International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 3, June 2023, pp. 3359~3366
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i3.pp3359-3366

Multivariate sample similarity measure for feature selection with a resemblance model

Tsehay Admassu Assegie[1], Ayodeji Olalekan Salau[2,5], Crescent Onyebuchi Omeje[3], Sepiribo Lucky Braide[4]
[1]Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara, Ethiopia
[2]Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria
[3]Department of Electrical and Electronics Engineering, Rivers State University, Port Harcourt, Nigeria
[4]Department of Electrical/Electronic Engineering, University of Port Harcourt, Port Harcourt, Nigeria
[5]Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

Article Info

Article history:

Received Mar 4, 2022
Revised Jul 6, 2022
Accepted Aug 18, 2022

Keywords:

Heart disease
Multivariate similarity
Random forest
University of California Irvine data repository
XGBoost

ABSTRACT

Feature selection improves the classification performance of machine learning models. It also identifies the important features and eliminates those with little significance. Furthermore, feature selection reduces the dimensionality of training and testing data points. This study proposes a feature selection method that uses a multivariate sample similarity measure. The method selects features with significant contributions using a machine-learning model. The multivariate sample similarity measure is evaluated using the University of California, Irvine heart disease dataset and compared with existing feature selection methods. The multivariate sample similarity measure is evaluated with metrics such as minimum subset selected, accuracy, F1-score, and area under the curve (AUC). The results show that the proposed method is able to diagnose chest pain, thallium scan, and major vessels scanned using X-rays with a high capability to distinguish between healthy and heart disease patients with a 99.6% accuracy.

Corresponding Author:

Ayodeji Olalekan Salau
Department of Electrical/Electronics and Computer Engineering, Afe Babalola University
Ado-Ekiti, Nigeria
Email: ayodejisalau98@gmail.com

1. INTRODUCTION

One of the machine learning (ML) classification problems is the diagnosis of chronic heart disease (CHD), in which an algorithm is trained to classify a given observation as a heart disease patient or a healthy patient [1]–[3]. The classification process begins with the use of ML to train different data samples with different rows of data and columns or features. The features (columns) are used to match the pattern between new data points in order to determine whether they are heart disease patients or healthy patients. ML has demonstrated promising results and has become one of the most widely used fields in the healthcare industry for medical diagnosis [4], [5]. Among the most common applications of ML in the healthcare industry is the medical diagnosis model. Automated medical diagnosis models play a critical role in supporting medical experts' decisions in high-precision medical diagnosis. Despite the fact that ML models have been widely studied and demonstrated significant success in decision support for medical diagnosis, CHD diagnosis is a very complicated task, and much research effort is still required to explore the existing methods for heart disease diagnosis [6]–[9]. CHD diagnosis is a problem for ML classification models. As a result, the classification algorithm is taught about the relationship between CHD features and targets.

In this study, CHD risk factors such as age, cholesterol, thulium scan, chest pain, and other results of heart disease tests are used to predict the presence of heart disease. The paper focuses on a sample similarity measure-based feature selection method for pre-processing CHD datasets before developing the predictive model using a classification algorithm. The study suggests using a sample similarity measure as a feature selection method to identify better feature combinations in order to develop a better-performing heart-disease diagnosis model. To accomplish this goal, simulation using multivariate sample similarity measures is performed on the CHD dataset obtained from the University of California Irvine (UCI) data repository. The simulation seeks to answer two research questions: What are the performance improvements of sample similarity measures for the CHD diagnosis model for the dataset under consideration, and what are the most significant CHD features?

Over the past few years, different approaches have been introduced to reduce the dimensionality of datasets with high-input space, and in addition, eliminate the less important features from the training process. Recently, the extreme boosting (XGBoost) algorithm was applied to the heart disease dataset to investigate the effect of heart disease features on the Cleveland dataset [10], [11]. The study looked into whether features like thallium scan, chest pain, and blood vessels are more important to the XGBoost model than other features. The experiment demonstrated that the XGBoost model was 90% accurate.

