TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 22, No. 3, June 2024, pp. 673~680
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v22i3.25560

Consistency, local stability, and approximation of Shapash explanation

Tsehay Admassu Assegie[1], Bommy Manivannan[2], Komal Kumar Napa[3], Bindu Kolappa Pillai Vijayammal[4], Rajkumar Govindarajan[3], Sangeetha Murugan[2], Atinkut Molla Mekonnen[5]
[1]School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Republic of Korea
[2]Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle, India
[3]Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, India
[4]Department of Science & Humanities (General Engineering Division), R.M.K. College of Engineering and Technology, Puduvoyal, India
[5]Department of Information Technology, College of Engineering and Technology, Injibara University, Injibara, Ethiopia

Article Info

Article history:

Received Jul 25, 2023
Revised Jan 27, 2024
Accepted Feb 2, 2024

Keywords:

Explainable artificial intelligence
Interpretable artificial intelligence
Interpretable machine learning
Model explanation
Predictive analytics

ABSTRACT

Consistency, scalability, and local stability properties ensure that a model or method produces reliable and predictable outcomes. The Shapash helps users understand how the model makes its decisions. With machine learning (ML) system, healthcare experts can identify individuals at higher risk and implement interventions to reduce the occurrence and severity of disease. ML had achieved higher prediction accuracy even though the accuracy of their prediction depends on the quality and quantity of the data used for training. Despite the wider application and higher accuracy of different ML for disease prediction, the explanation of their predictive outcome is much more important to the healthcare professional, the patient, and even their developers. However, most of the ML systems do not explain their outcomes. To address the explainability issue various techniques such as local model agnostic explanation (LIME), and shapley additive explanation (SHAP) have been proposed over the recent years. Furthermore, the consistency, local stability, and approximation of the explanation remained one of the research topics in ML. This study investigated the consistency, stability, and approximation of LIME and SHAP in predicting heart disease (HD). The result suggested that LIME and SHAP generated a similar explanation (distance=0.35), compared to the active coalition of variable (ACV) explanation (distance=0.43).

Corresponding Author:

Tsehay Admassu Assegie
School of Electronic and Electrical Engineering, Kyungpook National University
Daegu, Republic of Korea
Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

Heart disease (HD), or cardiac disease, is a term used to designate a variety of situations that affect the heart and blood vessels [1], [2]. HD comprises coronary artery disease, heart failure, arrhythmias, and valvular HD. Recently, machine learning (ML) algorithms have become one of the prominent components of healthcare in aiding medical decision-making [3]-[5]. Despite their effectiveness, ML algorithms, these algorithms do not explain their predictive result or outcome. To address the transparency issues of these algorithms, model explanation methods have been developed over the last few years to generate explanations for the predicted outcomes [6]. The explainability and interpretability of the ML model increase trust and produce explainable results.

Clinical decision-making systems should provide an explanation of their results for better adaptability and development for practical use in the medical domain [7], [8]. The explanation of the result makes the model's decision-making process clear to the medical practitioners and the patient. To that end, numerous model explanation (local model agnostic explanation (LIME) and shapley additive explanation (SHAP)) methods have been developed to make the decision-making process of the ML model transparent to healthcare practitioners and the patient [9], [10]. However, the result of the LIME and SHAP explanation techniques requires consistency, stability, and approximation quality to the development of a trustable and highly confidential ML model for particle use in HD risk prediction.

Over the past few years, ML has gained much research attention in the prediction of HD risk from various risk factors such as chest pain, age, and hypertension. Mohseni and Zarei [11] developed a model for predicting HD by employing different ML algorithms such as K-nearest neighbor (KNN), random forest (RF), logistic regression (LR), Naïve Bayes (NB), gradient, and adaptive boosting. The experimental result suggested that with grid searching techniques and preprocessing methods such as feature scaling, the developed ML model scored an accuracy of 95% for HD risk prediction.

