**1. Angati Kalyan KUMAR[1], 2. Tsehay Admassu ASSEGIE[2], 3. Ayodeji Olalekan SALAU[3,6], 4. Komal Kumar NAPA[1], 5. Suguna R[5]**

Department of Computer Science & Engineering (Data Science), Madanapalle, Andhra Pradesh, India (1),
Department of Computer Science, College of Engineering & Technology, Injibara University, Injibara, Ethiopia (2),
Department of Electrical/Electronics & Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria (3),
Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India (4)
Department of Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Chennai, Tamilnadu, India (5)
Saveetha Institute of Medical &Technical Sciences, Chennai, Tamil Nadu, India (6)
ORCID: 2. https://orcid.org/0000-0003-1566-0901; 3. https://orcid.org/0000-0002-6264-9783

doi:10.15199/48.2024.02.29

# Feature Contribution to an In-Depth Understanding of the Machine Learning Model Interpretation

***Abstract**. A transparent and understandable machine learning model refers to a model that is accurate, effective, explainable, and interpretable to humans. An interpretable model reduces the gap between complex algorithms and human understanding, allowing users to trust and comprehend the process of the model's decision-making. To that end, Machine-learning models can provide information about the importance of each input feature in making predictions. Model interpretation helps users understand the factors that have the most significant impact on the model's decisions. This study implements feature importance-based model interpretation by employing a heart disease dataset. The simulation result demonstrates that with feature importance analysis, the decision-making process of the extra tree classification algorithm is easily explainable.*

***Streszczenie.** Przejrzysty i zrozumiały model uczenia maszynowego odnosi się do modelu, który jest dokładny, skuteczny, zrozumiały i możliwy do interpretacji przez ludzi. Interpretowalny model zmniejsza lukę między złożonymi algorytmami a ludzkim zrozumieniem, pozwalając użytkownikom zaufać i zrozumieć proces podejmowania decyzji w modelu. W tym celu modele uczenia maszynowego mogą dostarczać informacji o znaczeniu każdej cechy wejściowej w tworzeniu prognoz. Interpretacja modelu pomaga użytkownikom zrozumieć czynniki, które mają największy wpływ na decyzje modelu. W tym badaniu zastosowano interpretację modelu opartą na ważności funkcji, wykorzystując zestaw danych dotyczących chorób serca. Wynik symulacji pokazuje, że dzięki analizie ważności cech proces decyzyjny algorytmu klasyfikacji dodatkowego drzewa jest łatwy do wyjaśnienia.(**Funkcja przyczyniająca się do dogłębnego zrozumienia interpretacji modelu uczenia maszynowego**)*

**Keywords:** Explainable AI, model explanation, local explanation.
**Słowa kluczowe:** Wyjaśnialna sztuczna inteligencja, wyjaśnienie modelu, wyjaśnienie lokalne.

## Introduction

Machine learning is a set of methods that a computer uses to make predictions and improve its predictive performance based on data [1]. For instance, to predict the patient outcome, the computer would learn patterns from the past patient history and test results. The learning assumes that the dataset employed in training the machine-learning model is available in the required quantity and meets a certain uniformity to identify the pattern [2]. Better evaluation and trust ability of the machine-learning model is an interpretation for credible use in the medical diagnosis and patient care systems.

Machine learning models surpass humans in many heart disease patient outcome predictions [3-4]. These models have achieved a 99.6% accuracy in predicting heart disease patient outcomes. Even if the machine-learning model is as good as a human expert, there remain great advantages in terms of speed, reproducibility, and scaling. A once-implemented machine-learning model can complete a heart disease patient outcome prediction faster than humans can and reliably deliver a consistent result. Replicating a machine-learning model on another machine is fast and cheap. However, the training of a human for a task takes decades and it is costly.

While the machine-learning model has the advantages of speed, replicability, and consistency of predictive results, it does not provide insights about the data and the task it solves [5]. Ensemble learning methods such as random forest, and extra tree consists of hundreds of decision tree that vote for prediction. To understand how the decisions are made, the structures of hundreds of tree has to be looked.

Interpretable machine learning refers to the methods and models that make the internal working and the prediction of machine learning systems understandable to humans [6]. The understandability of the machine-learning model depends on the degree of interpretability of its prediction outcome. The higher the interpretability of the machine learning model, the easier it is for the human to comprehend why certain decisions or predictions have been made. Thus, a model is better interpretable if its decisions are easier for humans to understand.