Haq *et al*. [12] investigated the application of various feature selection approaches on a CHD dataset. On the heart disease dataset, logistic regression, naïve Bayes, and the random forest algorithm are used to test statistical methods such as the chi-square test and principal component analysis (PCA). According to the findings of their research, the naïve Bayes algorithm achieves 85% accuracy. Similarly, chi-square is used on the Cleveland Heart Disease Dataset to optimize the performance of the logistic regression (LR) model with a feature that is more important in the LR model's training process. The experiment shows that when the model is trained using the most important feature selected using the chi-square method, it achieves 90.6% accuracy.

Wankhede *et al*. [13] and Muhammad *et al*. [14] used feature selection on a heart disease dataset to develop a linear support vector machine (SVM) model for heart disease diagnosis. The developed SVM model has an area under curve (AUC) score of 0.96 and a precision of 90.65%. To determine which features are important for the training process, feature ranking was used as a feature selection method.

Furthermore, Panda *et al*. [15] and George and Gaikwad [16] showed that feature selection [17] improves heart disease diagnosis by using ML models. The CHD diagnosis accuracy is 94.03% when using an embedded bagging feature selection method. A comparative study conducted in [18] on LR, SVM, k-nearest neighbor (*k*-NN), random forest (RF), naïve Bayes, and gradient boosting shows that RF achieves an accuracy of 92.04%, while the lowest, LR achieved an accuracy of 80.68%.

The importance of feature selection in improving the precision of machine learning models for heart disease diagnosis cannot be overstated. The effect of the feature on the diagnosis accuracy of different models such as gaussian naive Bayes (GNB), RF, LR, and extra tree classifier was investigated in [19]–[23]. The results of feature effect analysis on these models show that feature selection plays a variety of roles in improving the ML model's performance. The highest accuracy score was obtained with GNB, which is 94.92%.

The performance of various machine learning models on the heart disease dataset was examined in [24], [25]. The effectiveness of SVM, decision tree (DT), random forest, naïve Bayes, and logistic regression were compared. The results show that the SVM model outperforms other models, with an accuracy of 95%.

The review of literature presented in section 1 reveals that no previous work on the response of model optimization using multivariate sample similarity as a feature selection method has been reported. As a result, this study proposes a sample similarity-based feature selection method. The research is structured as follows: section 2 describes the methodology used for selecting features. Section 3 presents the results and discussion, and finally, the paper is concluded in section 4 with a conclusion and recommendations.

2. METHOD

To test the study of the heart disease features that contribute to the difference between samples, empirical research methodology was used. The UCI data repository was used to examine the characteristics that contribute to the similarity of the dataset samples. The following are the steps involved in conducting this research: First, the UCI machine learning repository is queried for the heart disease dataset. The collected dataset is then divided into two samples. The resemblance model is then created to determine whether or not the two samples are similar. In this case, the AUC is used to determine whether or not the samples are similar. An AUC of more than 0.5 is considered significant. Figure 1 shows the procedure for selecting features using the sample similarity measure.

The XGBoost model was developed with Python using the Jupyter Notebook integrated development environment (IDE). The dataset used is Cleveland. Cleveland is an open-source dataset that contains information about heart disease collected by the University of California, Irvine. It contains 1,025 observations and 14 variables (the first 13 are predictors and 'target', the patient's class, and the target variable). Table 1 describes all of the features of the Cleveland heart disease dataset.

