EasyChair Preprint

№ 15706

Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction Using CDC and BRFSS Data: a Focus on Oversampling and Ensemble Techniques.

Ridwan Amokun, Taofik Arowolo and John Eke

January 13, 2025

Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction Using CDC and BRFSS Data: A Focus on Oversampling and Ensemble Techniques.

[1]R.A. Amokun, [2,3]O.T. Arowolo, & [1]J.I. Eke

[1]Department of Mechanical Engineering, University of Lagos, Akoka
amokunridwan@gmail.com, ekejohnigwe@gmail.com
[2]Department of Mathematical Science, Lagos State University of Science and Technology, Ikorodu
taofik.arowolo@outlook.com

Abstract – *This work studies machine learning methods to predict heart disease based on data obtained from the CDC via the Behavioral Risk Factor Surveillance System. Accordingly, it compared various models, including Logistic Regression and Random Forest models, which can be further tuned for better outcomes in heart disease treatment and prevention.*
With a view to handling the class imbalance problem in heart disease classification, the SMOTE technique was applied, and model performance was evaluated on metrics such as Accuracy and Precision, among others. The high marks the F Score and ROC Area under the Receiver Operating Characteristic curve used in the evaluations were notably displayed by the XGBoost model with an F Score of 0.80. 0.92 ROC area Further application of SMOTE contributed to the identification of the minority cases; therefore, the models can assure balanced and reliable predictions. This study has shown how machine learning and techniques of oversampling can be used to better the diagnosis of heart disease, thus having health professionals equipped with tools for early diagnosis and timely treatment. Many algorithms are used in this study, as well as ensemble techniques, which provide a really strong basis for predictive modeling in the healthcare sector.

Keywords: CDC Data, BRFSS, SMOTE, Ensemble Techniques, XGBoost, Logistic Regression, Random Forest.

1. Introduction

Heart diseases still remain a significant health concern globally. It is ranked among the leading causes of death in the world and also contributes to a great number of preventable deaths yearly. In the United States of America, it causes one out of every four deaths, apart from its enormous impact on public health as described by Mohapatra, 2022. The WHO estimates that, worldwide, heart diseases could claim as many as 24 million lives annually by 2030 if urgent interventions are not affected, and as such, it may become one of the biggest challenges for health systems in the years to come. These numbers hint at a desperate call to bring improvement to early detection and treatment strategies in order to fight this emerging menace.

Heart disease can be grouped into coronary artery diseases, the most common of all; congenital heart defects; and arrhythmias-all three of them presenting different kinds of problems due to their subtle onsets and

variable symptoms. It may go along asymptomatically for many years until it causes fatal events, such as a heart attack or stroke, with little warning. The risk factors are unhealthy behaviors of smoking, poor diet, inactivity, and excessive use of alcohol, plus the chronic conditions of diabetes and hypertension. The CDC projects that by 2023 these interrelated factors will make early detection and treatment even more difficult because of the tangled web of risks involved. The current study is part of the growing interest in the application of machine learning techniques to healthcare, with an increasing special interest in the prediction of heart disease. One of the most difficult tasks related to heart disease prediction is class imbalance, since there are more patients without heart disease than diagnosed ones. In such an imbalanced dataset, model predictions always tend to be biased towards the majority class, in which patients are not suffering from heart disease, and predict poorly the actual patients who are at risk. Hence, SMOTE-the Synthetic Minority Oversampling Technique-will be used to create synthetic data for the minority class in order to balance the dataset. It increases not only the capability of the model to learn from the majority and minority cases but also, at the same time, improves SMOTE's predictions for the groups that are at risk.

Data from the CDC's Behavioral Risk Factor Surveillance System have been useful in getting information at the individual level on various health behaviours, conditions, and risks related to heart disease. Generally speaking, the ensemble techniques of Bagging and Stacking have tended to perform better in many cases because they combined outputs that boosts their accuracy and robustness. The current work investigates how such ensembling methods can perform better compared to state-of-the-art models for early detection in heart diseases. Among those, XGBoost has always been the top performer in prediction tasks for its efficiency and scalability; thus, the current research will cover performance evaluation for the classifier with other XGBoost classifiers.

The purpose of this study is to outline a pathway that should be followed by healthcare professionals in order for ML models to be safely introduced into clinical practice, thus offering higher accuracy and more reliability in the predictions concerning heart diseases. While it is true that with the influence of personalized medicine, healthcare is becoming more personalized and data-driven, the applications of machine learning bring a unique opportunity for fundamentally revising how we treat one of the most debilitating health concerns of our time: heart disease.

