

Prediction of the risk of developing heart disease using logistic regression

Ayodeji Olalekan Salau^{1,5}, Tsehay Admassu Assegie², Elisha Didam Markus³, Joy Nnenna Eneh⁴,
ThankGod Izuchukwu Ozue⁴

¹Department of Electrical and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria

²Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara, Ethiopia

³Department of Electrical, Electronic and Computer Engineering, Central University of Technology, Bloemfontein, South Africa

⁴Department of Electronic Engineering, University of Nigeria, Nsukka, Nigeria

⁵Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

Article Info

Article history:

Received Mar 4, 2023

Revised Sep 24, 2023

Accepted Oct 21, 2023

Keywords:

Automated decision making

Cardio vascular disease

Data analytics

Heart disease risk

Predictive analytics

ABSTRACT

Heart disease (HD) accounts for more deaths every year than other illnesses. World Health Organization (WHO) assessed 17.9 million life losses caused by heart disease in 2016, demonstrating 31% of all international life losses. Three-quarters of these life losses occur in low and middle-income nations. Machine learning (ML), due to advanced precision in pattern recognition and classification, demonstrates to be in effect in complementing decision-making and threat prediction from the huge number of HD data created by the healthcare sector. Thus, this study aims to develop a logistic regression model (LRM) for predicting the risk of getting HD in ten years. The study explores the different methodologies for improving the performance of base LRM for predicting whether a person gets HD after ten years or not. The result demonstrates the capability of LRM in predicting the risks of getting HD after ten years. The LRM achieves 97.35% accuracy with the recursive feature elimination and random under-sampling. This implies that the LRM can play an important role in precautionary methods to avoid the risk of HD.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Joy Nnenna Eneh

Department of Electronic Engineering, University of Nigeria

Nsukka 410001, Nigeria

Email: Nnenna.eneh@unn.edu.ng

1. INTRODUCTION

Heart disease (HD) denotes numerous kinds of disorders affecting the normal functioning of the human heart [1], [2]. The different types of HD include coronary artery, which disturbs the supply of blood to the heart, vascular HD that affects how the valves function to control blood flow, cardiomyopathies that affect heart muscles, heartbeat turbulences (arrhythmias) that affect the electrical conduction and hereditary heart illness where the heart has physical flaws that progress before birth [3].

Most HD are avoidable and modest lifestyle adjustments such as dropping tobacco use, consumption of healthily, fatness, and keeping fit and timely treatment improves their diagnoses [4], [5]. However, classifying high-risk patients is challenging because of the malfunctioning nature of several influential risk causes such as diabetes, high blood pressure, and cholesterol. Because of these limitations, scientists have turned concerning up-to-date methods of machine learning (ML) to predict the HD. The application of ML gained much research consideration in recognition of a pre-defined set of labeled data and classification, due to its proven effectiveness in assisting decision-making and risk assessment from the large quantity of data produced by the healthcare industry on HD [6]–[8].

This research aims to develop logistic regression model (LRM) for recognizing whether a patient has a 10-year chance of developing HD by employing the Framingham dataset. Various performance evaluation metrics, such as accuracy, precision, sensitivity, recall, and receiver operating characteristics (ROC), are employed to validate the LRM. The LRM was tested on the Framingham HD dataset. Moreover, the proposed model use of feature selection, and resampling have not been studied. Most of the studies focus on the early prediction of HD risks. However, preventative measures play a significant role in predicting the risks of HD. This study aims to develop a model for predicting the risks of getting heart disease in the next ten years, which helps the patient with high risks of getting HD to take preventative measure.

Machine learning-based classification algorithms have become one of the most widely researched problems in predictive analytics for preventative measures in the healthcare industry. For instance, logistic regression (LR) effectively predicts cardiovascular data given that the dataset is processed, and standardized [9], [10]. The study highlighted that the LR predicts heart disease risk with an accuracy score of 72.85%. While the study has suggested the use of data pre-processing (such as synthetic minority oversampling) and standardization as effective methods of improving the predicted power of the LR model, the obtained result is not accurate enough to predict the risk of heart disease precisely.