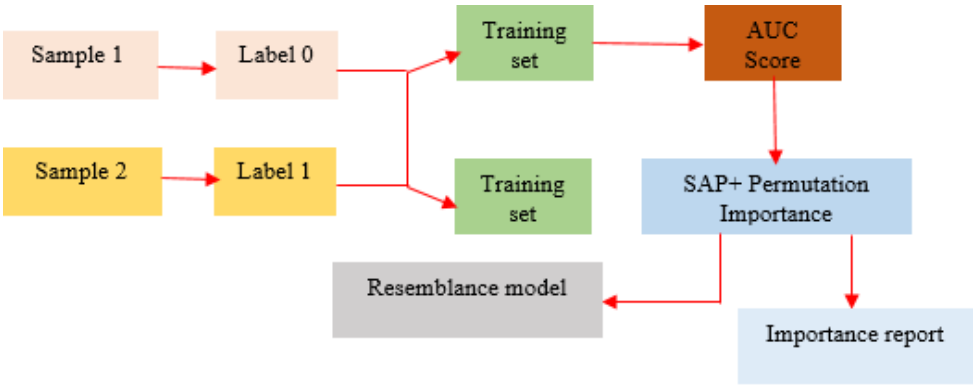

Figure 1. Feature selection using a multivariate sample similarity measure

Table 1. Chronic heart disease dataset features

	Feature
1.	Age
2.	Sex
3.	FBSfbs-fasting blood sugar
4.	Cp-chest pain type
5.	Chol-serum cholesterol
6.	Treetopsbps-resting blood pressure
7.	Restingecg-ecg at rest
8.	Thalach-maximum heart rate
9.	Exang-exercise-induced angina
10.	Ca-number of major vessels colored
11.	Slope-slope of the peak exercise ST segment
12.	Thal-defect type
13.	Oldpeak-ST depression induced by exercise
14.	Target-numeric

3. RESULTS AND DISCUSSION

This section contains a detailed discussion of the results discovered. The mean absolute Shapley additive (SHAP) values and mean SHAP values for the model's features are shown in Table 2. As shown in Table 2, mean absolute SHAP values provide information about the importance of a feature, while the mean SHAP values show the average direction in which the CHD feature influences prediction. A negative value denotes a negative class, while a positive value denotes a patient (positive) class. Furthermore, the findings highlight the potential clinical utility of the CHD classifier for CHD diagnosis.

Table 2. The mean absolute SHAP values and mean SHAP values for the model's features

Feature	Mean absolute SHAP value	Mean SHAP value
Chest pain	0.007050	-0.006214
oldpeak	0.006724	0.003684
thalach	0.004838	-0.004387
restecg	0.004801	-0.001373
trestbps	0.003306	-0.001510
age	0.003304	0.001094
exang	0.002491	0.000004
chol	0.002402	0.001618
ca	0.002182	-0.002093
slope	0.001968	0.000907
thal	0.001148	-0.000508
FBS	0.000159	-0.000011
sex	0.000126	-0.000069

Table 2 displays the sample similarity report for the CHD data sample. In addition to the SHAP value similarity report, the permutation resemblance model was developed and fitted using the two data samples. Figure 2 shows the permutation and AUC scores of the training and test sets.

The SHAP feature importance for the test set is shown in Figure 3. The figure shows that chest pain, thallium scan, and blood vessels are the top three most important features of the XGBoost model's learning process. In addition, Figure 3 shows that the train receiver operating characteristics curve has a value of 0.94, and the test AUC is 0.903.

