IPTEK The Journal for Technology and Science, 37(2), 2026 (e/pISSN:2088-2033, 0853-4098)
DOI: 10.12962/j20882033.v37i2.9251 | Received 9 February, 2026; Revised 30 June, 2026; Accepted 15 July, 2026

ORIGINAL RESEARCH

Accuracy Upgrade of BRFSS Dataset Subject Heart Disease Using the Hybrid Algorithm

Retno Aulia Vinarti* | Muhammad Ariiq Ramadhan | Ivan Tri Putra Hehanusa | Nethaneel Muluk | Amalia Utamima

[1]Department of Information Systems, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia, Email: zahra_17@is.its.ac.id

Correspondence

*Email: zahra_17@is.its.ac.id

Present Address

Department of Information Systems, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia

Abstract

Heart disease is a specific abnormal condition that includes heart-related and bloodstream-related illnesses. It can range from heart rhythm issues to blood vessel disorders. Medical practitioners use a patient"s medical history and tests like blood pressure, blood sugar, or cholesterol to make a diagnosis of heart disease. The Behavioral Risk Factor Surveillance System is the greatest telephone survey program in the United States for learning about Americans" health-related risk behaviors, chronic illnesses, and use of preventative drugs (BRFSS). We may use the data to identify a variety of illness-related traits and forecast an individual's likelihood of developing a specific condition. Following that, we compare and contrast the accuracy of the outcomes obtained using the methods described in the journal, earlier methods for previous projects, and hybrid algorithms. The objective of this research is to surpass the result of the journal reference which is the minimum accuracy of 78% and the maximum of 90%. Meanwhile the results of our Hybrid Algorithm methods are 90% for XGBoost+Randomize Search and 91% for the other algorithms. Based on the results, we can conclude that by combining the classification algorithm with the hyperparameter optimization method, we may enhance classification accuracy. Because the presence of hyperparameters makes the limit more optimal and optimal for classification, resulting in more accurate findings. It can be seen that the results obtained from our project are slightly higher compared to the Journal. The highest accuracy from the Journal is 90% while our result is 91%.

KEYWORDS:

accuracy, algorithms, cardiovascular disease, classification, hybrid

1 | INTRODUCTION

Heart disease is a specific abnormal condition that includes heart-related and bloodstream-related illnesses. It can range from heart rhythm issues to blood vessel disorders[1]. Medical practitioners use a patient's medical history and tests like blood pressure, blood sugar, or cholesterol to make a diagnosis of heart disease. In addition, cutting-edge medical diagnostics like the electrocardiogram (ECG), exercise stress tests, X-rays, echocardiogram, coronary angiography, radionuclide tests, MRI scans, and CT scans can help determine whether a person has a cardiac condition[2]. A timely and accurate diagnosis and treatment of this serious condition are essential for disease prevention and the effective use of medical resources, especially in light of the rising statistics. For the early diagnosis of diseases, machine learning is becoming increasingly significant. By spotting patterns concealed in observations, it focuses on drawing conclusions consistent with fresh data[3]. Particularly, machine learning (ML) has made significant strides in the field of cardiovascular medicine and has created a potential niche[4]. The fundamentals of machine learning are based on models that take input data (such as text or images) and, using statistical analysis and mathematical optimizations, produce the desired prediction outputs (such as disease, no disease, or neutral)[5]. Heart-related illnesses, for instance, are among the most lethal conditions in the world, claiming 17.9 million lives annually[6]. Many researchers from many backgrounds have been looking into the problem using this method because early detection of cardiac disease may reduce the death rate.