Similarly, several studies [11], [12] proposed standardization (min-max scaling) for improving the 63 performance of machine learning-based methods for predicting human HD. Compared to 64 the study [13], [14], the result achieved in the study was much more promising with a prediction accuracy of 96.72% using a support vector machine (SVM). The study has also proven that the performance of the machine learning method improves by an accuracy of 8.78%. Even though the study has achieved higher accuracy compared to the previous work, result 68 has still scope for improvement for more accurate prediction of HD risk. The LRM proves to have an accuracy score of 86.11% for coronary heart disease risk prediction [15], [16]. The accuracy of the LRM model improves when trained on features that are highly correlated to HD risk. Even though the proposed model was viable for the prediction of HD risks, the model has scope for future improvement. Additionally, research articles [17]–[19] have developed different machine learning systems for HD risk prediction. The HD risk prediction system was developed by employing boosting, and support vector machines for HD risk prediction [20], [21]. The boosting and SVM has been validated on the test set showing 99.75% accuracy. Although the boosting SVM has shown the highest accuracy in the literature, the model is not tested with other performance measures such as the receiver operating characteristic curve for testing the viability of the model's performance for real-time use.

The experimental result conducted in different studies [22]–[26] suggests that the performance of the machine learning method improves with feature selection, and preprocessing. During pre-processing, the missing values are replaced or removed, and the class distribution of the dataset is examined. Furthermore, the most significant features are selected, and the others are removed in the training phases. Some studies have employed ensemble methods, which combine multiple basic learning algorithms to improve HD risk prediction accuracy. However, the performance of ensemble methods can further be improved with feature selection, and by applying other data pre-processing techniques such as resampling the original dataset.

From the findings of the review of literature, the researchers have found the following research gaps to be addressed by this study. The following research gaps were identified in the literature survey: i) The result achieved by the previous study has scope for improvement. Most of the studies applied accuracy for HD risk prediction, which is the major flaw in the literature, as accuracy cannot measure the effectiveness of the model on imbalanced datasets such as the Framingham HD risks dataset; ii) While it is shown that feature selection, and resampling as effective methods for improving the performance of machine learning models for HD risk prediction, simultaneously methods use of feature selection, and resampling have not been studied; and iii) Most of the studies focus on the early prediction of HD risks. However, preventative measures play a significant role in predicting the risks of HD. This study aims to develop a model for predicting the risks of getting heart disease in the next ten years, which helps the patient with high risks of getting HD to take preventative measure. The organization of the study is as follows: section 2 discusses the methodology, section 3 presents the results and discussion, and section 4 presents the conclusion and recommendation for further work.

2. METHOD

The HD dataset is available on the Kaggle repository, containing a continuing heart study on residents of Framingham, Massachusetts. The goal of the prediction is to recognize if a patient has a 10-year risk of future HD. The dataset includes over 4,133 records and 15 features. Each characteristic represents a potential HD risk problem. Risk factors include demographic, behavioral, and homoeopathic factors. The Framingham HD dataset containing different HD risk indicator variables such as current medical conditions total cholesterol level, systolic and diastolic blood pressure, body mass index, heart rate and glucose level. In

addition to the current medical condition, medical history such as hypertension, diabetes, and blood pressure medication. The HD dataset also contains behavioral variables such smoking status, and number of cigarettes smoked per day. Some demographic information such as age and sex are included in the dataset.

The steps used to build the LRM for predicting heart disease in 10 years are discussed as follows. To begin, the HD dataset is acquired from the Kaggle repository. Secondly, the dataset is analyzed for missing values, the number of distributions in each class. Thirdly, the dataset is pre-processed with under-sampling, and recursive feature elimination (RFE) to improve the benchmark LRM on heart disease risk prediction. Figure 1 presents the flowchart of the proposed model.

The number of features associated with the risks of getting HD disease is indicated in Figure 2. As indicated in Figure 2, age is a highly correlated feature to the risks of getting HD in the next ten years. The other highly important features to the prediction of getting HD in the next ten years are the prevalence of hypertension, diastolic blood pressure, and diabetes.