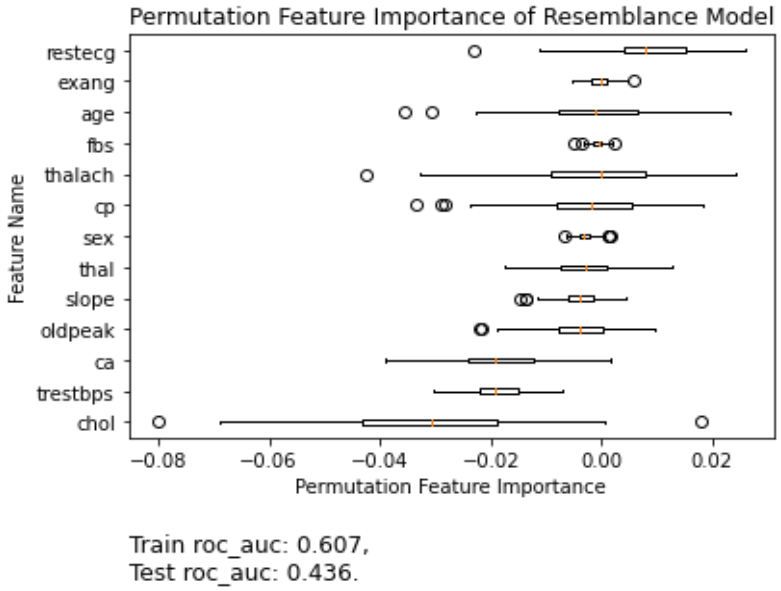

Figure 2. Permutation feature importance of sample similarity measure resemblance mode

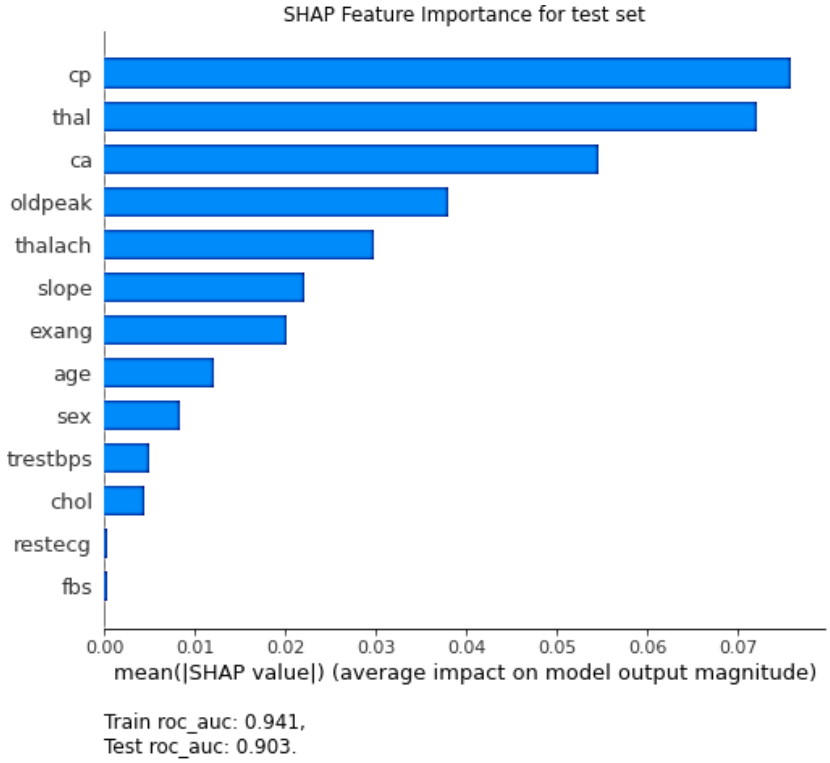

Figure 3. SHAP feature importance for the test set

Figure 4 depicts the impact of the most important feature and also shows that a patient with chest pain value 2 (a-typical angina) has a higher risk of developing heart disease than a patient with chest pain type of typical angina, non-angina pain, or symptomatic angina. Figure 5 depicts the most important feature, the thallium scan effect on the development of heart disease, and also shows that a patient with a fixed defect and a normal thallium scan value has a higher risk of developing heart disease than a patient with a reversible defect and a normal thallium scan value.