While the ML models have achieved promising results in terms of HD prediction accuracy, explainability of the predicted result is a relatively new research area that needs further research for high-performance models to implement in healthcare analytics [12], [13]. One of the research topics in the explainability of predictive outcomes is the consistency of the explanation result for similar input features. In this study, consistency refers to the similarity between the explanation generated by different model explanation methods such as LIME, SHAP, and active coalition of variable (ACV) to the same model prediction outcome.

Furthermore, the research article [14]-[16] suggested that a summary plot illustrating the contribution of HD risk factors to model output provides an interpretation of the ML decision-making process. The study further investigated that the explanation generated by SHAP with the help of a feature contribution plot helps medical practitioners and patients to understand why the model has reached a particular diagnostic result. The simulation results highlighted that the ML model assists in decision-making achieving an accuracy is 78.81% for HD risk prediction.

The use of the ML models in the detection of HD risk has gained much research work. A research article [17], [18] introduced a LR model for detecting HD. The study suggested that ML models such as RF, LR, and KNN can be effectively used with higher precision to detect HD risks. However, these higher precision model developed for detecting HD do not provide an explanation and understandability for their decision and prediction outcome.

Despite the impressive application of ML systems in HD prediction and diagnosis, several challenges exist in the practicality of ML systems, for instance, automated HD risk prediction [19], [20]. The higher performance of the ML system obtained from its internal confidence score is not trusted and model introspection methods (using simpler models) do not help to achieve higher predictive performance (reliability-explainability trade-off). Thus, building ML systems for practical cases requires either incorporating the explanation component into the existing complex ML systems or developing post hoc algorithms that could generate an explanation for their prediction outcome.

While the impact of high-precision predictive systems has attracted much of the research attention, the development of a predictive model that provides transparent and understandable explanations and interpretations for the patient prediction outcome has also been studied in various research papers [21], [22]. The explanation generated by the existing methods of model explanation (LIME, and SHAP) to the prediction result of the ML model in HD. LIME and SHAP have demonstrated these methods provide more insight into how ML models such as XGBoost reach a certain decision while predicting HD [23]-[25].

Driven by the success of the ML systems in healthcare (prediction of patient outcome and diagnosis), significant efforts exist to exploit ML systems to analyze the HD dataset. However, understanding why ML systems have reached a certain prediction outcome is crucial, since it is the understanding that provides the confidence to decide on clinical intervention to care for the patient. Thus, this study aims to assess the degree of confidence in explainability methods with consistency, local stability, and approximation metrics. Overall, this study aims to explore the answers to the following research questions: i) What is the average consistency of different model explainability methods for the HD dataset?; ii) What are the HD features that drive the RF repressor model on positive and negative patient prediction outcomes?; and iii) How to build confidence in the model explainability method? This study aimed to investigate the consistency, scalability, and approximation of the explanations provided by the SHAP, and LIME in explaining the predictive outcomes of RF regression. Overall, the contributions of this work are outlined as follows: i) to the best of our knowledge, no existing work investigated the consistency, stability, and approximation of RF regressor model explanation highlighting its applications and their importance for HD risk prediction, ii) to explore the consistency among LIME, and SHAP explanation methods on the HD dataset collected from the UCI data repository, iii) to study the local stability of LIME, and SHAP by investigating the similarity of explanation provided by LIME, and SHAP for similar instances of the HD dataset, and iv) to examine the approximation of the explanation by exploring the influence of HD dataset features, on the predictive outcomes of the RF regressor.

The rest of this work is arranged as follows. Section 2, provides the background of different ML algorithms employed to predict HD risk. It also discusses the method. and the results achieved by this study by comparing the various explanation methods. Section 3, presents the summary of the findings and implications as well as the recommendation for future work

2. METHOD

In the investigation of the consistency, and local stability of Shapash explanation of the RF regressor prediction outcome, this study suggests the use of a random forest repressor (RFR). To explain the prediction outcome of RFR, the study employed LIME, SHAP, and ACV. Figure 1 highlights the study's general method. In the evaluation of the explanation generated by these methods, the study used consistency, stability, and approximation.

