INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 11, Issue - 07, July - 2025

DOIs:10.2015/IJIRMF/202507012 --:-- Research Paper / Article / Review

Using LIME and Tree SHAP to explain decision trees: A case study in Explainable AI

G. Roch Libia Rani,
Assistant Professor, Department of Computer Applications, De Paul College, Mysuru, India
Email: rochgilter@gmail.com

Abstract: *The increasing complexity of machine learning models, particularly in high-stakes domains such as healthcare, finance, and autonomous systems, has highlighted the need for transparency and interpretability. Understanding the reasons behind the AI predictions becomes the critical aspect here. Explainable AI (XAI) methods aim to bridge the gap between model performance and human understanding, enhancing trust, accountability, and decision-making. The study will focus on how these methods make black-box models more transparent and facilitate human-centered decision-making. In particular, this study explores the application of XAI methods—Local Interpretable Model-Agnostic Explanations (LIME) and Tree SHAP—to a decision tree classifier trained on a heart disease dataset. Although decision trees are generally interpretable, LIME and Tree SHAP offer deeper, instance-level and global insights into model behavior. Our analysis reveals that key clinical features such as the number of major vessels (ca), thalassemia status (thal), and chest pain type (cp) consistently influence model predictions. In this way, we expect to arrive at a suggestion that explainability is not only crucial for fostering trust in AI systems but also enhances the quality and transparency of decisions, especially when there is human oversight. The paper would also outline the challenges and future direction for integrating XAI in decision-making processes, highlighting the importance of designing systems that are both accurate and interpretable.*

Key Words: *explainable AI, decision making systems, interpretability, transparency, LIME, Tree SHAP.*

1. INTRODUCTION:

In recent years, Artificial Intelligence (AI) has revolutionized decision-making systems across diverse domains, including healthcare, finance, transportation and governance. However, as AI models grow increasingly complex, their decision-making processes often resemble "black boxes," where outputs are generated without clear insight into the underlying reasoning. This leads to lack of transparency in predicting the results. Such opaqueness also poses significant challenges, including reduced trust, limited adoption, and ethical concerns in different kinds of AI applications, especially in critical areas like healthcare, finance related fraud detection etc.

Explainable AI (XAI) emerges as a solution to bridge the gap between AI complexity and human interpretability. By providing clear, understandable explanations of model predictions and decisions, XAI enhances trust, accountability, and user acceptance(Molnar, 2025). Moreover, it empowers stakeholders, whether they are technical experts or end-users, to understand better, validate clearly, and collaborate fruitfully with AI systems.

This paper explores the potential of XAI in transforming decision-making systems. It examines how explainability can improve accuracy, fairness, and ethical accountability. This study will aim to analyse one of the current methodologies and its certain application in the field of healthcare. For this, I will use LIME (Local Interpretable Model-agnostic Explanations) and Tree SHAP (a SHAP [SHapley Additive exPlanations] variant tailored for tree-based models), two popular techniques in XAI that allow users to understand how the machine learning model, which, otherwise, makes predictions by without providing interpretable explanations, can be tweaked to provide those explanations for individual predictions. Implementing these techniques is supposed to open the ‘black box’ of complex models by highlighting which features influence the outcome most for a specific data point. This is particularly important in regulated domains like healthcare, where accountability, fairness, and transparency are as crucial as accuracy. Through this, the study will aim to highlight some of the crucial differences between AI decision making and XAI decision making.

2. LITERATURE REVIEW:

XAI is one of the most researched areas in the fields of Artificial Intelligence and Machine Learning with many works published regularly advancing the research focus. It has progressed much further from what was suggested in the beginning (van Lent, Fisher, & Mancuso, 2004). However, given the limited scope of this research, a combination of basic and comprehensive presentations on the different areas have been reviewed for the purpose. They include one of the earliest and influential papers on the methods of xAI, an overview on the history, another brief survey of the different methods of xAI, the original article introducing the LIME method amd another that introduced SHAP and finally an article that surveys the various methods with their application in the area of healthcare.