2. Literature Review

These works predominantly focus on the integration of data-driven insights to facilitate early diagnosis that is highly accurate, sensitive, and reliable. Ahsan et al. (2022) further emphasize that machine learning might be a game-changing innovation in healthcare since clinicians are empowered to make more knowledgeable decisions based on complex patient data. Bhatt, 2023, has made a similar point by mentioning flaws in traditional diagnostic methods, which range from misdiagnosis to delaying the timely detection of fatal heart diseases. Several machine learning models were tried on heart disease datasets, such as

logistic regression, Decision Trees, and Random Forest, with some impressive predictive scores. For example, Ali et al. (2021) demonstrated that even simple machine learning algorithms, such as KNN and Decision Trees, if complemented with feature importance analysis, provided high sensitivity and accuracy in early-stage heart disease detection, while the accuracy of some classifiers was as high as 100%, thereby revealing the potential of these models in building effective predictions.

Similarly, Mohapatra et al. (2022) examined the ensemble approach, more precisely stacking, and noticed that the combination of more models definitely raised the level of prediction up to an accuracy of 92%. This underlines the fact that a combination of models improves their generalization for different datasets. Newer algorithms such as XGBoost and CatBoost have taken a leading position in heart disease prediction due to the fact that they support high volume and complex feature interactions not possible in simple algorithms. Rajendran et al. (2022) identified favorable attributes of such models by identifying that EFE can be employed with XGBoost for optimizing important predictor selection to strengthen robustness in the models.

Another challenge that keeps arising in heart disease prediction relates to class imbalance, since usually there are many more patients without heart disease in a dataset than those who actually have the disease. The main effect of this class imbalance is that it biases predictions toward the majority class and hence misses the at-risk patients. Recent research has tried to balance this inequality. The common solution includes the Synthetic Minority Over-sampling Technique, SMOTE, suggested by Chawla et al. (2002), which creates artificial data from the minority class in order to balance the dataset. Hasanova et al. (2022) applied SMOTE in conjunction with blockchain technology for ensuring the security and integrity of patient data management to ensure that minority cases fell suitably within a model.

However, oversampling techniques run the risk of overfitting of the models to the synthetic data; the resultant models fail to generalize well. Rajendran 2022 illustrated that this can be minimized by careful model tuning and stringent cross-validation to allow high accuracy in detecting minority cases without loss of model robustness. This has remained so far one of the most successful approaches to improving heart diseases predictions. Many models come together to become the best prediction. In fact, Breiman's Random Forest (2001) was the first great ensemble method; showing that aggregating decision trees was much more predictive as it captured a wider range of patterns in the data, and reduces overfitting, making it ideal for the prediction of heart diseases.

More recent studies have extended the work to include more sophisticated ensemble methods. In that respect, Ozcan and Peker (2023) explored stacking and bagging ensembles, a method of combining weak learners into one stronger predictor. Herein, it was established that the stacking models outcompeted the base models both in accuracy and ROC-AUC score. Similarly, Liang and Guo 2023 reported that ensemble methods, especially in deep learning ensembling, improved handling high-dimensional heart disease datasets. Techniques included the Bagging and Stacking heart disease data, where Bagging, through its average predictions over multiple

subsets, reduces variance; further, stacking trains a meta-model to learn from the predictions of base models, enhancing accuracy and stability.

Although much has been achieved in improving heart disease forecasting with the incorporation of machine learning, deep learning algorithms still hold many unseen variables. Deep learning has been a great potential in many fields, but in heart disease prediction, it's under-explored when compared with classic models. Further studies may combine deep learning techniques with ensemble methods to construct a model capable of processing more challenging datasets with high accuracy.

3. Methodology

The methodology for this study is highlighted as follows:

3.1 Data collection

The dataset used in the research is from the Behavioral Risk Factor Surveillance System, taken from Kaggle. These variables include BMI, smoking, and alcohol consumption in order to predict heart disease. The data preparation involves imputation of missing values, encoding categorical data using one-hot encoding, and standardization of numerical features in order to be consistent. SMOTE was used to handle the imbalanced class distribution in this dataset and generated synthetic samples of the minority class. The preparation of the dataset was thus complete, ensuring a balanced and robust dataset for heart disease prediction.