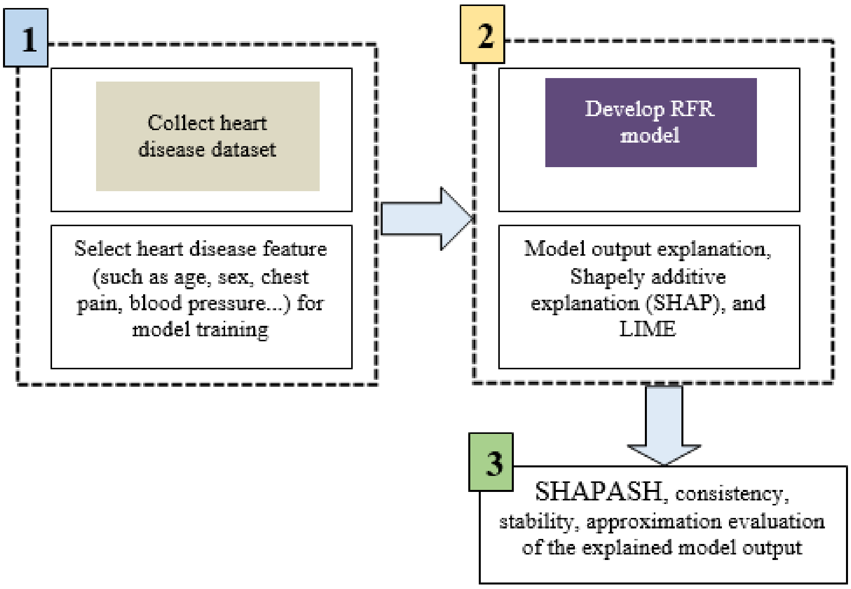
1
Collect heart disease dataset
Select heart disease feature (such as age, sex, chest pain, blood pressure...) for model training
2
Develop RFR model
Model output explanation, Shapely additive explanation (SHAP), and LIME
3
SHAPASH, consistency, stability, approximation evaluation of the explained model output

Figure 1. The schematic diagram of the procedures for the study

The three metrics employed for comparing the explainability generated by the LIME, SHAP, and ACV explanation techniques. These metrics include consistency, local stability, and approximation. The consistency metric compares how close the explanations are to each other. To measure the consistency, the Euclidean distance between the generated explanations is measured; the smaller distance leads to the assumption that the explanation provides similar results. The consistency of the explanation is calculated with the formula given in (1).

$$Consistency = Dis\,t(x, y) = \sqrt{\frac{\sum_{i=1}^{N} xi-yi)}{N}} \quad (1)$$

Where X and Y denote the explanations generated by the explanation method, I denotes the number of the model's explainability methods, and n denotes the number of explanations produced.

To build confidence upon the explanations provided by explainability methods, their local stability is significant as it shows whether the generated explanations are similar for similar samples or not. Local stability is a significant factor in building trust in the explanation because, for similar instances, the explanations are expected to be similar. Thus, the model explanation that generates a similar explanation for a similar instance is trusted compared to the one that generates a different explanation for a given instance. The third important metric for building confidence in the generated explanation is approximation. The approximation metric tests the impact of features on the model's prediction outcome.

2.1. Consistency

The consistency metric compares how close the explanations are to each other. It evaluates the similarity of explanations generated from different explainability methods. The similarity between the explanations is determined based on the average distance between the generated explanations. Figure 2 shows the consistency among LIME, SHAP, and ACV.