Adadi and Berrada (2018) is considered as one of the earliest papers to discuss the importance of transparency in AI systems in a lucid manner in the background of a survey of the methods prevailing at that point of time. This paper described two of the important group of researchers who were pursuing XAI. Their research seems to have laid the foundation for the present XAI research. First one was a group of researchers working under the acronym FAT. FAT here refers to Fairness, Accountability and Transparency. Their research was aimed at providing details of algorithmic decisions and the information about the data which brings in those decisions in non technical terms.The second group of researchers was funded by DARPA (Defense Advanced Research Projects Agency). They were working on increasing explainability for security applications. According to the authors these contributed in producing the third wave in AI research where XAI became an important objective and goal.

The history of xAI and its different methods are found in many works. A brief, but well researched overview is found in the workshop and conference proceedings of 'xxAI -- beyond explainable AI: International Workshop'. The editorial and another paper in this volume explore the history of the methods and the different methods respectively. In the editorial, Holzinger et al (2022a) describe how the focus has been shifting in the research through decades. According to these editors, the deep learning methods, known as DL, brought new vigor to the research in machine learning. They state, 'by demonstrating its power in learning from vast amounts of data in order to solve complex tasks…often even beyond human level performance'. This is one of the reasons deep learning methods have made AI extremely popular, according to them. However, they point out to the fact that it is the complexity of the DL models, where millions of parameters are involved, that led to the black box experience. This made the researchers and users struggle at different points. The editors see the development of different toolboxes to bail out the struggling researchers as the starting point of XAI. The editors then trace the further development of XAI models. According to them, there have been at least two trajectories. The models developed through the first path look at whether an explanation is local or global, that is, whether it explains a single AI decision (local) or gives insights into how the whole model works (global). Here, the global model gives understanding to an entire layer or network. In the second trajectory, the models examine whether the method is post-hoc or ante-hoc. The post-hoc explanations come after a model is trained, whereas the ante-hoc methods build explainability directly into the AI's design. This helps make AI decisions clearer and more understandable for users. While the field of XAI concentrated on the tools which resulted in 'safe, responsible, ethical and accountable deployment of AI technology', the future, of course, would have to deal with wider and newer scenarios, according to the editorial. The future would most often have to deal with the phenomenon of unsupervised and intensified learning and would have to offer well structured and easy explanations with which humans could make decisions. This is also a scenario, where the editorial indicates the healthcare field to become a major beneficiary of XAI where counterfactuals would arise providing explanations to independent factors that affect the decision making or prediction.

The second paper Holzinger et al (2022b) describe the different methods of XAI. Overall, seventeen methods have been individually described under three criteria, namely, idea, source repository and discussion. Under the 'idea', the authors explain the origin of the model and development. The different aspects of the model are also explained briefly. They give the 'GitHub' address from where the files can be accessed under the 'source repository'. Under 'discussion', the authors compare the particular model with other models. They discuss the advantages and disadvantages of the models in comparison.

Band et al. (2023) have surveyed various methods of explainable AI and their applications in the field of healthcare. The article explains how each method is used individually for various disease diagnostics and what their results imply in each case. They mention how in the beginning CNN method was combined with LIME, SHAP etc. to achieve the explainability in Vivo Gastral Images. The various methods reviewed by Band et al. (2023) include layer-wise relevance propagation (LRP), Uniform Manifold Approximation and Projection (UMAP), Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP), Gradient-weighted Class Activation Mapping (Grad-CAM) etc. They also have studied some newer models in the field like Contextual importance and Utility (CIU) which performs faster than LIME and SHAP. However, they note that the model has been going through the early stages of development and therefore lacks a comparable evolution with the other two methods. This study has particularly looked at their systematic review of the two methods that have been considered in this study,

namely, SHAP and LIME. About SHAP, Band at al. explain how Color Plot is used by the method to predict the health condition as safe or unsafe. They explain how it classifies images into “normal” or “abnormal” for Covid-19 detection or classifies antifungal peptides (positive samples) and non-antifungal peptides (negative samples) in different studies. The value of each feature is determined based on the thickness of the connecting line in the given neural network which is common to many methods. They present its working through figures. Different features are assigned values in a step-by-step process using SHAP. If the values assigned by SHAP is low, it would indicate that the respective feature has a much smaller impact than other features on the process of prediction. If the value assigned through SHAP is high, then it may mean that the corresponding feature has a high and strong influence on the prediction. By this way, the method explains how different features have influenced the prediction process. However, Band et al (2023) also caution that the interpretation of values as low and high would vary depending on the specific context of the disease diagnosis. The interpretation would also be different according to the dataset used and the nature of the problem that is being analysed.