2 | LITERATURE REVIEW

Heart disease is a specific abnormal condition that includes heart-related and bloodstream-related illnesses. It can range from heart rhythm issues to blood vessel disorders[1]. Medical practitioners use a patient's medical history and tests like blood pressure, blood sugar, or cholesterol to make a diagnosis of heart disease. In addition, cutting-edge medical diagnostics like the electrocardiogram (ECG), exercise stress tests, X-rays, echocardiogram, coronary angiography, radionuclide tests, MRI scans, and CT scans can help determine whether a person has a cardiac condition[2]. A timely and accurate diagnosis and treatment of this serious condition are essential for disease prevention and the effective use of medical resources, especially in light of the rising statistics. For the early diagnosis of diseases, machine learning is becoming increasingly significant. By spotting patterns concealed in observations, it focuses on drawing conclusions consistent with fresh data[3]. Particularly, machine learning (ML) has made significant strides in the field of cardiovascular medicine and has created a potential niche[4]. The fundamentals of machine learning are based on models that take input data (such as text or images) and, using statistical analysis and mathematical optimizations, produce the desired prediction outputs (such as disease, no disease, or neutral)[5]. Heart-related illnesses, for instance, are among the most lethal conditions in the world, claiming 17.9 million lives annually[6]. Many researchers from many backgrounds have been looking into the problem using this method because early detection of cardiac disease may reduce the death rate. Recent studies show that heart disease prediction has increasingly moved toward machine learning and deep learning approaches to improve diagnostic accuracy, especially when working with large-scale health datasets. The BRFSS dataset is particularly relevant because it provides population-level behavioral and health-risk information that has been widely used for cardiovascular disease and stroke prevention programs[7]. More recent work using the 2022 BRFSS dataset demonstrated that XGBoost can effectively handle imbalanced cardiovascular disease data and improve predictive performance in risk classification tasks[8]. This confirms that BRFSS-based prediction remains valuable but also highlights the need for algorithmic enhancement to address common challenges such as class imbalance, heterogeneous features, and limited classification sensitivity. Current literature also indicates that hybrid and advanced models have become a dominant direction in heart disease classification. Transformer-based diagnosis tools, deep neural networks, hybrid quantum-neural networks, and CNN–GRU fusion models have been proposed to improve early detection, classification accuracy, and diagnostic objectivity[9–13]. These studies suggest that no single conventional algorithm is always sufficient for complex cardiovascular prediction tasks. Therefore, this article attempts to improve predictive accuracy by combining the strengths of multiple algorithms. The hybrid approach is expected to provide better classification performance than standalone models, particularly for BRFSS data that contains behavioral, demographic, and clinical risk-factor variables.

TABLE 1 Variable used for the study.

Name	Type	Name	Type
Heart Disease (CVDCRHD4)	Nominal	Age Category (AGEG5YR)	Nominal
BMI (BMI5)	Continuous	Sex (COLGSEX)	Nominal
Kidney Disease (CHCKDNY2)	Nominal	Walking Difficulty (DIFFWALK)	Nominal
Asthma (ASTHMA3)	Nominal	Mental Health (MENTHLTH)	Continuous
Sleep Time (SLEPTIM1)	Continuous	Physical Health (PHYSHLTH)	Continuous
General Health (GENHLTH)	Nominal	Stroke (CVDSTRK3)	Nominal
Exercise (EXERANY2)	Nominal	Alcohol Consumption (MAXDRNKS)	Nominal
Diabetes (DIABETE4)	Nominal	Smoking (SMOKE100)	Nominal
Race (IMPRACE)	Nominal	Skin Cancer (CHCSCNCR)	Nominal

3 | MATERIAL AND METHODS

This section explains data and methods used in this research article. Datasets are described and pre-processed in section 3.1. Thereafter, algorithms used for classified heart diseases are explained in section 3.2.