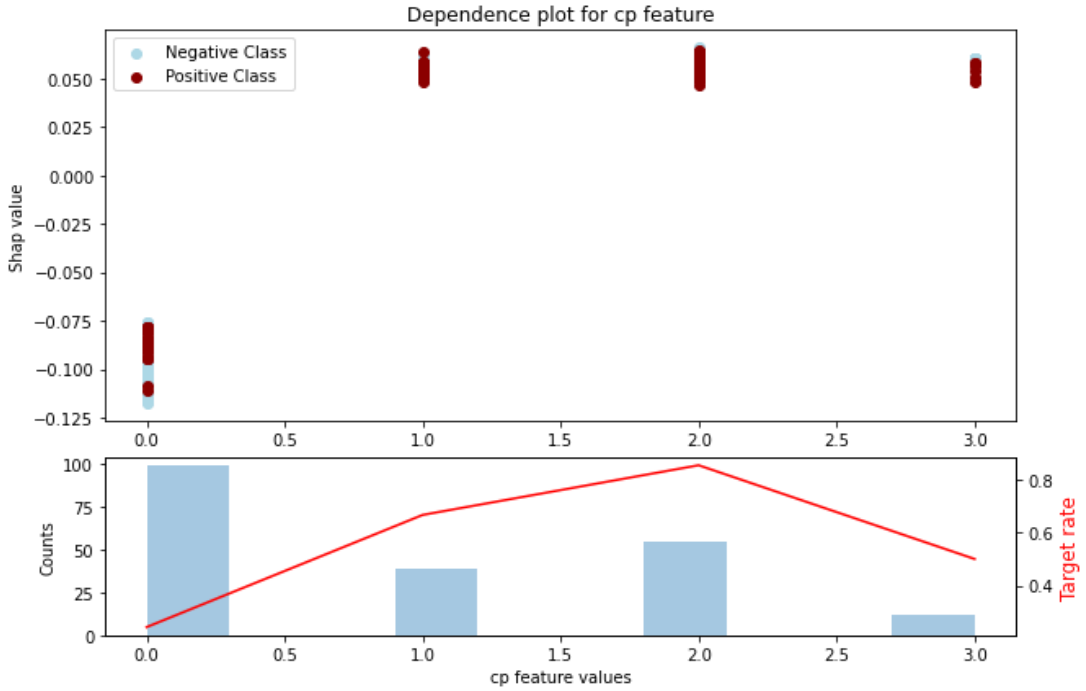

Figure 4. Dependency plot for chest pain (the topmost important feature)

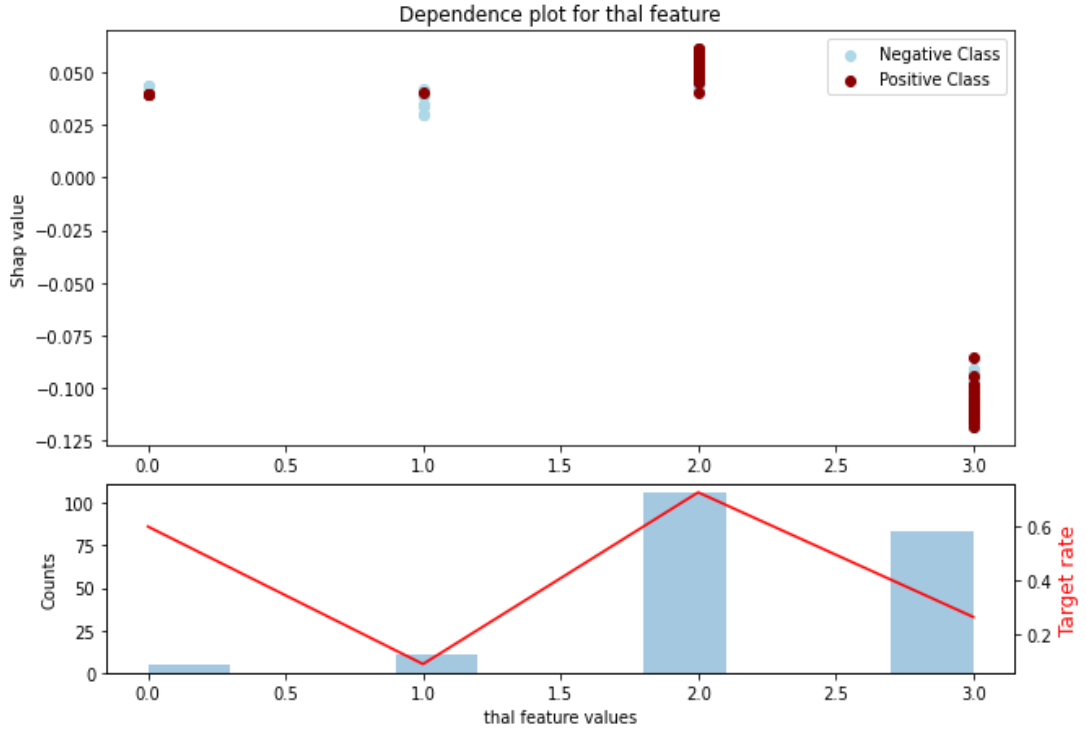

Figure 5. Dependency plot for thallium scan (the second most important feature)