LIME (Ribeiro et al., 2016) is an explainability technique designed to interpret individual AI model predictions. Since many complex models, such as deep learning and ensemble methods, function as "black boxes," LIME helps make their decisions more understandable by approximating them with simpler, interpretable models. Being model-agnostic, LIME can be applied to various machine learning models, including logistic regression, random forests, decision trees and neural networks. In healthcare, for example, it can explain why an AI system predicts a patient is at high risk for heart disease by highlighting influential features like cholesterol levels or blood pressure, improving trust and transparency.

The SHAP method of interpreting predictions was first introduced in Lundberg and Lee (2017). The authors found that the models that were used to interpret the predictions at that time lacked one or the other feature and were different from each other in that way. The authors considered it a necessity therefore to combine the models to create a unified approach and SHAP was the result of that. The models integrated included the popular ones like LIME and DeepLIFT. In this way it became one of the most used approaches for interpreting predictions. The authors used game theory concepts as a theoretical ground for building their method. SHAP will assign a value to each of the features in any prediction. The total gain generated as a result of predictions is considered as the one that is shared by each feature and higher the value assigned to a feature, the more influencing it was considered in the prediction process.

However, computing exact Shapley values can be computationally intensive, especially for models with many features or instances. To address this limitation for tree-based models, Tree SHAP was introduced by Lundberg, Erion, and Lee (2019) as a highly efficient, model-specific variant of SHAP. Tree SHAP leverages the internal structure of decision trees to compute exact Shapley values in polynomial time, eliminating the need for sampling or approximation. Unlike the general SHAP method—which relies on techniques like Kernel SHAP or permutations—Tree SHAP provides fast, accurate explanations tailored specifically for decision trees, random forests, and gradient boosting models. This makes Tree SHAP particularly well-suited for use in real-time or large-scale applications involving tree-based classifiers.

3. DESCRIPTION OF THE DATASET:

I have used the heart disease dataset available at https://www.kaggle.com/code/prasenjitsharma/beginner-heart-disease-prediction/input for our research in this paper. The heart-disease dataset serves as a prominent resource in medical research and machine learning. It is designed to predict heart diseases based on clinical and demographic parameters. The dataset includes features such as age, chest pain type, resting blood pressure, cholesterol level, maximum heart rate, and others that are relevant for cardiovascular diagnosis. It comprises 1,025 patient records with 14 attributes which reflects various risk factors associated with cardiovascular conditions.The target column is a binary classification variable that indicates whether a patient has a heart disease or not. It is represented as: 1 (Presence of heart disease), 0 (Absence of heart disease). This column serves as the dependent variable in predictive modeling. This helps to classify patients based on their cardiovascular risk.

4. METHODOLOGY:

The methodology adopted in the study is structured to support both technical depth and clarity. This study employs the heart disease dataset obtained from Kaggle. Prior to model development, a series of preprocessing steps were conducted to ensure data quality and consistency. Missing or anomalous values were handled appropriately, categorical variables were encoded using suitable techniques and numerical features were scaled to standardize the input space. The processed dataset was then divided into training and testing sets using an 80:20 ratio, ensuring a balanced distribution of the target classes. This prepared dataset serves as the foundation for training the decision tree model and applying interpretability techniques like LIME and Tree SHAP. I used Google Colab environment to execute the code.

5. RESULTS AND DISCUSSION:

5.1 LIME Results:

The model predicted that Prediction Probability for class 0 (No Disease): 0.00 and Prediction Probability for class 1 (Disease): 1.00. The model is 100% confident that the patient has heart disease.It assigned zero probability to the "No Disease" class. This is a very strong prediction — especially common with Decision Trees, which can overfit the training data.

In LIME method results for a specific instance predicted as having heart disease, the most influential features included the absence of major vessels (ca<= 0), normal thalassemia (thal = 2), and the presence of typical angina (cp = 0). These three features alone accounted for more than 80% of the model's prediction weight, as revealed through LIME-based local explanation. Additional but smaller contributions came from high heart rate, no ST depression, and patient age."