3.1 | Dataset Description and Pre-Processing

The Behavioral Risk Factor Surveillance System is the greatest telephone survey program in the United States for learning about Americans' health-related risk behaviors, chronic illnesses, and use of preventative drugs (BRFSS). Using the data, we may categorize numerous illness-related features and predict an individual's probability of developing a certain condition. Classification is one of the most common data mining strategies. The act of developing a model in the form of classification rules, decision trees, or mathematical formulas using a training set of data that comprises samples supposed to belong to a preset class is referred to as "classification" (a hybrid). A labeled dataset is used in the training model of the classification process to teach it the relevant attributes of the labeled objects (sequential clustering). The dependent variable in the study is heart disease. The following variables in Table 1 are chosen from conditions that can lead to heart disease. As BMI rises, so does the chance of obesity, which leads to heart disease. People who have kidney illness are more likely to have heart disease. Because asthma impacts many bodily functions, it is a condition that also affects heart disease. Lack of sleep could cause heart disease since our bodies cannot accomplish numerous functions. Individual's overall health is reflected in their general health. Physical activity is essential for the cardiovascular system. Diabetes is recognized as a risk factor for heart disease.It has been found that racial differences exist in the risk of developing heart disease. It is also known that being older raises your risk of having heart disease. Gender has an impact on heart disease. Heart disease is more common in men than in women. Seniors are sought after because elderly folks with heart problems may have difficulty walking. Heart problems can raise the risk of stroke. The risk of heart disease is increased by both alcohol use and cigarette smoking. The final factor is skin cancer, which can spread to the heart and result in heart disease[14].

3.2 | Classification Algorithms

In this section we are going to talk about the algorithms we use to classify the dataset. In this study we are combining the algorithm with two different hyperparameters namely, Random Forest + Randomized Search, Random Forest + Grid Search, XGBoost Randomized Search, and XGBoost + Grid Search.

3.2.1 | Random Forest

Random Forest is a machine learning algorithm used for classification and regression tasks.It is an ensemble approach, which means it integrates the results of different models to provide a prediction that is more precise. Decision trees serve as the models in a Random Forest.[15]. A decision tree is a flowchart-like tree structure that makes a prediction based on the values of the input features. Each leaf node in the tree represents a class label, whereas each internal node represents a "test" on an input feature. The path from the root node to the leaf node denotes a choice based on the input features. The decision trees of a Random

Forest are trained using various subsets of the training data, and the predictions from all the trees are pooled using an average or voting approach. This collection of forecasts aids in lowering overfitting and enhancing the model's general performance. The algorithm's performance can be optimized by adjusting a number of factors, such as the depth of each individual tree and the number of trees in the forest. Numerous applications, such as picture categorization, spam filtering, and credit fraud detection, have made extensive use of Random Forest. Additionally, it has been applied to biological applications including forecasting molecular activity and locating significant DNA sequence patterns.[16].

3.2.2 | XGBoost

XGBoost (eXtreme Gradient Boosting) is an open-source software library for gradient boosting, a machine learning technique for regression and classification tasks. It was developed by Tianqi Chen and provides efficient implementations of the gradient boosting algorithm. Due to its great performance and efficiency, XGBoost has grown to be a well-liked option for creating machine learning models. It has been utilized in many different applications, including as natural language processing, computer vision, and recommendation systems, and it can handle big datasets. Gradient boosting involves training a model to make forecasts based on the mistakes of earlier models. Each weak model in the process is successively trained, making them each relatively unreliable predictors on their own. The weak models' predictions are then combined to produce a strong predictor. Decision tree models are used by XGBoost as the weak learners, and regularization and other methods are used to optimize the training process. The capacity of XGBoost to parallelize the training process, which enables it to operate well on big datasets, is one of its important characteristics. Additionally, it features a variety of hyperparameters that may be adjusted to enhance the performance of the model[17].