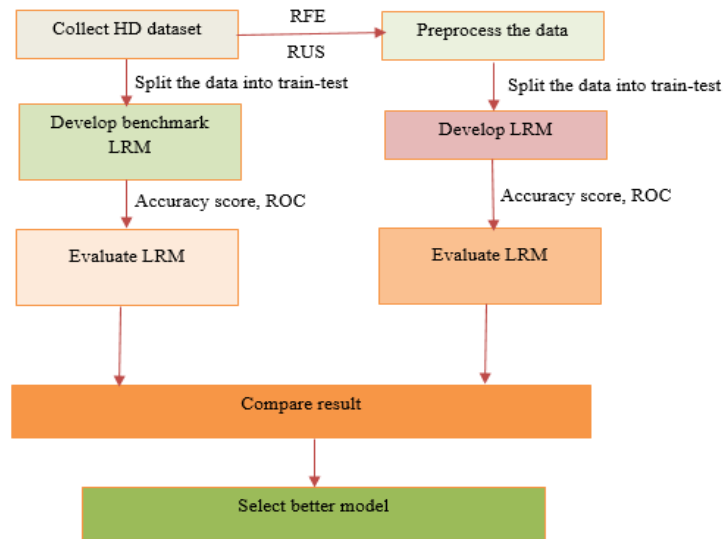


Figure 1. Proposed methodology

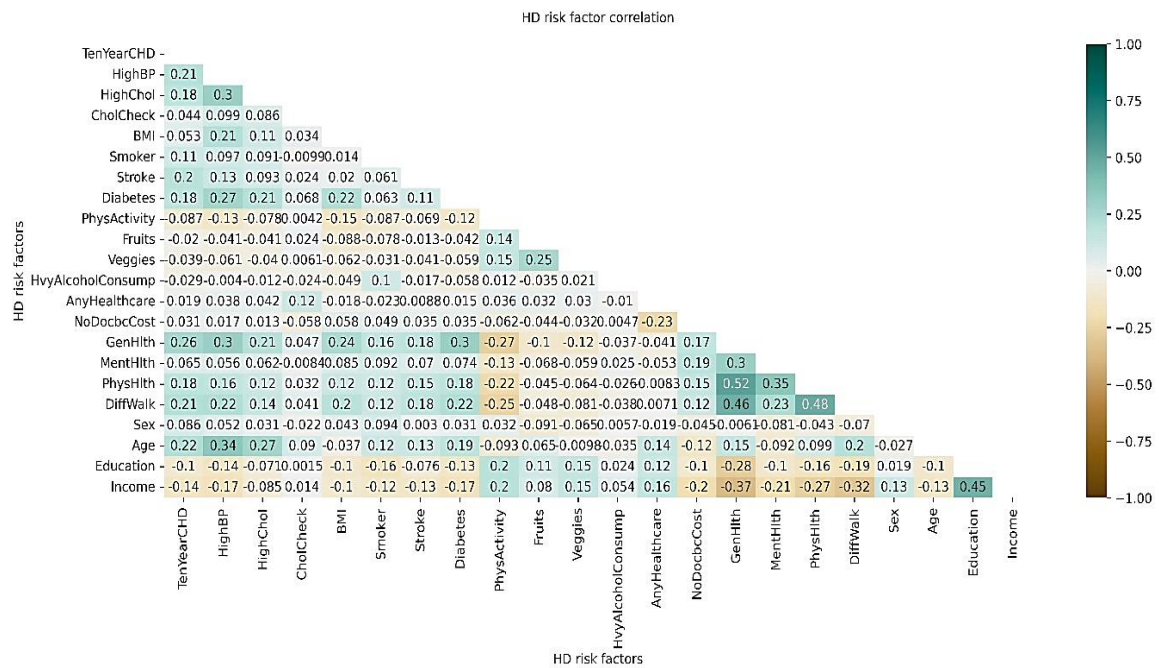


Figure 2. HD feature correlation

3. RESULTS AND DISCUSSION

The ROC curve scores of 7-fold stratified cross-validation on the dataset obtained by the LRM indicate that the LRM outperforms the under-sampled dataset compared to the RFE, and the benchmark dataset. The results indicate 85.13%, 87.85%, and 87.68% on the benchmark, under-sampled, and RFE respectively. However, the model achieves high accuracy on the benchmark imbalanced dataset; the distribution of the positive and negative classes is unequal. This study applied under sampling for balancing the class destitution of positive and negative classes. Thus, the under-sampling method achieves more ROC, and accuracy scores compared to the benchmark and RFE. Moreover, the results show improvements in the accuracy compared to previous research outcomes on a similar dataset.

3.1. Result of the benchmark dataset

The performance of the proposed LRM on predicting the risk of getting HD after ten years has been tested on the benchmark Framingham dataset. Figure 3 indicates the 7-fold cross-validation receiver operating characteristics curve of the LRM on the benchmark dataset. The LRM model achieves higher ROC value with the 6-fold Figure 3 indicates.