Model interpretation plays a crucial role in making predictions understandable to humans. Because the process of interpreting machine learning algorithms into patient prediction requires interpretability to increase trust and acceptance [7]. Understandability refers to the explanation of how the algorithm learns from the data and the relationship it can learn between the feature and the output.

Research on the application of the ensemble-learning model [9] suggested that machine learning achieved promising results in detecting heart disease. With pre-processing methods such as chi-squared statistics, the performance of the ensemble learning system improves, achieving h 93.44% predictive accuracy. However, these model lack transparency and do not interpret their predictive outcome even if their performance is high.

To address, the explainability problem of complex learning such as ensemble learners, numerous works [9, 10] have been conducted on the explainability of machine learning for building confidence in the predictive performance of different machine learning models a research article [11] proposed interpretable machine learning model for predicting cardiovascular disease risk. The study demonstrated that interpretable models provide more accuracy and can be used as potential predictors of cardiovascular disease risk.

Furthermore, research on artificial intelligence (AI) and machine learning (ML) in medicine has increased in the last decade [12]. However, implementation of this system into clinical practice lags due to a lack of trust and explainability [13]. Most effective machine learning models are still black-box. The users and developers are unable to interpret and understand the reason behind their decision. The absence of such a transparent system can result in a lack of trust in the predictive outcomes of those systems.

Additionally, interpretable machine learning systems make transparent decisions [14]. The interpretation of the system is achieved via visualization of the parameters necessary for prediction and interpreting how to explore the relationship between the input feature and the output. Researchers use the terms interpretability and explainability interchangeably [15]. This study aims to investigate the visualization of the input features and their impact on model prediction outcomes by employing an extra tree regressor on the heart disease risk dataset.

The organization of this study is as follows. The methodology section discusses the interpretation technique employed for explaining the predictive outcome of the extra tree algorithm on heart disease patient outcomes. The result section presents the results achieved. The conclusion section presents the summary and future work.

## Methodology

The Shapley Additive Explanation (shapash) Python library is employed for interpreting the extra tree classifier prediction outcome of heart disease patients. The library provides several types of visualization, which display explicit labels to the impact of each heart disease feature on the prediction outcome of the extra tree classifier.

Shapash has been widely employed for interpreting the predictive outcomes of ensemble learning methods such as random forest and extra tree classifiers [16, 17]. The shapash library provides a module for explaining the predictive outcome of a single instance (instance-based) local and general explanation of the overall predictive outcome of predictions made by an ensemble learning classifier. Figure 1 indicates the following chart of the proposed extra tree-based heart disease risk prediction system.

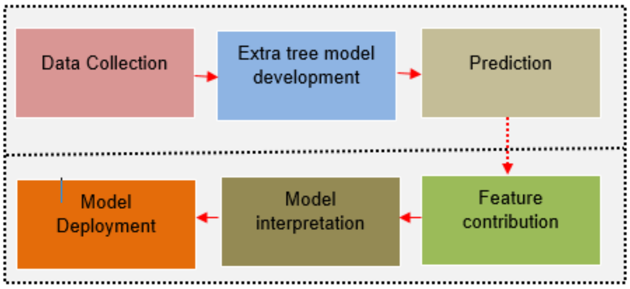

Fig.1. Flowchart of the proposed system

## Result

The data collected from the Kaggle machine learning data repository were used in training and testing the extra tree classifier for heart disease detection. The training set had 243 records (80 percent of the dataset) and the test set had 60 records (20 percent of the dataset). The research was conducted with the machine learning software Python [18, 19]. The classification algorithm employed for training and testing was an extra tree classifier. The simulations were conducted in two steps: Step 1: building models on the training set. Step 2: Interpret the model prediction outcome with feature contribution.

Figure 1 indicates the contribution of the heart disease features on the extra tree classifier for heart disease patient prediction. As confirmed in Figure 1, heart disease features such as chest pain (cp), previous history of cardiovascular disease (ca), old peak, sex, and angina due to exercise have a higher influence on the extra tree classifier. However, heart disease features such as fasting blood sugar (fbs), resting electrocardiograph (resetecg), and slope have less effect on predicting heart disease patient outcome.