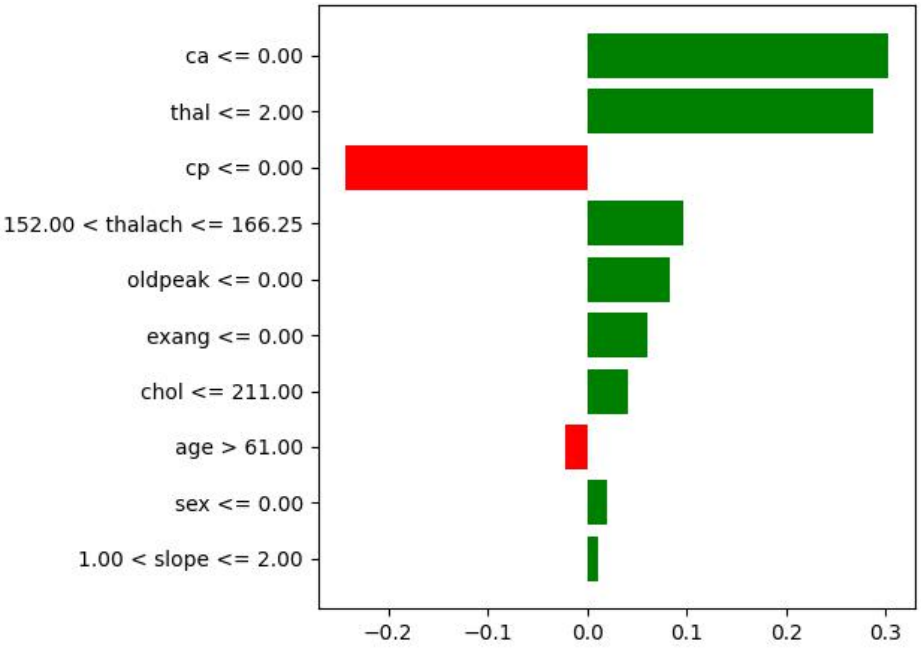

Figure 1. Local explanation for class disease

This output is a LIME explanation for a single prediction made by Decision Tree model. It shows how each feature condition influences the model's prediction — either increasing or decreasing the likelihood of the predicted class (likely "Disease" in this case). The model predicts "Disease" for this patient based on the following key factors:

- Absence of major vessels colored by fluoroscopy (ca = 0)
- Normal thalassemia result (thal = 2)
- Presence of typical angina (cp = 0)

5.1.1 LIME-Based Interpretation Summary

These features are the most influential in driving the prediction. Additional but smaller contributions come from other attributes such as age, sex, and the slope of the ST segment.

To better understand the decision-making process of the trained decision tree model, Local Interpretable Model-Agnostic Explanations (LIME) was applied to a correctly classified instance of heart disease. The explanation revealed that several key features significantly influenced the model's prediction.The most influential factors were the absence of major vessels colored by fluoroscopy (ca ≤ 0.00), the presence of normal thalassemia (thal ≤ 2.00), and the occurrence of typical angina (cp ≤ 0.00), each contributing substantially to the predicted probability of disease. Additional features such as no ST depression (oldpeak ≤ 0.00), absence of exercise-induced angina (exang ≤ 0.00), and a high maximum heart rate (thalach > 152) also provided moderate support for the disease classification.

Smaller contributions came from features like normal cholesterol levels (chol ≤ 211.00), female sex (sex = 0), and the slope of the ST segment. Notably, all the top contributing conditions had positive influence weights, indicating that they consistently pushed the model's decision toward the "disease" class.This instance-level explanation highlights LIME's ability to uncover feature-level reasoning in black-box models and reinforces the model's alignment with domain knowledge, thereby enhancing its trustworthiness in clinical decision support.

5.1.2 Feature value summary:

The instance selected for local interpretability analysis corresponds to a 62-year-old female patient presenting with typical angina, no exercise-induced angina, and no ST depression, all of which are potentially indicative of lower cardiac risk. However, the model strongly predicted heart disease, primarily due to the absence of colored vessels (ca = 0), normal thalassemia (thal = 2), and a high heart rate (thalach = 163).