3.2.3 | Hybrid Algorithm Theory

A hybrid algorithm is an algorithm that combines two or more different approaches or techniques in order to achieve better performance than would be possible by using any single approach alone. Hybrid algorithms are often used in machine learning and data mining to improve the accuracy, speed, or scalability of the algorithm[17]. In order to perform better than any of the component algorithms alone, a hybrid algorithm mixes two or more separate algorithms. In order to increase the effectiveness or accuracy of the generated model, hybrid algorithms are frequently utilized in machine learning and other disciplines. Combining algorithms to produce a hybrid algorithm can be done in a variety of ways. Using a meta-algorithm, which is an algorithm that accepts as input a collection of several algorithms and somehow integrates them, is one typical strategy. Meta-algorithms can be used to alter the parameters of the individual algorithms, integrate the predictions of many models, or choose the best model for a particular task. Another approach is to use a combination of different types of algorithms, such as combining a decision tree with a neural network. This can be done in a number of ways, such as using the decision tree to select the most relevant features and then training a neural network on the resulting subset of features or using the decision tree as a pre-processing step and then training the neural network on the output of the decision tree. Hybrid algorithms can also be created by combining different techniques from within a single algorithm. For example, a hybrid genetic algorithm might combine elements of both crossover and mutation in the evolution process. Overall, the goal of a hybrid algorithm is to take advantage of the strengths of different algorithms and combine them in a way that leads to improved performance[18].

3.2.4 | Randomized Search

Randomized search is a method for hyperparameter optimization in machine learning algorithms. It entails selecting a random subset of possible hyperparameter values and assessing the model's performance for each conceivable set of parameters. The objective is to identify the set of hyperparameters that results in the model performing at its peak. Randomized search is often used as an alternative to grid search, which involves evaluating the model for every possible combination of hyperparameter values within a specified range. Grid search can be computationally expensive, particularly for models with many hyperparameters, so randomized search can be a more efficient method in these cases. To use randomized search, the range of possible values for each hyperparameter is defined, and a random sample of combinations of values is drawn from this range. The model is then trained and evaluated for each combination of values, and the combination that leads to the best performance is selected. The

TABLE 2 XGBoost + Randomized Search confusion matrix.

	Precision	Recall
0 (No Risk)	0.91	0.98
1 (Risky)	0.26	0.07
Accuracy	0.90	

TABLE 3 Random Forest + Grid Search confusion matrix.

	Precision	Recall
0 (No Risk)	0.91	1.00
1 (Risky)	1.00	0.00
Accuracy	0.91	

TABLE 4 XGBoost + Grid Search confusion matrix.

	Precision	Recall
0 (No Risk)	0.91	1.00
1 (Risky)	0.42	0.00
Accuracy	0.91	

process can be repeated multiple times to improve the chances of finding the optimal combination of hyperparameters. Randomized search has been shown to be effective in a variety of machine learning applications and can be used with any type of model that has tuneable hyperparameters [18].

3.2.5 | Grid Search

Grid search is a method for hyperparameter optimization in machine learning algorithms. It involves specifying a grid of possible hyperparameter values, and then training and evaluating the model for each combination of values. The goal is to find the combination of hyperparameters that leads to the best performance of the model. Grid search involves defining the range of feasible values for each hyperparameter and creating a grid of all feasible value combinations. The value combination that results in the model performing the best is then chosen after the model has been trained and tested for each combination of values. Grid search is frequently used in conjunction with other techniques, like cross-validation, to decrease the number of model evaluations that are necessary because it can be computationally expensive, especially for models with multiple hyperparameters. Grid search has been widely used in a variety of machine learning applications and can be used with any type of model that has tuneable hyperparameters [19].

4 | RESULT AND ANALYSIS

The result of the experiment is divided into three: precision, recall, and accuracy. The closest the results to one, the better the algorithm performance. The result from XGBoost and Randomized Search can be seen in Table 2 . The result from Random Forest and Grid Search can be seen in Table 3 . The result from XGBoost and Grid Search can be seen in Table 4 . The result from Random Forest and Randomized Search can be seen in Table 5 .

TABLE 5 Random Forest + Randomized Search confusion matrix.

	Precision	Recall
0 (No Risk)	0.91	1.00
1 (Risky)	0.26	0.00
Accuracy	0.91	

TABLE 6 Accuracy result based on journal.

Algorithm	Accuracy
LR	83%
SVM	84%
NB	78%
k-NN	81%
DT	84%
AdaBoost	85%
MLP	81%
XGB	90%

TABLE 7 Result of hybrid algorithms accuracy.