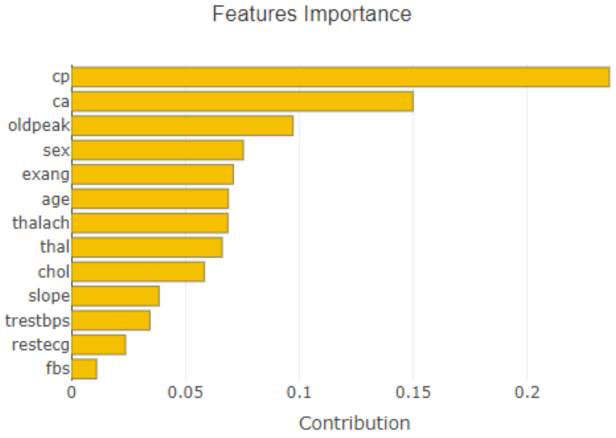

Fig.2. The contrition of heart disease feature on extra tree classifier

Table 1 shows the contribution of each feature to the predicted patient outcomes of the extra tree classifier. Chest pain contributes differently in different instances. The chest pain contrition was high for the test instances (163, 58, and 29) while the contrition was low for instance 189. Similarly, the sex feature has a higher contribution of the heart disease patient (positive class) as the contributions are high for the entire four test instances considered in the experiment. Moreover, the extra tree model correctly predicted the entire four heart disease patient outcomes.

Table 1. The contribution of top heart disease features

| Instance | Feature | Value | Contribution | Predicted | Actual |
|---|---|---|---|---|---|
| 163 | cp | 2 | 0.127294 | 1 | 1 |
| 189 | cp | 0 | -0.115029 | 1 | 0 |
| 58 | cp | 3 | 0.072536 | 1 | 1 |
| 29 | cp | 2 | 0.111976 | 1 | 1 |
| 67 | Sex | 0 | 0.098899 | 1 | 1 |
| 49 | Sex | 0 | 0.110258 | 1 | 1 |
| 84 | Sex | 0 | 0.183626 | 1 | 1 |
| 53 | Sex | 0 | 0.100252 | 1 | 1 |

Figure 3 reveals that the chest pain types (1=chest pain atypical angina, 2= typical angina pain, and 3=asymptomatic) contribute more to the heart disease patient being predicted as a heart disease patient. However, the typical angina=0 has little contribution in detecting heart disease patients. This result agrees with the chest pain contribution illustrated in Table 1. In Table 1, the extra tree classifier model incorrectly predicted the actual instance 189 in the test set which is negative as a heart disease patient for typical angina chest pain.

To compare the individual contribution of each heart disease feature and interpret how the extra tree classification model predicts heart disease patient outcome, comparison plots of both the positive class (target=1) and the negative class (target=0) instance were used in the simulation. Figure 4 indicates that chest pain contributes more for 159 compared to instance 204. Other heart disease features have also different contributions for different instances. In contrast, age and cholesterol (chol) have very little influence on the predictive outcomes of the extra tree model. The age and cholesterol features contributed similarly to the instances shown in Figure 4.