Other supporting features included downsloping ST segment, borderline-high cholesterol, and normal resting blood pressure, each contributing moderately to the overall prediction. The consistency between these input features and the model's high-confidence prediction, as confirmed by LIME and SHAP explanations, illustrates the model's ability to integrate multiple subtle indicators in its decision-making process.

Feature	ca	thal	cp	oldpeak	exang	thalach	chol	Sex	trestbps	slope
Value	0.00	2.00	0.00	0.00	0.00	163.00	209.00	0.00	124.00	2.00

Table 1. Feature - value output

	Precision	Recall	F1-score	Support
No Disease	0.97	1.00	0.99	102
Disease	1.00	0.97	0.99	103
Accuracy			0.99	205
Macro avg	0.99	0.99	0.99	205
Weighted avg	0.99	0.99	0.99	205

Table 2. Evaluation metrics for Tree SHAP

5.2 SHAP results:

The big difference between SHAP and LIME is the weighting of the instances in the regression model. LIME weights the instances according to how close they are to the original instance. There are three ways to estimate Shapley values for explaining predictions: KernelSHAP, Permutation Method, and TreeSHAP. Lundberg, Erion, and Lee (2019) proposed TreeSHAP, a variant of SHAP for tree-based machine learning models such as decision trees, random forests, and gradient-boosted trees. TreeSHAP was introduced as a fast, model-specific alternative to KernelSHAP. To complement LIME's local perspective, Tree SHAP was applied to the entire test set. Tree SHAP assigns Shapley values to each feature, reflecting its average contribution to predictions across all instances. The accuracy of the model is 98.5%. The most influential global features were ca (number of major vessels), thal (thalassemia), cp (chest pain type), oldpeak and thalach (ST depression and max heart rate).

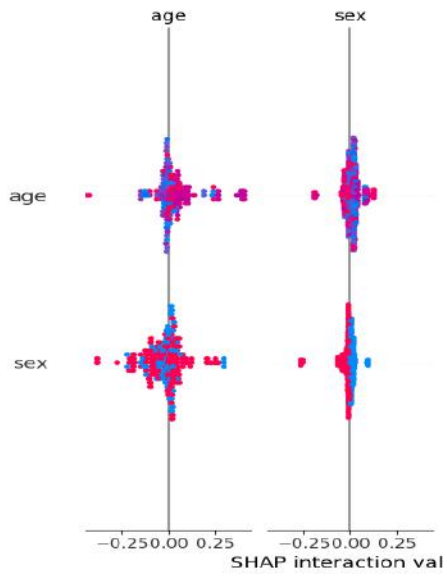

Figure 2. SHAP interaction val

The findings aligned closely with the LIME analysis, reinforcing the model's reliance on clinically significant features.

6. CONCLUSION

Explainable AI (XAI) is essential for ensuring that AI-driven decisions are transparent, trustworthy, and actionable—particularly in healthcare. In this study, I applied LIME and Tree SHAP to interpret predictions made by a decision tree classifier trained on heart disease data.

LIME offered human-understandable, local explanations, while Tree SHAP provided consistent, model-specific insights across the dataset. Both techniques identified the same core features (ca, thal, cp) as dominant contributors to prediction, and their alignment with clinical understanding enhances the interpretability and reliability of the model.

This case study underscores the value of combining XAI techniques to better understand model behavior, even for models like decision trees that are traditionally seen as interpretable. Future work can extend this analysis to more complex black-box models and include feedback from clinical users to evaluate the practical usability of AI explanations in real diagnostic settings.

REFERENCES:

1. Molnar, C. (2025). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Christoph Molnar.
2. van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence (IAAI'04)*,900–907.
3. Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).*IEEE Access*, *6*,52138-52160.
4. Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (2022). xxAI - Beyond Explainable Artificial Intelligence. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *XxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 3-10). Springer International Publishing.
5. Holzinger, A., Saranti, A., Molnar, C., Biacek, P., & Samek, W. (2022). Explainable AI Methods - A Brief Overview. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *XxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 13-38). Springer International Publishing.
6. Band, S. S., Yarahmadi, A., Hsu, h.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., & Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, *40*, 101286.
7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).
8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
9. Lundberg, S. M., Erion, G. G. & Lee, S.-I. (2017). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv*, https://doi.org/10.48550/arXiv.1802.03888.