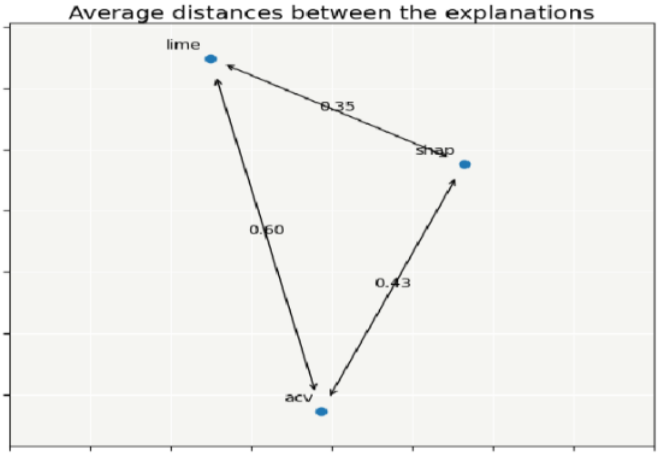

Figure 2. Consistency of explainability methods

The consistency metric demonstrated in Figure 2 shows the similarity between explanations generated by different explainability methods. The consistency is determined based on the average distance between the generated explanations by various explainability methods. The explanation between LIME, and SHAP generated a similar explanation (distance=0.35), compared to the ACV explanation (distance=0.43). In conclusion, for this particular sample, SHAP, and LIME are more similar than ACV. Moreover, Figure 3 demonstrates the consistency of the explainability methods pairwise plot for the explanation generated by different explanation methods.

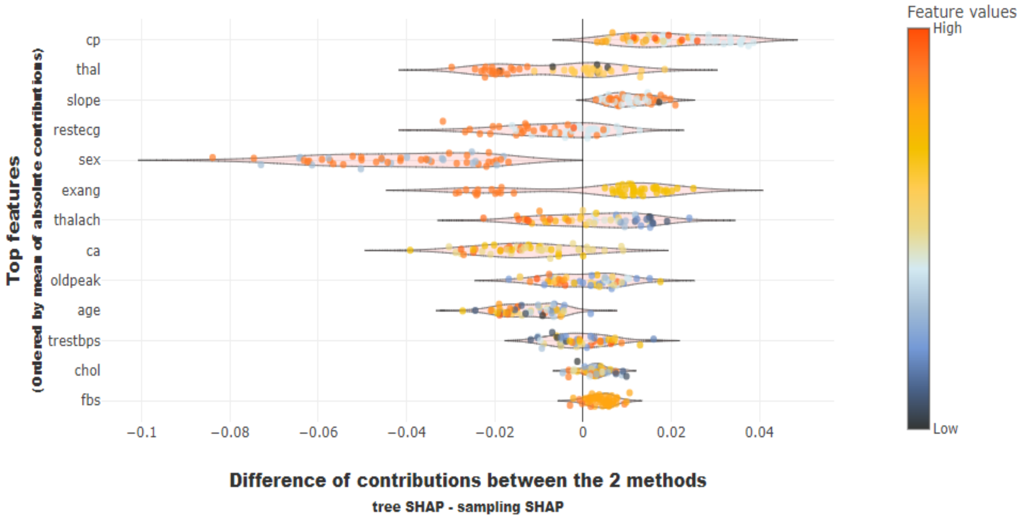

Figure 3. The praise-wise comparison of consistency between tree and sampling SHAP

The plot takes as input two explainability methods and outputs the difference of the contributions for each feature across the HD dataset or a sample. The HD features according to the mean of absolute contributions are displayed, ranked from the most important to the least. The position on the x-axis shows how different the contributions are in each direction: points centered on zero indicate little to no difference between the explainability methods, as opposed to points far away. The color bar represents the feature values. Based on that, it is possible to understand if differences between methods have a recurring pattern, helping to identify groups of data points with similar contributions. For example, looking at the sex feature, SHAP seems to constantly overestimate contributions for males concerning lime; which is not necessarily the case for the chest pain feature.

Table 1 indicates the values of each HD feature, the explanation generated by tree SHAP on feature contribution, and the difference between sampling and tree SHAP explanations. As indicated in Table 1, the tree and sampling SHAP generated contribution of the 13 HD features have differences varying between 0.00 and 0.01. Figure 4 indicates the distance between multiple explainability methods across all the HD features. As revealed in Figure 4, sampling and kernel SHAP are similar or closer on test instance 0 (id:0) compared to other test instances such as instance 1 and instance 2.