Algorithm	Accuracy
XGBoost + Randomized Search	90%
Random Forest + Grid Search	91%
Random Forest + Randomized Search	91%
XGBoost + Grid Search	91%

Tabel 2 , 3 , 4 , and 5 shows the result from running the presented algorithms. As seen from those Tables, the best result we got is in the first experiment on Table 2 with the results of 90% accuracy, 91% precision on 'No Risk' and 26% on 'Risky'. For the recall we got 98% on 'No Risk' and 7% on 'Risky'. For the comparisons with the reference are shown in Table 6 and 7 . Table 7 also shows this research result using four Hybrid Algorithms.

After that, we examine and contrast the accuracy of the results obtained using the methods employed in the journal, earlier methods for midterms, and methods utilizing hybrid algorithms. The objective of this research is to surpass the result of the journal reference which we can see in Table 6 with the minimum accuracy of 78% and the maximum of 90%. Meanwhile the results of our Hybrid Algorithm methods in Table 7 are 90% for XGBoost+Randomize Search and 91% for the other algorithms. We can see that from the journal reference the highest is using XGBoost algorithm the same as our Hybrid Algorithm using XGBoost+Grid Search with the same 0.2 test size and 10-fold. From the result we can see that our final project has the highest accuracy, but the result is not very detailed because the method only uses 1 method. Figure 1 depicts the distribution of participants in the dataset with and without heart disease based on the results. The number of people with heart disease detected in the sample is 9.1%, or up to 6566 people, out of a total of 72103 people.

5 | CONCLUSION

This article aims to classify heart disease risk using several combinations of classifiers: (1) XGBoost + Randomized Search, (2) Random Forest + Grid Search, (3) Random Forest + Randomized Search, (4) XGBoost + Grid Search. Each classifier using nine

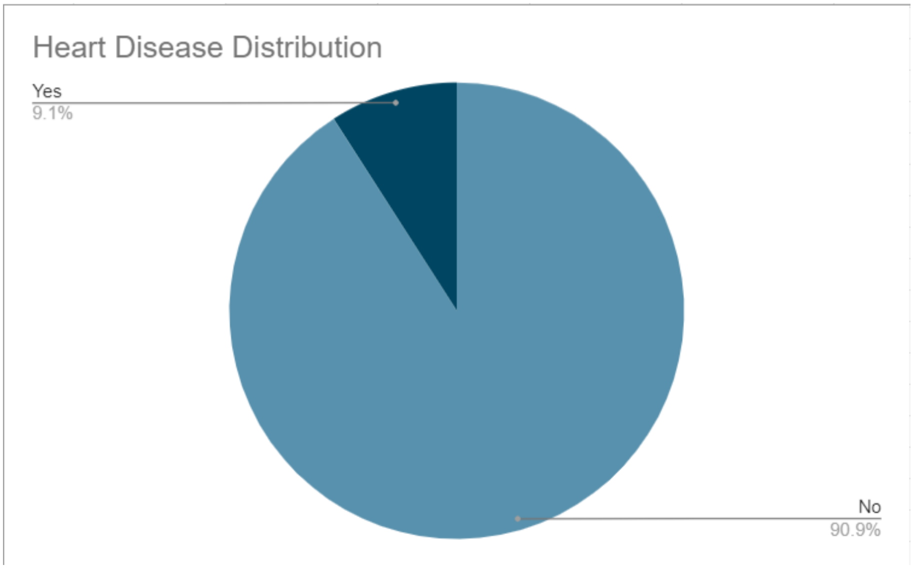

FIGURE 1 Distribution of people with and without heart disease.

variables to perform the risk clasification. The results show that all four combinations outperform other classifier algorithms (e.g., Logistic Regression, Support Vector Machine, Naive Bayes, k-Nearest Neighbor, Decision Tree, AdaBoost, Multilayer Perceptron, and XGBoost) by having 90-91% accuracy. However, from the presented result, the distribution of people having risk of heart disease distribution is only 9.1%. This may imply that the contribution of this research may not largely needed by human population.