3.2 Model Development

Several machine learning algorithms have been applied to come up with models that can predict the presence or absence of a heart disease. The models were tuned with hyperparameter tuning using cross-validation. The following algorithms were employed in the model development:

- Logistic Regression
- Random Forest
- Decision Trees
- Support Vector Machines
- K-Nearest Neighbors
- Gradient Boosting Methods
- Ensemble Methods

3.3 Machine Learning Models

A mathematical summary of the machine learning models is as follows:

3.3.1 Logistic Regression

The logistic regression model is to use in this case to predict binary outcomes. The probability $P(y=1|X)$ that a patient has heart disease is modeled using the well-known logistic (sigmoid) function:

$$P(y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \quad (1)$$

Where:

- β_0 is the intercept,
- $\beta_1,.....,\beta_n$ are the coefficients,
- $X_1,.....,X_n$ are the features.

The model is then trained using maximum likelihood estimation to minimize the error between predicted probabilities and actual labels.

3.3.2 Random Forest

Random forest is one of the most popular ensembles learning algorithm that constructs multiple decision trees and then aggregates their results (Breiman, 2001). The prediction for a classification problem is based on a majority vote from all decision trees.

$$\bar{y} = mode(T_1(x), T_2(x), ..., T_m(x)) \quad (2)$$

$T_1(x), ..., T_m(x)$ are decision trees trained on different subsets of the data.

The Gini index is used to split the nodes in decision trees:

$$Gini(p) = 1 - \sum_{i=1}^{c} p_i^2$$

Where p_i is the probability of a class at a particular node.

3.3.3 CATBoost (Category Boosting)

CATBoost is a gradient boosting algorithm designed for categorical data. It minimizes the following loss function over trees:

$$L(\theta) = \sum_{i=1}^{N} l(y_i, f(x_i, \theta)) + \lambda \| \theta \|^2 \quad (3)$$

Where:

- l is the loss function (log loss for classification),
- $f(x_i, \theta)$ is the model prediction,
- λ is a regularization parameter to control overfitting.

3.3.4 KNN (K-Nearest Neighbour)

KNN is a non-parametric algorithm that helps to classify a data point been focused on based on the majority class among its k nearest neighbors. Typically, its known to use a distance metric to find the k closest data points to the input and assigns the label based on the majority class of those neighbors.

Then we have that the Euclidean distance d between points x and y is given by:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

Where x_i and y_i are the feature values for points x and y in the i-th dimension.

3.3.5 Support Vector Machines (SVM)

The SVM algorithm aims to find the optimal hyperplane that maximally separates the classes while minimizing classification errors.

The objective function in SVM is to minimize:

$$\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \quad (5)$$

Where:

- w is the weight vector defining the orientation of the hyperplane,
- C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors,
- ξ_i are slack variables that allow some of the data points to be misclassified, thereby enabling soft margin classification.

For the class prediction, the SVM decision function f(x) for a point x is given by:

$$f(x) = w \cdot x + b \quad (6)$$

where b is the bias term, and the classification boundary is defined by f(x)=0.

3.3.6 Ensembling

Ensembling is a technique that involves combining the predictions from multiple models to improve overall predictive performance. By aggregating the output of multiple classification models, ensemble methods aim to produce a more accurate and robust model.

Some of the common types of ensembling techniques are:

- **Bagging (Bootstrap Aggregating)**: Combines predictions from multiple models trained on different subsets of data to reduce variance.
- **Boosting**: Sequentially involves trains models, focusing on the misclassified points to reduce possible bias.
- **Stacking**: This combines the predictions of the base models using a meta-model, which learns from these predictions to improve accuracy of predictions.

The final prediction in an ensemble model is typically:

$$y = \frac{1}{M}\sum_{m=1}^{M} y_m \tag{7}$$

where M is the number of models in the ensemble and y is the prediction of the m-th model

3.4.3 Evaluation Metrics

ROC curves have significance in the context of prediction models. This portrays the capability of the model in differentiating patients with and without heart disease by plotting the true positive rate as a function of the false positive rate. The performance is summarized by AUC; 1 represents an AUC close to 1 if the predictive model is strong. Since the class imbalance is inherently available, with more people not having heart disease in the heart disease dataset, ROC-AUC will be particularly effective. It considers both sensitivity and specificity-identifying actual non-heart disease cases, offering a more complete picture than accuracy can show alone. The ROC-AUC, as applied in this study, balances the dataset with SMOTE and ensures a fair model performance on both the minority and majority classes, showing that it is strong for detecting heart diseases in early stages.