Table 1. Differences in contribution distributed across HD features

Feature	Feature value	Tree SHAP	Sampling SHAP
Chest pain (cp)	2	0.06	0.05
Thallium (thal)	2	0.07	0.07
Slope	2	0.05	0.05
Restecg	0	-0.06	-0.06
Sex	1	-0.02	-0.01
exang	0	0.07	0.06
thalach	166	-0.01	-0.01
ca	0	0.01	0.01
oldpeak	1.6	-0.00	-0.01
Age	46	-0.02	-0.02
totalrestbps	135	0.01	0.00
chol	263	-0.00	-0.01
fbs	0	-0.00	-0.01

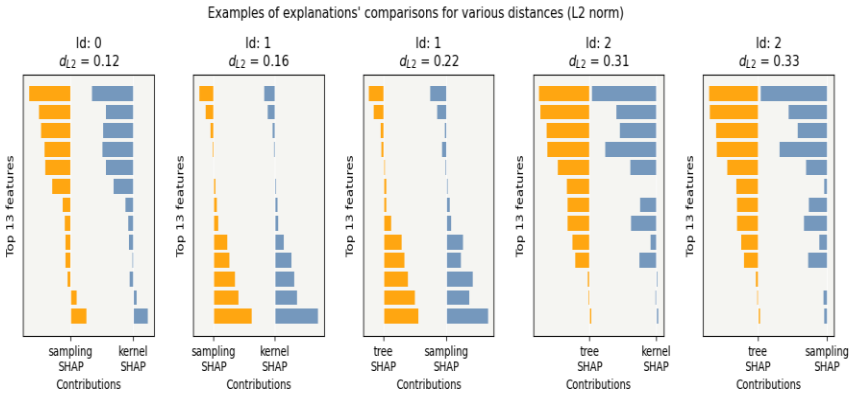

Figure 4. The average distance between multiple explanations

Figure 5 indicates the approximation or the number of features required to generate an explanation for the HD dataset sample. The number of HD features required to produce an accurate explanation. The number of features required explaining 85% of the model's output, and the percentage of the model output explained by the 13 HD features per instance is indicated in Figure 5. The top seven HD features explain at least 85% of the model for 100% of the HD dataset instance. However, all 13 features explain at least 100% of the model for 100% of the instances.

Figure 6 reveals the stability of the explainability methods. The stability is demonstrated in the neighborhood around each provided instance (reminder: neighborhood in terms of HD features and model output) which shows the average importance of the HD feature across the dataset based on its contributions (y-axis). Stability also shows the average variability of the feature across the instances' neighborhood (x-axis). The left features are stable in the neighborhood, unlike those on the right. The top features are important, unlike the bottom ones demonstrated in Figure 6. In conclusion, HD features such as "chest pain (cp)", "angina pain due to exercise (exang)", thallium scan (thal), and "slope" tends to have strong and relatively stable contributions. Thus, one might be more confident in using them for explanations. However, HD features such as "fasting blood sugar (fbs)", "total blood pressure at rest (trestbps), and "cholesterol (chol)" are much more unstable, and we might want to be careful before interpreting explanations around those features.