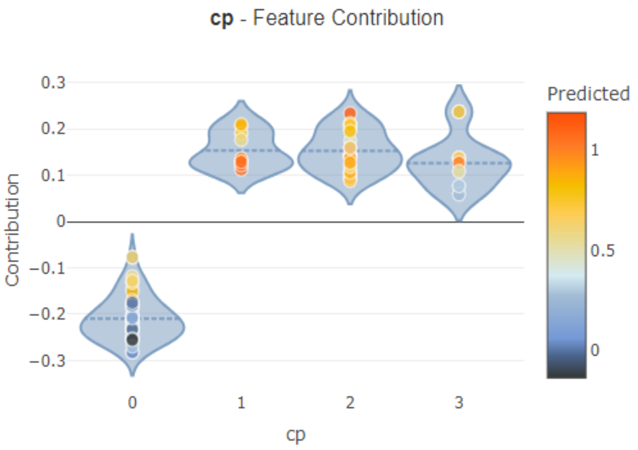

Fig.3. The contribution of chest pain (cp) on the predicted outcome

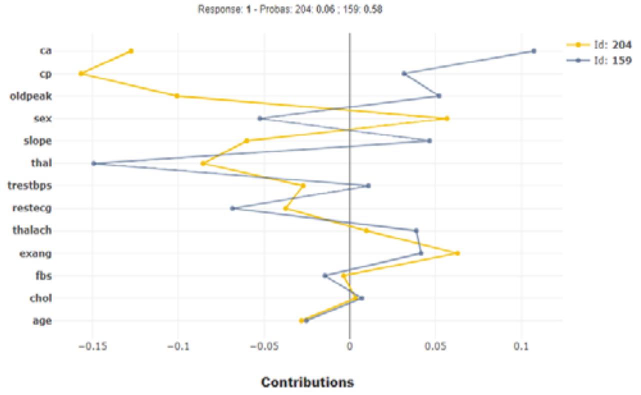

Fig.4. Comparison of heart disease feature contribution for positive class prediction

Figure 5 indicates that the comparison of heart disease features contributes to the heart disease negative class. As shown in Figure 5, the chest pain feature has a higher contribution to the predictive outcome of the extras tree classification model. In comparison to the positive class prediction demonstrated in Figure 4, the chest pain feature has a higher contribution for instance 204 compared to instance 159.

The contribution of the previous history of cardiovascular disease (ca) is validated in Figure 6. As Figure 6 reveals, when the number of blood vessels colored by fluoroscopy (ca) is 0 the patient has a high probability of getting heart disease risk.

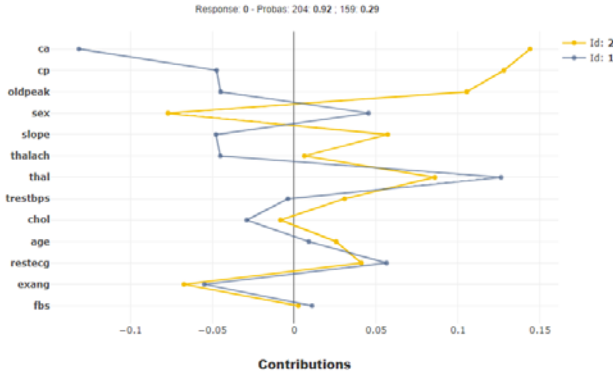

Fig.5. Comparison of heart disease feature contribution for negative class prediction

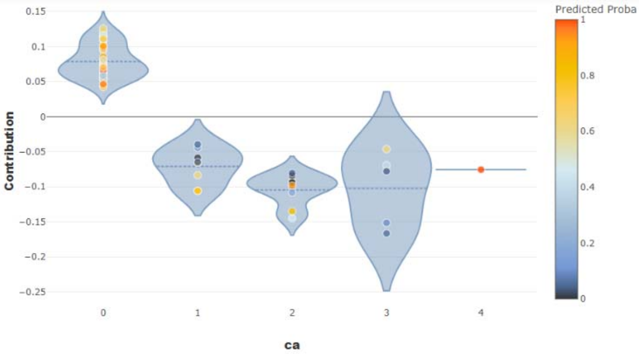

Fig.6. The contribution of the number of blood vessels on the prediction outcome of an extra tree classifier

## Conclusion

Based on this study, it is evident that feature contribution visualization is significant in building confidence in the prediction outcomes of heart disease patients by using an extra tree classifier. Moreover, the study explored that feature impact visualization with the SHAPASH Python library solves transparency and interpretability issues in predicting heart disease patient outcomes using an extra tree classifier. The feature contributing visualization revealed that the number of blood vessels, and chest pain particularly typical, asymptomatic, and atypical angina has a higher impact on the prediction of heart disease.

***Authors**: Angati Kalyan KUMAR, Department of Computer Science & Engineering (Data Science), Madanapalle, Andhra Pradesh, India.*
*Mr. Tsehay Admassu ASSEGIE Department of Computer Science College of Engineering and Technology, Injibara University, Injibara, Ethiopia. Corresponding Author E-mail: tsehayadmassu2006@gmail.com*
*Dr. Ayodeji Olalekan SALAU, Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria.*
*Komal Kumar NAPA, Department of Computer Science & Engineering (Data Science), Madanapalle, Andhra Pradesh, India.*
*Suguna R, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Chennai, Tamilnadu, India.*