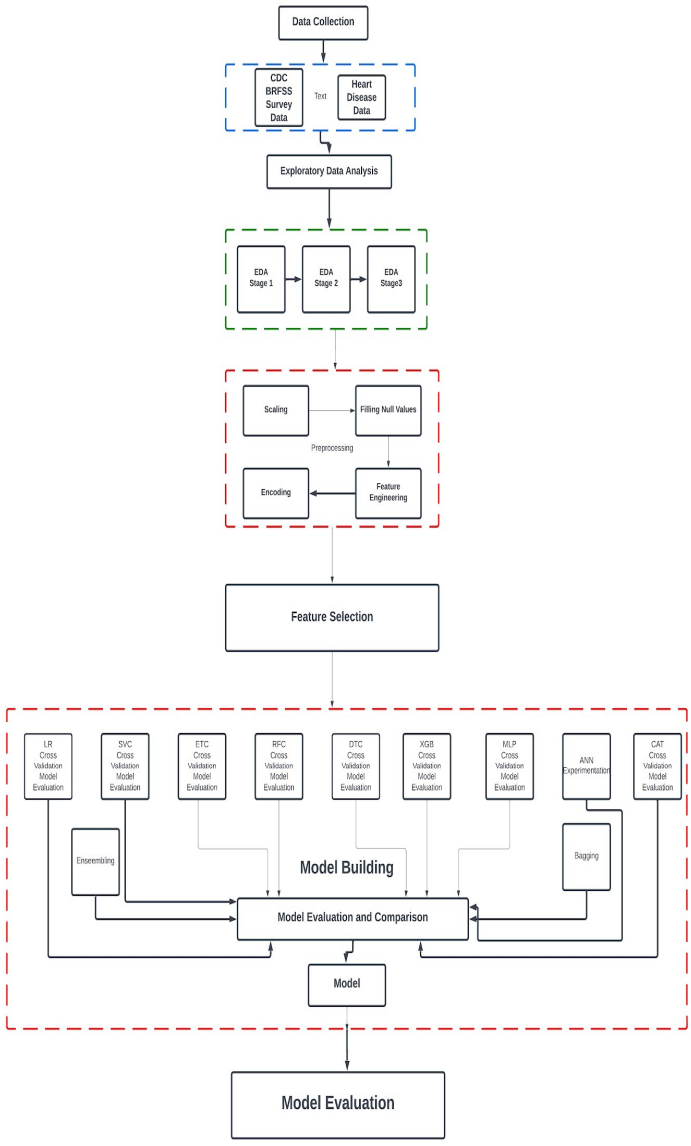

Figure 1: Systematic flow chart for model development and comparison

4. Results and Discussion

Tables 1-3 depict in detail how different models fare before and after the application of SMOTE to ensure a good comparison that

can be used to analyze and conclude which models perform best for heart disease prediction.

The comparison of results across different models shows fewer critical observations that emerged on the performance of the classifiers employed.

Table 1 Model Performance Before and After SMOTE (Positive Class - heart disease)

Model	Precision (Before SMOTE)	Recall (Before SMOTE)	F1-Score (Before SMOTE)	Precision (After SMOTE)	Recall (After SMOTE)	F1-Score (After SMOTE)
LR	0.51	0.08	0.14	0.72	0.72	0.72
KNN	0.32	0.07	0.11	0.85	0.91	0.88
RF	0.57	0.05	0.08	0.86	0.82	0.83
ET	0.59	0.03	0.06	0.45	0.09	0.15
DT	0.45	0.09	0.15	0.81	0.8	0.8
XGBoost	0.57	0.07	0.13	0.84	0.8	0.82
CatBoost	0.56	0.1	0.16	0.85	0.88	0.87
SVM	0	0	0	0.36	0.36	0.36
ANN	0.54	0.06	0.11	0.77	0.74	0.75