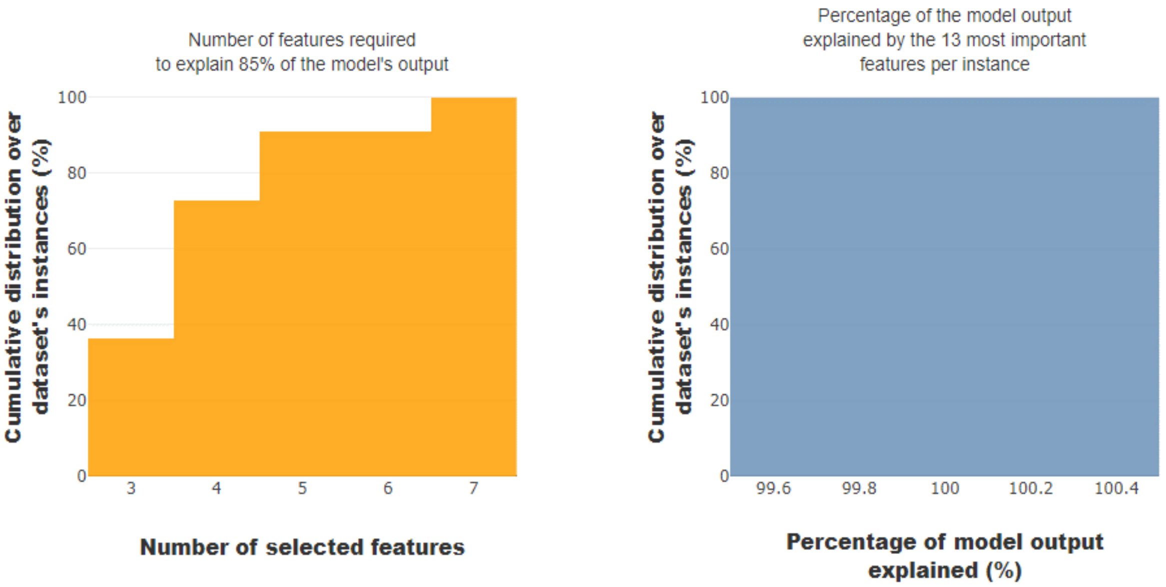

Figure 5. The average distance between explanation

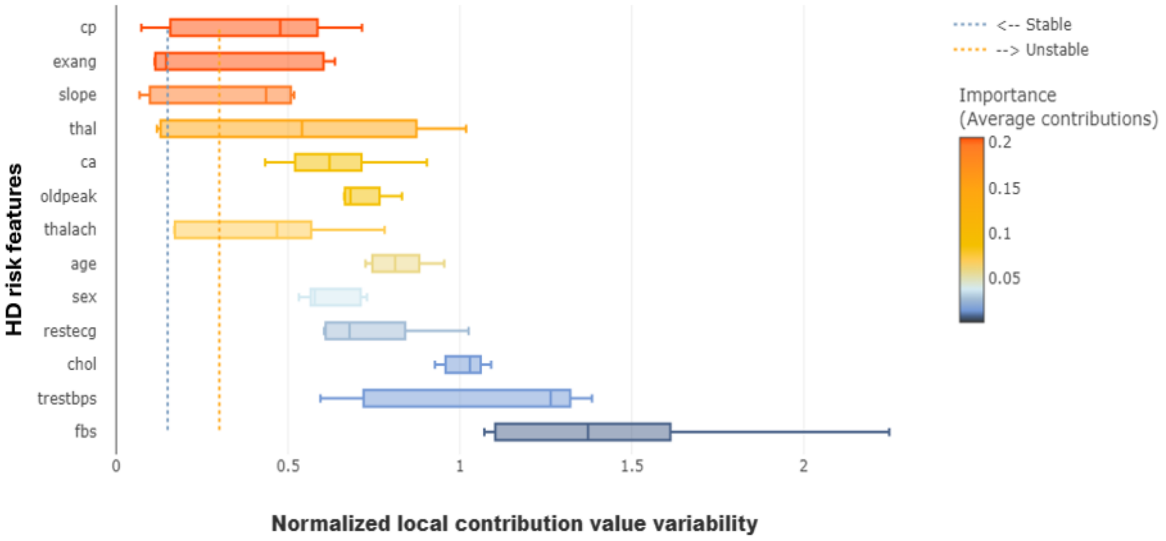

Figure 6. The average distance between multiple explanations

3. CONCLUSION

This study investigated the constituency, local stability, and approximation of LIME, SHAP, and ACV for the explanation generated on the RF regressors model using the UCI HD dataset. In conclusion, this paper shows that Shapash has desirable properties such as consistency, local stability, and approximation on HD datasets. However, it should be noted that further research is needed to evaluate Shapash in more realistic scenarios using other real-world datasets. The stability of the explanation generated by explainability methods, compactness, and consistency are crucial parameters for building confidence in the explanation generated by these methods. The explanation helps in the verification of patient outcomes predicated by the ML model.