3.1. Comparison with existing feature selection methods

This section presents the comparative analysis of the performance of the multivariate sample similarity measure-based feature selection method with other existing feature selection approaches. These approaches include methods such as extreme boosting and random forest feature importance. The comparative analysis of the proposed method with existing works is presented in Table 3.

Table 3. Comparison of the existing and proposed feature selection approach

Method	Dataset	Algorithm employed	Accuracy achieved (%)
[7]	Heart disease	XGBoost	90
[8]	Heart disease	Naïve Bayes	85
[9]	Heart disease	LR	90.6
[10]	Heart disease	SVM	90.65
[11]	Heart disease	*k*-NN	90
[12]	Heart disease	SVM	98.7
[13]	Heart disease	RF	92.04
[14]	Heart disease	GNB	94.92
[15]	Heart disease	SVM	95
Proposed method	Heart disease	XGBoost	99.32

4. CONCLUSION

This paper presented the investigation of a multivariate feature selection method with sample similarity measures to discriminate features that contribute to the difference between data samples. The multivariate sample similarity measures are evaluated for feature selection on the UCI dataset with the XGBoost model. The experiment reveals that the multivariate sample similarity method achieves better performance compared to the XGBoost and random forest-based feature selection. In the future, the authors recommend extending this work by exploring the effectiveness of the multivariate sample similarity in determining the optimal subset of features with different medical datasets such as diabetes and liver disease by using deep learning to further quantify the performance and the sample similarity measure for feature selection.

REFERENCES

[1] G. Saranya and A. Pravin, "An approach for optimal feature selection in machine learning using global sensitivity analysis," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 647–656, 2021, doi: 10.14569/IJACSA.2021.0120676.

[2] S. J. Sushma, T. A. Assegie, D. C. Vinutha, and S. Padmashree, "An improved feature selection approach for chronic heart disease detection," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3501–3506, Dec. 2021, doi: 10.11591/eei.v10i6.3001.

[3] T. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, "A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1831–1838, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.

[4] N. Rajinikanth and L. Pavithra, "Heart diseases prediction for optimization based feature selection and classification using machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, pp. 636–643, 2021, doi: 10.14569/IJACSA.2021.0120280.

[5] D. Khanna, R. Sahu, V. Baths, and B. Deshpande, "Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease," *International Journal of Machine Learning and Computing*, vol. 5, no. 5, pp. 414–419, Oct. 2015, doi: 10.7763/IJMLC.2015.V5.544.

[6] T. Chandrasegar and S. B. N. Vutukuri, "Optimized machine learning model using decision tree for cancer prediction," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Mar. 2019, pp. 1–4. doi: 10.1109/i-PACT44901.2019.8960129.

[7] P. Anuradha and V. K. David, "Feature selection and prediction of heart diseases using gradient boosting algorithms," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Mar. 2021, pp. 711–717. doi: 10.1109/ICAIS50930.2021.9395819.

[8] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digital Health*, vol. 6, p. 205520762091477, Jan. 2020, doi: 10.1177/2055207620914777.

[9] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.

[10] H. M. Le, T. D. Tran, and L. van Tran, "Automatic heart disease prediction using feature selection and data mining technique," *Journal of Computer Science and Cybernetics*, vol. 34, no. 1, pp. 33–48, Aug. 2018, doi: 10.15625/1813-9663/34/1/12665.