Table 2 Model Accuracy and ROC-AUC Before and After SMOTE

Model	Accuracy (Before SMOTE)	ROC-AUC (Before SMOTE)	Accuracy (After SMOTE)	ROC-AUC (After SMOTE)
LR	91.48%	0.8	71.76%	0.8
KNN	87.49%	0.58	87.49%	0.87
RF	91.55%	0.83	83.82%	0.92
ET	91.54%	0.82	91.54%	0.83
DT	91.29%	0.78	80.42%	0.89
XGBoost	91.61%	0.84	81.93%	0.91
CatBoost	91.62%	0.84	86.53%	0.95
SVM	91.46%	-	35.71%	-
ANN	91.53%	0.8	75.36%	0.83

Table 3 Summary of Classifier Performance After SMOTE (Positive Class Only)

Model	Precision	Recall	F1-Score	ROC-AUC
LR	0.72	0.72	0.72	0.8
KNN	0.85	0.91	0.88	0.87
RF	0.86	0.82	0.83	0.92
ET	0.45	0.09	0.15	0.83
DT	0.81	0.8	0.8	0.89
XGBoost	0.84	0.8	0.82	0.91
CatBoost	0.85	0.88	0.87	0.95
ANN	0.77	0.74	0.75	0.83
Stacking (Ensemble)	0.85	0.84	0.85	0.85
Bagging (Ensemble)	0.9	0.85	0.88	0.92

4.1 Performance of Classifiers without SMOTE (Imbalanced Data)

Figure 2: F1-Score for positive class after sampling

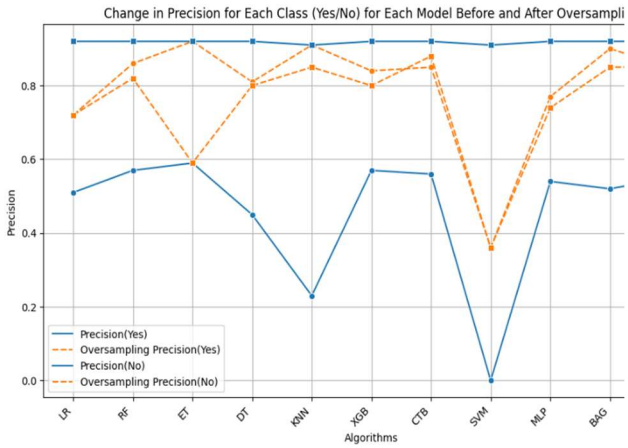

Figure 3: ROC areas of Models before and after sampling

First, the models were trained on an imbalanced dataset in the first stage, where the majority class-without heart disease-outweighed the minority class for patients with heart disease. These led to models that were performing high accuracies-most of them above 90%-which might seem good in the first case. However, a closer look reveals that this high accuracy arises almost entirely because the models can predict the majority class of patients, i.e., those without heart disease.

For example, Random Forest and Logistic Regression had a very high precision at 0.92 for the negative class, yet both had much lower precisions for the positive class, at 0.51. What was very low across all these models in this phase was the recall for positive cases-that is, hearts suffering from the disease. The K-Nearest Neighbor and Support Vector Machine models showed very poor performance for the two classes, returning low recall of 0.07 and 0.00, respectively. In a medical application where the aim is to identify all possible heart diseases for timely interventions, it is just unacceptable. The F1-scores were comparably quite low for the positive classes across the models, ranging from 0.11 to 0.16, and further established that without preprocessing due to data imbalance, the models were comparably bad at correctly predicting heart diseases.

4.2 Impact of SMOTE on Classifier Performance (Balanced Data)

In fact, for the case of the SMOTE oversampling technique, which was used to balance the dataset, there were quite a few improvements in the models presenting the positive class-that is, heart disease cases. In most of the models, precision and recall for the positive class went up noticeably. For example, KNN, which performed badly when the data was imbalanced, improved its precision to 0.85 and recall to 0.91 following oversampling.

Among these, the CATBoost algorithm performed really well, with a precision of 0.85 and recall of 0.88, hence a really good balance between precision and recall for both classes. Similarly, in the result, balancing significantly increased the F1-scores across all models, hence confirming that the classifiers became much more reliable in detecting positive cases.

4.3 ROC-AUC Score Improvements

Performance evaluation by the ROC curve and its AUC is a very important metric when dealing with imbalanced classification problems. Relatively high scores from 0.80 to 0.84 resulted for models trained using the imbalanced data. However, this was mostly due to the goodness of the models in predicting the majority class. Once SMOTE was implemented, for instance, ROC-AUC scores improved, including 0.92 for a Random Forest Model and 0.95 for a

CATBoost model, to name but a few, evidencing better discrimination between positive and negative cases.