ACKNOWLEDGEMENTS

The authors would like to acknowledge almighty God and Miss Aster Belay for her help in completing this work.

REFERENCES

[1] K. Vishnu *et al*., “Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators,” *Application Science*, vol. 11, p. 8352, 2021, doi: 10.3390/app11188352.

[2] K. Wang *et al*., “Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP,” vol. 137, p. 104813, 2021, *Computers in Biology and Medicine*, doi: 10.1016/j.compbiomed.2021.104813.

[3] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, 2021, doi: 10.3390/e23010018.

[4] A. M, Westerlund, J. S. Hawe, M. Heinig, and H. Schunker,” Risk Prediction of Cardiovascular Events by Exploration of Molecular Data with Explainable Artificial Intelligence,” *International Journal of Molecular Science*, vol. 22, p. 10291, 2021, doi: 10.3390 ijms221910291.

[5] V. Belle and I. Papantonis,” Principles and Practice of Explainable Machine Learning,” *Frontiers in Big Data*, vol. 4, pp. 1-25, 2021, doi: 10.3389/fdata.2021.688969.

[6] N. Chandrasekhar and S. Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization,” *Processese*, vol. 11, no. 1210, pp. 1–31, 2023, doi: 10.3390/pr11041210.

[7] A. Elkhawaga, O. Elzeki, M. Abuelkheir, and M. Reichert, “Evaluating Explainable Artificial Intelligence Methods Based on Feature Elimination: A Functionality-Grounded Approach,” *Electronics*, vol. 12, p. 1670, 2023, doi: 10.3390/electronics12071670.

[8] A. D Samaras *et al*., “Classification models for assessing coronary artery disease instances using clinical and biometric data: an explainable man-in-the-loop approach,” *Scientific Reports*, vol. 13, no. 1, pp. 1-15, 2023, doi: 10.1038/s41598-023-33500-9.

[9] G. Vilone and L. Longo, “Notions of explainability and evaluation approaches for explainable artificial intelligence,” *Information Fusion*, pp. 89-106, 2021, doi: 10.1016/j.inffus.2021.05.009.

[10] B. Alsinglawi1 *et al*., “An explainable machine learning framework for lung cancer hospital length of stay prediction,” *Scientific Reports*, vol. 12, no. 607, pp. 1-10, 2022, doi: 10.1038/s41598-021-04608-7.

[11] S. Mohseni and N. Zarei, “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,” *ACM Transactions on Interactive Intelligent Systems*, vol. 11, pp. 1-45, 2021, doi: 10.1145/3387166.

[12] P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, “From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies,” *Hindawi Mobile Information Systems Volume*, 2022, doi: 10.1155/2022/8167821.

[13] S. Das *et al.*, “XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI,” *The Journal of Supercomputing*, pp. 18167–18197, 2023, doi: 10.1007/s11227-023-05356-3.

[14] M. A. Kadhim and A. M. Radhi, “Heart disease classification using optimized Machine learning algorithms,” *Iraqi Journal for Computer Science and Mathematics*, 2023, doi: 10.52866/ijcsm.2023.02.02.004.

[15] A. Quadir *et al*., “Enhanced preprocessing approach using ensemble machine learning algorithms for detecting liver disease,” *Biomedicines*, vol. 11, no. 581, 2023, doi: 10.3390/ biomedicines11020581.

[16] M. Brucoli and T. C. Green, “Data mining in predicting liver patients using classification model," *Health and Technology*, vol. 12, no. 12, pp. 11–1235, 2022, doi: 10.1007/s12553-022-00713-3.

[17] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, “Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection,” *Informatics in Medicine Unlocked*, 2019, doi: 10.1016/j.imu.2019.100255.