[11] J. Nourmohammadi-Khiarak, M.-R. Feizi-Derakhshi, K. Behrouzi, S. Mazaheri, Y. Zamani-Harghalani, and R. M. Tayebi, "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health and Technology*, vol. 10, no. 3, pp. 667–678, May 2020, doi: 10.1007/s12553-019-00396-3.

[12] A. U. Haq, J. Li, M. H. Memon, M. Hunain Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Mar. 2019, pp. 1–4. doi: 10.1109/I2CT45611.2019.9033683.

[13] J. Wankhede, M. Kumar, and P. Sambandam, "Efficient heart disease prediction-based on optimal feature selection using DFCSS and classification by improved Elman-SFO," *IET Systems Biology*, vol. 14, no. 6, pp. 380–390, Dec. 2020, doi: 10.1049/iet-syb.2020.0041.
[14] L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine learning predictive models for coronary artery disease," *SN Computer Science*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/s42979-021-00731-4.
[15] D. Panda, R. Ray, A. A. Abdullah, and S. R. Dash, "Predictive systems: Role of feature selection in prediction of heart disease," *Journal of Physics: Conference Series*, vol. 1372, no. 1, Nov. 2019, doi: 10.1088/1742-6596/1372/1/012074.
[16] J. P. George and S. M. Gaikwad, "Simulation modeling for heart attack patient by mapping cholesterol level," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 1, pp. 16–23, Apr. 2020, doi: 10.11591/ijeecs.v18.i1.pp16-23.
[17] S. Jain and A. O. Salau, "An image feature selection approach for dimensionality reduction based on kNN and SVM for AkT proteins," *Cogent Engineering*, vol. 6, no. 1, Jan. 2019, doi: 10.1080/23311916.2019.1599537.
[18] M. A. Tamal, M. Saiful, M. Jisan, M. Abdul, P. Miah, and K. Mohammed, "Heart disease prediction based on external factors: A machine learning approach," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 446–451, 2019, doi: 10.14569/IJACSA.2019.0101260.
[19] M. Jan, A. A. Awan, M. S. Khalid, and S. Nisar, "Ensemble approach for developing a smart heart disease prediction system using classification algorithms," *Research Reports in Clinical Cardiology*, vol. 9, pp. 33–45, Dec. 2018, doi: 10.2147/RRCC.S172035.
[20] A. I. Sapitri, S. Nurmaini, S. Sukemi, M. N. Rachmatullah, and A. Darmawahyuni, "Segmentation atrioventricular septal defect by using convolutional neural networks based on U-NET architecture," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 3, pp. 553–562, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp553-562.
[21] S. Elyassami and A. Ait Kaddour, "Implementation of an incremental deep learning model for survival prediction of cardiovascular patients," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 101–109, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp101-109.
[22] K. S. Nugroho, A. Y. Sukmadewa, A. Vidianto, and W. F. Mahmudy, "Effective predictive modelling for coronary artery diseases using support vector machine," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 345–355, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp345-355.
[23] A. A. Hussein, "Improve the performance of k-means by using genetic algorithm for classification heart attack," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 2, pp. 1256–1261, Apr. 2018, doi: 10.11591/ijece.v8i2.pp1256-1261.
[24] A. Mohd Yusof, N. A. Md. Ghani, K. A. Mohd Ghani, and K. I. Mohd Ghani, "A predictive model for prediction of heart surgery procedure," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 15, no. 3, pp. 1615–1620, Sep. 2019, doi: 10.11591/ijeecs.v15.i3.pp1615-1620.
[25] W. Wiharto, H. Kusnanto, and H. Herianto, "System diagnosis of coronary heart disease using a combination of dimensional reduction and data mining techniques: A review," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 7, no. 2, pp. 514–523, Aug. 2017, doi: 10.11591/ijeecs.v7.i2.pp514-523.