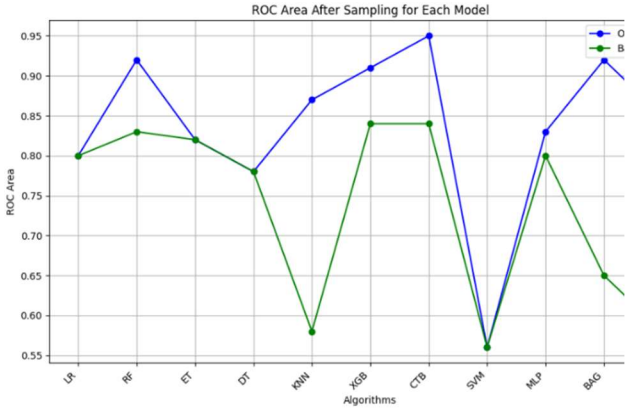

Figure 3: ROC areas of Models after sampling

4.4 Effectiveness of Ensemble Models

Another interesting approach was ensemble methods, such as bagging and stacking. By this, the stacking methodology combines the powers of several base models and yields a better performance metric: 84.79% accuracy, 0.85 precision, and 0.84 recall after SMOTE application. Similarly, the Bagging methodologies with particular scope on K-Nearest Neighbors as a base learner were strong and equated well in terms of accuracy and recall.

5. Conclusion

With implemented various machine learning algorithms for the prediction of the likelihood of heart disease through addressing class imbalance with SMOTE. The dataset initially was so imbalanced that the performance of the models was skewed toward the majority class, as most algorithms were doing great in predicting those patients that did not have heart diseases. This resulted in high accuracy and poor detection of actual heart disease cases, which can be represented by the low recall and F1-scores for the positive classes.

With SMOTE applied, the models clearly improved in their ability to predict heart disease with a lot better accuracy. Models such as Random Forest, CatBoost, and XGBoost turned out best with really high precision, recall, and overall F1-scores for both the majority and minority classes. Other ensemble methods combining the powers of individual algorithms, such as stacking and bagging, also helped increase the performance of models.

The class imbalance problem, as seen in this study, has emerged as a very critical issue in building reliable models on medical diagnosis for heart disease prediction. Algorithms that integrate boosting techniques with ensemble learning and other strategies such as SMOTE have proved to offer robust solutions in effective recognition of people at risk. Further refinements of these models can be deployed in real-world healthcare systems for early detection and intervention.

6. Recommendations

The result obtained shows that oversampling is vital to the generalization of the models. Consequently, an observable increase in the accuracy of prediction for the positive cases detected.

Based on this fact Bagging and Ensemble techniques are one of the optimal methods of obtaining models that generate well.

References

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning

algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare, 10(3), 541. https://doi.org/10.3390/healthcare10030541

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Centers for Disease Control and Prevention. (2023). Heart disease facts. U.S. Department of Health and Human Services.

Chawla, N., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357

Harp Documentation, Harp Random Forests https://dsc-spidal.github.io/harp/docs/getting-started/Accessed: 29th of March, 2024LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436444.https://doi.org/10.1038/nature14539

Hasanova, H., Tufail, M., Baek, U., Park, J., & Kim, M. (2022). A novel blockchain-enabled heart disease prediction mechanism using machine learning. Computers and Electrical Engineering, 101, 108086. https://doi.org/10.1016/j.compeleceng.2022.108086

Liang, Y., & Guo, C. (2022). Heart failure disease prediction and stratification with temporal electronic health records data using patient representation. Biocybernetics and Biomedical Engineering, 43(1), 124-141. https://doi.org/10.1016/j.bbe.2022.12.008

Mohapatra, S., Maneesha, S., Patra, P. K., & Mohanty, S. (2022). Heart Diseases Prediction based on Stacking Classifiers Model. Procedia Computer Science, 218, 1621-1630. https://doi.org/10.1016/j.procs.2023.01.140

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3*, 100130. https://doi.org/10.1016/j.health.2022.100130

Rajendran, R., & Karthi, A. (2022). Heart disease prediction using entropy-based feature engineering and ensembling of machine learning classifiers. Expert Systems with Applications, 207, 117882. https://doi.org/10.1016/j.eswa.2022.117882

World Health Organisation. (2021), cardiovascular diseases (CVDs).