[18] A. O. Salau, T. A. Assegie, E. D. Markus, J. N. Eneh, T. I. Ozue, “Prediction of the risk of developing heart disease using logistic regression," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 1809-1815, 2024, doi: 10.11591/ijece.v14i2.pp1809-1815.

[19] A.S. Rahman, J. M. Shamrat, Z. Tasnim, J. Roy, and S.A. Hossain, “A comparative study on liver disease prediction using supervised machine learning algorithms,” *International Journal of Scientific & Technology Research*, vol. 8, no. 11, Nov. 2019.

[20] B. Khan, R. Naseem, M. Ali, M. Arshad, and N. Jan, “Machine Learning Approaches for Liver Disease Diagnosing,” *International Journal of Data Science and Advanced Analytics*, vol. 1, no. 1, pp. 27-31, 2019.

[21] K. Hamid, A. Asif, W. Abbasi, D. Sabih and F. -u. -A. A. Minhas, “Machine Learning with Abstention for Automated Liver Disease Diagnosis,” *2017 International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, 2017, pp. 356-361, doi: 10.1109/FIT.2017.00070.

[22] 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, June 2023, pp. 3359-3366, doi: 10.11591/ijece.v13i3.pp3359-3366.

[23] M. Abdar *et al*., “Performance analysis of classification algorithms on early detection of Liver disease,” *Expert Systems with Applications*, doi 10.1016/j.eswa.2016.08.065, 2016.

[24] N. Tanwar and K. F. Rahman, “Machine Learning in liver disease diagnosis: Current progress and future opportunities,” *Materials Science and Engineering*, 2021, doi: 10.1088/1757-899X/1022/1/012029.

[25] S. Tokala *et al.*, “Liver disease prediction and classification using machine learning techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, 2023.

BIOGRAPHIES OF AUTHORS

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 networks. He has published over 50 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.

Bommy Manivannan is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India. She can be contacted at email: bommym@mits.ac.in.

Komal Kumar Napa is currently working as an Assistant Professor in the Department of Computer Science & Engineering at Madanapalle Institute of Technology & Science, Madanapalle. His research interests include machine learning, data mining, and cloud computing. He can be contacted at email: komalkumarnapa@gmail.com.

Bindu Kolappa Pillai Vijayammal completed UG in Electrical and Electronics Engineering in the year 2001 and ME in Power Electronics and Drives in 2005. She obtained her PhD in Electrical Engineering from Anna University in 2019. Received “Best Circuit Faculty” for the year 2018 in the South Indian Association of Scientists, Developers and Faculties Awards 2018 held in Bangalore. Received Project Grant from Tamil Nadu State Council for Science and Technology (2021-22) under the SPS scheme. Published 45 Research Articles in various International Journals and Conference Proceedings. She has published 5 patents and also has four professional memberships. Her research interests include Power Quality, optimization techniques, artificial intelligence, and smart grid. She served in the field of Teaching from 2005 and is currently working as an Assistant Professor at RMK College of Engineering and Technology. She can be contacted at email: bindu@rmkcet.ac.in.

Rajkumar Govindarajan is currently working as an Assistant Professor in the Department of Computer Science and Engineering (Data Science) at Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India. His research interests include machine learning, data mining, and networking. He can be contacted at email: kumar3544@gmail.com.

Sangeetha Murugan is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India. Her research interests include machine learning and data mining. She can be contacted at email: sangee525@gmail.com.

Atinkut Molla Mekonnen received his B.Sc. degree, in Information Technology Wollo University Kombolcha Institute of Technology (KIoT), Ethiopia 2017. He received his M.Sc. degree, in Computer Networks and Communications at Wollo University Kombolcha Institute of Technology (KIoT), Ethiopia in 2019. He is currently working as a lecturer at the Department of Information Technology, College of Engineering and Technology, Injibara University, Injibara, Ethiopia. His research interest includes computer networks and machine learning. He can be contacted at email: atinkut19@gmail.com.