3.2. Result of the balanced dataset

In addition to the ROC analysis presented in section A, the LRM model is evaluated on the under-sampled dataset. To resample the majority class, random under-sampling techniques are applied to the original benchmark dataset. The resampling does not show improvement in terms of classification accuracy. However, the ROC curve of the model has shown significant improvement. Thus, resampling and other data pre-processing such as the scaling dataset feature substantially improve the performance of LRM for predicting the HD risk of a patient. Figure 4 indicates the 7-fold cross-validation ROC curve of the LRM for predicting the risk of getting HD after ten years.

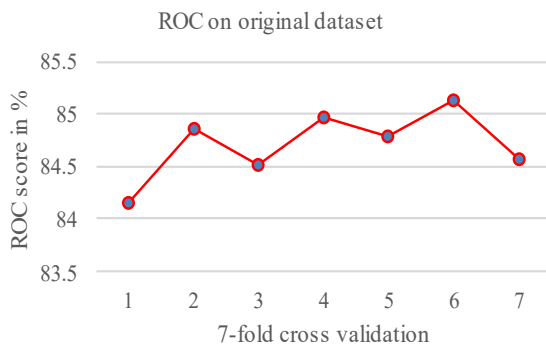


Figure 3. ROC on original dataset

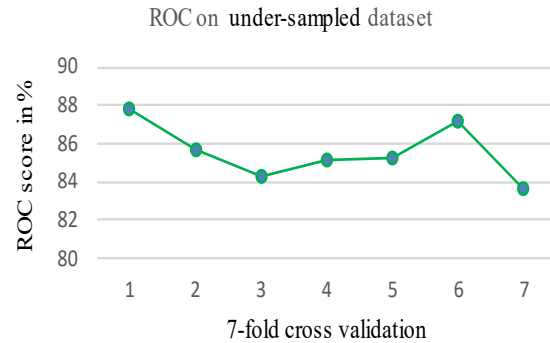


Figure 4. ROC on under-sampled dataset

3.3. Result of the RFE dataset

The top ten HD risk features or factors are selected by employing the RFE method. The LRM is then trained on the selected feature subset by the RFE. The top ten features for getting HD in ten years are demonstrated in Table 1. The 7-fold cross-validation receiver operating characteristics curve is used for performance measures to evaluate the LRM effectiveness for predicting the risk of getting HD in the next ten years. Figure 5 indicates the ROC for different folds of the LRM on ten years of HD risk prediction.

Table 1. Top nine heart disease risk

Feature	Rank
Age	1
Previous medication	2
Stroke	3
Diabetes	4
High blood pressure	5
Sex	6
Income	7
High cholesterol	8
Mental health	9

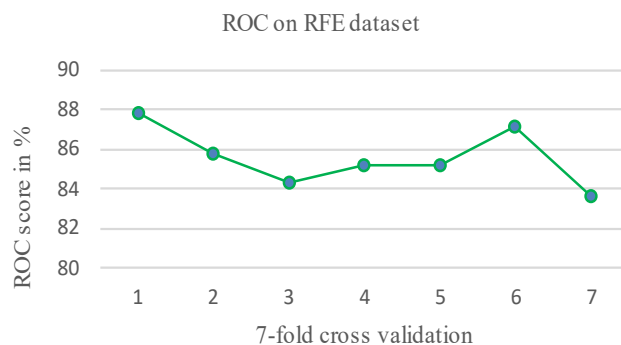


Figure 5. ROC of the model with RFE

4. CONCLUSION AND RECOMMENDATION FOR FUTURE WORK

This study proposed LRM for the recognition of HD in the next ten years. The study applied different techniques of data pre-processing to increase the performance of the proposed LRM. The RFE techniques showed and confirmed that clinical features and risk factors such as age, previous medication of HD, Stroke, Diabetes, and high blood pressure are among the most important features that help in the prediction of the presence of HD risk from medical records. Cardiologists can take advantage of the investigative data analysis conducted on the dataset to show correlations and relationships between patients' data.