## REFERENCES

[1] K.N. Kunze, A.V. Karhade, A.J. Sadauskas, J.H. Schwab, and B.R. Levine, " Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty," The Journal of Arthroplasty, 2020. DOI: https://doi.org/10.1016/j.arth.2020.03.019.
[2] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y. Han, "Reliable and explainable machine-learning methods for accelerated material discovery," *Computational Materials, 2019,* DOI: https://doi.org/10.1038/s41524-019-0248-2.
[3] L. a Kohoutová et al., "Toward a unified framework for interpreting machine-learning models in neuroimaging," *Nature Protocols, 2020,* DOI: https://doi.org/10.1038/s41596-019-0289-5.
[4] 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, vol. 13, no. 3, 2023, pp. 3359-3366, DOI: 10.11591/ijece.v13i3.pp3359-3366
[5] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Advanced Review Willey, 2020,* DOI: 10.1002/widm.1379
[6] Z. Li et al., "Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost," Computers, Environment and Urban Systems, *2022,* DOI: https://doi.org/10.1016/j.compenvurbsys.2022.101845

[7] A. Chatzimparmpas, R.M. Martins, L. Jusufi, and A. Kerren, “A survey of surveys on the use of visualization for interpreting machine learning models,” Information Visualization 2020, Vol. 19(3) 207–233, DOI: https://doi.dox.org/10.1177/1473871617751245.

[8] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y. Han, “Hybrid Feature Selection Algorithm and Ensemble Stacking for Heart Disease Prediction,” International Journal of Advanced Computer Science and Applications, Vol. 14, No. 2, 2023.

[9] J. Zacharias, M. Zahn, J. Chen, and O. Hinz, “Designing a feature selection method based on explainable artificial intelligence Electronic Markets, 2022, 32:2159–2184 DOI: https://doi.org/10.1007/s12525-022-00608-1.

[10] Q. Qiao, A.Y. Kaltungo, and R.E. Edwards, “Developing a machine learning based building energy consumption prediction approach using limited data: Boruta feature selection and empirical mode decomposition,” Enegery Reports, 2023, DOI: https://doi.org/10.1016/j.egyr.2023.02.046.

[11] H. Salah, and S. Srinivas, “Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents,” Scientific Reports (2022) 12:21905, DOI: https://doi.org/10.1038/s41598-022-25933-5.

[12] N.I. Papandrianos et al., “An Explainable Classification Method of SPECT Myocardial Perfusion Images in Nuclear Cardiology Using Deep Learning and Grad-CAM,” Applied Science, 2022, 12, 7592. DOI: https://doi.org/ 10.3390/app12157592.

[13] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y. Han, “Explainable Information Retrieval using Deep Learning for Medical images,” Computer Science and Information Systems, 2019, DOI: https://doi.org/10.2298/CSIS201030049S.

[14] R.K. Sheu, and M.S.I Pardesh, “A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System,” Sensors 2022, 22, 8068, DOI: https://doi.org/10.3390/ s22208068.

[15] G. Abdulsalam, S. Meshoul2, and H. Shaiba, “Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach,” Intelligent Automation & Soft Computing DOI: 10.32604/iasc.2023.032262.

[16] P. Guleria et al., “XAI Framework for Cardiovascular Disease Prediction Using Classification Techniques,” International Journal of Advanced Computer Science and Applications, Vol. 14, No. 2, 202Electronics 2022, 11, 4086, DOI: https:// doi.org/10.3390/electronics112440863.

[17] A. Pedro, and M. Sanchez, “Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees,” IEEE, 2020, DOI: 10.1109/BigData50022.2020.9378460.

[18] A.O. Salau, E.D. Markus, T.A. Assegie, C.O. Omeje, and J. N. Eneh, “Influence of Class Imbalance and Resampling on Classification Accuracy of Chronic Kidney Disease Detection,” Mathematical Modelling of Engineering Problems, vol. 10, no. 1, February, 2023, pp. 48-54, DOI: https://doi.org/10.18280/mmep.100106.

[19] T.A. Assegie, “Evaluation of Local Interpretable Model-Agnostic Explanation and Shapley Additive Explanation for Chronic Heart Disease Detection,” Proceedings of Engineering and Technology Innovation, vol. 23, 2023, pp. 48-59, DOI: https://doi.org/10.46604/peti.2023.10101.