References

1. Mayo Clinic Staff. Heart Disease. Mayo Clinic 2022;https://www.mayoclinic.org/diseasesconditions/heart-disease/symptoms-causes/syc-20353118, Accessed: Dec. 06, 2022.

2. National Heart, Lung, and Blood Institute (NHLBI). Coronary Heart Disease - Diagnosis. NHLBI Health Topics 2022;https://www.nhlbi.nih.gov/health/coronary-heart-disease/diagnosis, Accessed: Dec. 06, 2022.

3. Williams R, Shongwe T, Hasan AN, Rameshar V. Heart Disease Prediction using Machine Learning Techniques. In: 2021 International Conference on Data Analytics for Business and Industry (ICDABI) IEEE; 2021. p. 118–123.

4. Haq AU, Li JP, Memon MH, Nazir S, Sun R, Garciá-Magariño I. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems 2018;2018:1–15.

5. Gavhane A, Kokkula G, Pandya I, Devadkar K. Prediction of Heart Disease Using Machine Learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) IEEE; 2018. p. 1275–1278.

6. World Health Organization. Cardiovascular Diseases. World Health Organization Fact Sheets 2022;https://www.who.int/health-topics/cardiovascular-diseases, Accessed: Dec. 06, 2022.

7. Greenlund KJ, Denny CH, Mokdad AH, Watkins N, Croft JB, Mensah GA. Using Behavioral Risk Factor Surveillance Data for Heart Disease and Stroke Prevention Programs. American Journal of Preventive Medicine 2005;29(5):81–87.

8. Imani M, et al. Predicting cardiovascular diseases using imbalanced data: An XGBoost-based analysis of the 2022 BRFSS dataset. American Heart Journal Plus: Cardiology Research and Practice 2026;63:100719.

9. Arif S, Son SH, Kim HY, Kim SC, Lee JY. A diagnosis tool for early detection and classification of heart disease in individuals using transformer mechanisms. Computer Methods and Programs in Biomedicine 2026;277:109248.

10. Han S, Lee J, Park E. Towards objective heart disease classification: A deep neural network approach for comprehensive diagnosis. Biomedical Signal Processing and Control 2026;117:109502.

11. Amir R, et al. Association Between a novel Adult Congenital Heart Disease Patient-Reported Outcome Metric and New York Heart Association Functional Classification. CJC Pediatric and Congenital Heart Disease 2026;p. 100200.

12. Verdone A, Succetti F, Ceschini A, Rosato A, Fioravanti A, Panella M. A hybrid quantum-neural network for heart disease classification. Biomedical Signal Processing and Control 2026;113:109185.

13. Renukaradhya PC, Reddi R, Simon BK, Venkta Satyanarayana N. Multichannel CNN-based wavelet attention GRU Fusion-Net with triplet context learning for dynamic heart disease classification. Biomedical Signal Processing and Control 2026;113:109155.

14. Akkaya B, Sener E, Gursu C. A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques. In: HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings IEEE; 2022. p. 1–6.

15. Breiman L. Random Forests. Machine Learning 2001;45(1):5–32.

16. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York; 2009.

17. Wang J, editor. Encyclopedia of Data Warehousing and Mining. 2nd ed. IGI Global; 2009.

18. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 2012;13(10):281–305.

19. Džeroski S, Ženko B. Machine Learning, Ensemble Methods in. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning Boston, MA: Springer US; 2004.p. 511–517.

How to cite this article: Retno A. V., Muhammad A. R., Ivan T. P. H., Nethaneel M., Amalia U., (2026), Accuracy Upgrade of BRFSS Dataset Subject Heart Disease Using the Hybrid Algorithm, *IPTEK The Journal for Technology and Science*, *37(2):92-99*.

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)