Hence, by considering the various factors that affects the HDs, the proposed approach aimed at implementing the HD prediction approaches using the ML models. The dataset from UCI repository was adopted to evaluate the performance of the ML models in the present approach. Among the various models, the RF model showed the remarkable performance regarding the prediction accuracy than the other models namely DT, NN and XGB models. Further the hybrid RF model shall be adopted to predict HDs in the future work. Moreover, the performance of the ML models can be improved by Hyperparameter Tuning method. Hyperparameters control the behavior of the models and can significantly impact their performance.

References

[1] "Cardiovascular Disease: Types, Causes & Symptoms," Cleveland Clinic, Available: https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease

[2] Encyclopedia Britannica, "Heart," Encyclopedia Britannica. [Online]. Available: https://www.britannica.com/science/heart.

[3] K. Kwakye, & E. Dadzie, Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases, North Carolina Agricultural and Technical State University, USA, 2021, http://arxiv.org/abs/2112.01503

[4] B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, & E. Krishna Rao Patro, Early and Accurate Prediction of Heart Disease Using Machine Learning Model. Turkish Journal of Computer and Mathematics Education 4516 Research Article, vol. 12, no. 6, 4516-4528, 2021.

[5] S. Ahmed, S. Shaikh, F. Ikram, M. Fayaz, H. S. Alwageed, F. Khan and F. H. Jaskani, Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models,

Journal of Sensors, vol. 2022, Article ID 3730303, 2022, https://doi.org/10.1155/2022/3730303

[6] 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, Computational and Mathematical Methods in Medicine, vol. 2022, Article ID 6517716, 2022. https://doi.org/10.1155/2022/6517716.

[7] R. R. Sarra, A. M. Dinar, M. A. Mohammed and K. H. Abdulkareem, Enhanced heart disease prediction based on machine learning and $\chi 2$ statistical optimal feature selection model, Designs, vol. 6, no. 5, 2022, https://doi.org/10.3390/designs6050087

[8] N. Xiao, Y. Zou, Y. Yin, P. Liu and R. Tang, DRNN: Deep residual neural network for heart disease prediction, Journal of Physics: Conference Series, vol. 1682, no. 1, 2020, https://doi.org/10.1088/1742-6596/1682/1/012065

[9] L. Chandrika and K. Madhavi, A hybrid framework for heart disease prediction using machine learning algorithms. E3S Web of Conferences, vol. 309, 2021, https://doi.org/10.1051/e3sconf/202130901043

[10] M. J. Nayeem, S. Rana, and M. R. Islam, Prediction of heart disease using machine learning algorithms, European Journal of Artificial Intelligence and Machine Learning, vol. 1, no. 3, pp. 22-26, 2022, http://dx.doi.org/10.24018/ejai.2022.1.3.13

[11] G. Malavika, N. Rajathi, V. Vanitha and P. Parameswari, Heart disease prediction using machine learning algorithms, Biosc. Biotech. Res. Comm., SI. Vol. 13, no. 11, pp. 24-27, 2020.

[12] P. Manjula, U. R. Aravind, M. V. Darshan, M. H. Halaswamy and E. Hemanth, Heart attack prediction using machine learning algorithms, International Journal of Engineering Research & Technology, vol 10, no. 11, pp. 324-327, 2022.

[13] D. Yewale, S. P. Vijayaragavan and M. Munot, Decision support system for reliable prediction of heart disease using machine learning techniques: An exhaustive survey and future directions, International Journal of Engineering Trends and Technology, vol. 70, no. 4, pp. 316-331, 2022, https://doi.org/10.14445/22315381/IJETT-V70I4P228

[14] H. Jindal, S. Agrawal, R. Khera, R. Jain and P. Nagrath, Heart disease prediction using machine learning algorithms, IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, pp. 1-11, 2021, https://doi.org/10.1088/1757-899X/1022/1/012072

[15] U. Desai and S. Mantri, Hybrid Model of Machine Learning Algorithms for Prediction of Cardiovascular Disease, Journal of Positive School Psychology, vol. 6, no. 6, pp. 10551-10560, 2022.

[16] "UCI Heart Disease Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[17] Shubham Gupta, "Architecture Of Artificial Neural Network," Knoldus Blogs, 28-Dec-2021, https://blog.knoldus.com/architecture-of-artificial-neural-network/.