BIOGRAPHIES OF AUTHORS

Tsehay Admassu Assegie holds a Master of Science degree in Computer Science from Andhra University, India in 2016. He received his B.Sc. in computer science from Dilla University, Ethiopia in 2013. His research includes machine learning, data mining, health-formatics, network security, and software-defined networks. He has published over 38 papers in reputed international journals and international conferences. Tsehay is an active member of the International Association of Engineers (IAENG), with membership number 254711. He can be contacted at tsehayadmassu2006@gmail.com.

Ayodeji Olalekan Salau received a B.Eng. in Electrical/Computer Engineering from the Federal University of Technology, Minna, Nigeria. He received his M.Sc. and Ph.D. degrees from the Obafemi Awolowo University, Ile-Ife, Nigeria. His research interests include research in the fields of computer vision, image processing, signal processing, machine learning, control systems engineering, and power systems technology. Dr. Salau serves as a reviewer for several reputable international journals. His research has been published in many reputable international conferences, books, and major international journals. He is a registered Engineer with the Council for the Regulation of Engineering in Nigeria (COREN), a member of the International Association of Engineers (IAENG), and a recipient of the Quarterly Franklin Membership with ID number CR32878 given by the Editorial Board of London Journals Press in 2020 for top quality research output. More recently, Dr. Salau's research paper was awarded the best paper of the year 2019 in Cogent Engineering. In addition, he is the recipient of the International Research Award on New Science Inventions (NESIN) under the category of "Best Researcher Award" given by Science Father with ID number 9249, 2020. Currently, Dr. Salau works at Afe Babalola University in the Department of Electrical/Electronics and Computer Engineering. He can be contacted at ayodejisalau98@gmail.com.

Crescent Onyebuchi Omeje received his bachelor's degree in Electrical Engineering, in 2004 from the University of Nigeria, Nsukka. He also obtained his Master of Engineering (M.Eng.) and Doctor of Philosophy (Ph.D.) in 2011 and 2019 respectively in electrical engineering from the same university. He is a member of Nigeria Society of Engineers (MNSE), a registered member Council for the Regulation of Engineering in Nigeria (COREN), a member of the Institute of Electrical/Electronic Engineering (IEEE), and a full-time lecturer in the Department of Electrical/Electronic Engineering, University of Port Harcourt, Rivers State, Nigeria. He has published widely in local and international journals. His research interests include but are not limited to power electronics, new energy conversion systems, multilevel inverter applications, electric motor drives and control, and power systems modeling. He can be contacted at crescent.omeje@uniport.edu.ng.

Sepiribo Lucky Braide is a senior lecturer in the Department of Electrical Electronics Engineering, Rivers State University, Port Harcourt, Nigeria. He is the immediate past Head of the Department of Electrical Engineering from 2018 to 2021, a fellow of the Nigeria Institute of Electrical Electronics Engineers (FNIEEE), and currently the Post Graduate Coordinator in the Department, from 2021 to date. He is a member of several professional bodies/organizations among many includes Institute of Electrical Electronics Engineers (IEEE), the Nigeria Society of Engineers (MNSE), the Council for the Regulation of Engineering in Nigeria (COREN), the Nigeria Institute of Electrical Electronics Engineers (MNIEEE), International Association of Engineers (MIAENG). He had the following degrees in the electrical engineering profession which include a Bachelor of Technology (B.Tech.), Master of Technology (M.Tech), and Doctor of Philosophy (Ph.D.). Dr. Braide has attended several conferences including ICEPT, NSE, COREN, IEEE, WCECS (San Francisco, USA Oct. 2019), and WCE (London UK, July 2019). His strong pragmatic confrontation to analyses of finding solutions to challenging engineering problems made him compete in many highly reputable international journals, most recently in 2018, the International Journal of Engineering and Science Invention (IJESI) classified one of his articles as Best Paper Award - 2018 certified and titled "A Mathematical Model of Double Exponential Wave Shape (Impulse Generator) for Power Sub-station using Laplace Transform". He has received so many honors and awards both local and international. He can be contacted at braidesepiribo@yahoo.com.