HD is a foremost health concern of the world, and accurate prediction of the risk of developing HD can aid in preventive measures and personalized care. This paper explored the application of logistic regression for predicting the risk of developing heart disease in ten years. The performance of LRM depends on the nature of the data representing the HD under consideration. Even though the collected HD dataset has, the risk factors for predicting HD, with a set of features, under-sampling, and features selection with RFE potentially improving the prediction results of LRM. In future work, we plan to apply other machine-learning approaches to different HD datasets. We also plan to deploy the model with user interfaces application to allow medical experts to validate the model for real-world application.




REFERENCES

- [1] M. V. Dogan, I. M. Grumbach, J. J. Michaelson, and R. A. Philibert, "Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham heart study," *PLOS ONE*, vol. 13, no. 1, Jan. 2018, doi: 10.1371/journal.pone.0190549.
- [2] M. Ordikhani, M. S. Abadeh, C. Prugger, R. Hassannejad, N. Mohammadifard, and N. Sarrafzadegan, "An evolutionary machine learning algorithm for cardiovascular disease risk prediction," *PLoS ONE*, vol. 17, Jul. 2022, doi: 10.1371/journal.pone.0271723.
- [3] H. Byeon, "Developing a model to predict the occurrence of the cardio-cerebrovascular disease for the Korean elderly using the random forests algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 494–499, 2018, doi: 10.14569/ijacsa.2018.090962.
- [4] M. Bergamini *et al.*, "Mapping risk of ischemic heart disease using machine learning in a Brazilian state," *PLoS ONE*, vol. 15, Dec. 2020, doi: 10.1371/journal.pone.0243558.
- [5] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2022, pp. 1–9, Mar. 2022, doi: 10.1155/2022/1410169.
- [6] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *Journal of Healthcare Engineering*, vol. 2017, pp. 1–13, 2017, doi: 10.1155/2017/2780501.
- [7] E. Owusu, P. Boakye-Sekyerehene, J. K. Appati, and J. Y. Ludu, "Computer-aided diagnostics of heart disease risk prediction using boosting support vector machine," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–12, Dec. 2021, doi: 10.1155/2021/3152618.
- [8] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," *Applied Sciences*, vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188352.
- [9] 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, pp. 1–14, May 2022, doi: 10.1155/2022/6517716.
- [10] K. Kumar *et al.*, "Identification of cardiac patients based on the medical conditions using machine learning models," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–15, Jul. 2022, doi: 10.1155/2022/5882144.
- [11] B. S. Shukur and M. M. Mijwil, "Involving machine learning techniques in heart disease diagnosis: a performance analysis," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 2177–2185, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2177-2185.
- [12] O. Sami, Y. Elsheikh, and F. Almasalha, "The role of data pre-processing techniques in improving machine learning accuracy for predicting coronary heart disease," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 816–824, 2021, doi: 10.14569/IJACSA.2021.0120695.




- [13] H. Khedair and N. M. Dasari, "Exploring machine learning techniques for coronary heart disease prediction," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 28–36, 2021, doi: 10.14569/IJACSA.2021.0120505.
- [14] H. Salah and S. Srinivas, "Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents," *Scientific Reports*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-25933-5.
- [15] T. A. Assegie, A. O. Salau, C. O. Omeje, and S. L. Braide, "Multivariate sample similarity measure for feature selection with a resemblance model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 3, pp. 3359–3366, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3359-3366.
- [16] V. Chang, V. R. Bhavani, A. Q. Xu, and M. A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100016.
- [17] W. Li, M. Zuo, H. Zhao, Q. Xu, and D. Chen, "Prediction of coronary heart disease based on combined reinforcement multitask progressive time-series networks," *Methods*, vol. 198, pp. 96–106, Feb. 2022, doi: 10.1016/j.ymeth.2021.12.009.
- [18] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, 2020, doi: 10.1016/j.imu.2020.100402.
- [19] G. Saranya and A. Pravin, "A comprehensive study on disease risk predictions in machine learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4217–4225, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4217-4225.
- [20] S. Krishnan, P. Magalingam, and R. Ibrahim, "Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5467–5476, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5467-5476.
- [21] S. Molla *et al.*, "A predictive analysis framework of heart disease using machine learning approaches," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 5, pp. 2705–2716, Oct. 2022, doi: 10.11591/eei.v11i5.3942.
- [22] J. Al Nahian, A. K. M. Masum, S. Abujar, and M. J. Mia, "Common human diseases prediction using machine learning based on survey data," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 6, pp. 3498–3508, Dec. 2022, doi: 10.11591/eei.v11i6.3405.
- [23] T. A. Assegie, "Evaluation of the Shapley additive explanation technique for ensemble learning methods," in *Proceedings of Engineering and Technology Innovation*, Apr. 2022, vol. 21, pp. 20–26, doi: 10.46604/peti.2022.9025.
- [24] W. Wiharto, E. Suryani, and V. Cahyawati, "The methods of duo output neural network ensemble for prediction of coronary heart disease," *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, vol. 7, no. 1, Mar. 2019, doi: 10.52549/ijeii.v7i1.458.
- [25] S. Kutiami, R. Millham, A. F. Adekoya, M. Tettey, B. A. Weyori, and P. Appiahene, "Application of machine learning algorithms in coronary heart disease: a systematic literature review and meta-analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 153–164, 2022, doi: 10.14569/IJACSA.2022.0130620.
- [26] T. A. Assegie, A. O. Salau, C. O. Omeje, and S. L. Braide, "Multivariate sample similarity measure for feature selection with a resemblance model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 3, pp. 3359–3366, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3359-3366.

BIOGRAPHIES OF AUTHORS






Ayodeji Olalekan Salau    received the B.Eng. in electrical/computer engineering from the Federal University of Technology, Minna, Nigeria. He received the 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 by email using: ayodejisalau98@gmail.com.






Tsehay Admassu Assegie    holds a master of science degree in computer science from Andhra University, India 2016. He received his B.Sc., in computer science from Dilla University, Ethiopia in 2013. His research includes machine learning, data mining, health informatics, network security, and software-defined network. He has published over 44 papers in reputed international journals and international conferences. Tsehay is an active member of the International Association of Engineers (IAENG), with membership number: 254711. Tsehay is an active reviewer of different MDPI journals. He has reviewed many research articles in MDPI, IEEE Access, and other reputed international journals verified by the Web of Science. He can be contacted at email: tsehayadmassu2006@gmail.com.






Elisha Didam Markus    received the B.Eng. and M.Sc. degrees in electrical and electronics engineering from Abubakar Tafawa Balewa University: Bauchi, Nigeria and Ph.D. degree in electrical engineering from Tshwane University of Technology, Pretoria, South Africa. He is currently an associate professor in the Department of Electrical, Electronic and Computer Engineering, Central University of Technology, Free State South Africa. Dr. Markus research interest are in the fields of control engineering, robotics, artificial intelligence, IoTs, smart networks, and assistive mobility. Dr. Markus has a record of over 10 years of research and published over 100 papers in reputable journals, conferences and books. He can be contacted by email using: emarkus@cut.ac.za.



Joy Nnenna Eneh    received a Ph.D. in electronic and computer engineering from Nnamdi Azikiwe University Awka and also a master's degree and a bachelor's degree in electrical/electronic engineering from the Enugu State University of Science and Technology. She is a Senior Academia and a researcher in the Department of Electronic and Computer Engineering University of Nigeria, Nsukka. She is also on parallel appointment with the African Center of Excellence for Sustainable Power and Energy Development (ACESPED) UNN. Engr. Dr. Joy Eneh has numerous publications in reputable international journals and conferences. She is a member of several professional bodies and has held many leadership positions in those organizations. She is a registered engineer with the Council for Regulation of Engineering in Nigeria (COREN). She is a member of the Nigerian Society of Engineers. She is a past chairman of the Association of Professional Women Engineers of Nigeria Enugu Chapter (APWEN) (2014-2016). She is the current Chairman of the Nigerian Institute of Electrical Electronic Engineers (NIEEE) Enugu Chapter. Her research interests are in the areas of control systems, model predictive control, optimal control applications in industrial process, power systems and renewable energy, artificial intelligence applications and robotics, signal and image processing among others. She can be contacted by email using: nnenna.eneh@unn.edu.ng.



ThankGod Izuchukwu Ozue    obtained B.Eng. in electrical electronic and telecommunication engineering and M.Eng. in electronic and computer engineering from Nnamdi Azikiwe University Awka Nigeria. He is lecturer and researcher within University of Nigeria and has been collaborating with indigenous Tech. start-ups. His current research interests include cyber physical systems, microgrids, IoT embedded control systems, and cybersecurity. He can be contacted by email using: izuchukwu.ozue@unn.edu.ng.