# Explainable Artificial Intelligence: A Comprehensive Review Of Methods And Applications For Transparent Machine Learning With Heart Disease Prediction Case Study

1st Anika Shaji
*Independent Researcher*
Kerala, India
shajianika@gmail.com

2nd Sachin Gavankar
*Independent Researcher*
Mumbai, India
sachin@computer.org

***Abstract*—Explainable Artificial Intelligence (XAI) is a critical research area that addresses the opacity of modern machine learning systems. As AI becomes more complex and is deployed in high-stakes fields like healthcare and finance, transparency and interpretability are paramount. This paper reviews XAI methodologies, categorizing them into local explanation methods (e.g., SHAP, LIME), global interpretation techniques (e.g., permutation importance), and inherently interpretable models. A practical case study on heart disease prediction demonstrates the value of using multiple XAI techniques to understand model behavior. This work contributes to the field by providing a structured taxonomy of explanation methods and highlighting future research directions for creating transparent and trustworthy AI systems.***

***Keywords*—Explainable AI, Machine Learning Interpretability, Model Transparency, SHAP, LIME, Responsible AI, Heart Disease Prediction***

## I. Introduction

The rapid advancement of artificial intelligence and machine learning technologies has led to unprecedented capabilities in pattern recognition, prediction, and decision-making across various domains [1]. However, the increasing complexity of these systems, particularly deep neural networks and ensemble methods, has created what researchers commonly refer to as the "black box" problem – where the decision-making process remains opaque to human understanding [2]. This opacity presents significant challenges when AI systems are deployed in critical applications where understanding the reasoning behind decisions is essential for trust, accountability, and regulatory compliance [3].

Explainable Artificial Intelligence (XAI) has emerged as a fundamental research area aimed at addressing these challenges by developing methods and techniques that make AI decision-making processes transparent and interpretable [4]. The field encompasses a diverse range of approaches, from post-hoc explanation techniques that analyze existing models to inherently interpretable algorithms designed with transparency as a core feature [5]. The importance of XAI extends beyond technical considerations, encompassing ethical, legal, and social implications of AI deployment in society [6].

The growing regulatory landscape, exemplified by the European Union's General Data Protection Regulation (GDPR) and proposed AI Act, mandates explanations for automated decision-making systems, particularly those affecting individual rights and freedoms [7]. Furthermore, industries such as healthcare, finance, and criminal justice require explainable AI systems to ensure fairness, prevent bias, and maintain human oversight in critical decisions [8].

This paper provides a comprehensive review of the current state of XAI research, examining the methodological landscape and categorizing explanation techniques based on their scope and approach. We analyze local explanation methods that provide insights into individual predictions, global techniques that reveal overall model behavior, and inherently interpretable models that maintain transparency by design. Through a practical heart disease prediction case study, we demonstrate the real-world application of these techniques and their clinical relevance. Through this systematic review and empirical analysis, we aim to contribute to the understanding of XAI methodologies and identify future research directions for advancing transparent and trustworthy AI systems.

## II. Literature Review

The concept of explainable artificial intelligence traces its roots to the early development of expert systems in the 1970s and 1980s, where rule-based systems inherently provided explanations through their logical reasoning chains [9]. However, the modern XAI movement gained momentum with

the rise of complex machine learning models, particularly deep learning, which achieved remarkable performance at the cost of interpretability [10].

Ribeiro et al. introduced LIME (Local Interpretable Model-agnostic Explanations) in 2016, marking a significant milestone in post-hoc explanation methods by providing a model-agnostic approach to explaining individual predictions [11]. This work established the foundation for local explanation techniques and demonstrated the feasibility of explaining complex models without requiring access to their internal structure.

Concurrently, Lundberg and Lee developed SHAP (SHapley Additive exPlanations), which provided a unified framework for feature attribution based on cooperative game theory [12]. SHAP addressed several limitations of existing methods by ensuring additive consistency and providing a solid theoretical foundation for explanation techniques [13]. The introduction of SHAP values revolutionized the field by offering both local and global explanations through a single, mathematically principled approach.

Researchers have proposed various taxonomies to categorize XAI methods. Guidotti et al. presented a comprehensive survey distinguishing between model-specific and model-agnostic approaches, as well as local versus global explanations [14]. This categorization has become widely adopted in the XAI community and provides a structured framework for understanding different explanation methodologies.

Molnar's work on interpretable machine learning established a foundational framework that separates intrinsic interpretability from post-hoc explanations, emphasizing the trade-off between model complexity and interpretability [15]. This distinction has influenced subsequent research directions and policy discussions regarding the deployment of AI systems in regulated environments.

Local explanation methods focus on understanding individual predictions and have received significant attention in the XAI literature. Beyond LIME and SHAP, researchers have developed specialized techniques such as Anchors, which generate high-precision rules that guarantee specific predictions [16]. These rule-based explanations provide intuitive understanding by identifying minimal sufficient conditions for model decisions.

Integrated Gradients, introduced by Sundararajan et al., addresses attribution in deep learning models by accumulating gradients along a path from a baseline input [17]. This method satisfies important axioms such as sensitivity and implementation invariance, making it particularly suitable for neural network interpretability [18]. The technique has found widespread application in computer vision and natural language processing domains.

Counterfactual explanations have emerged as another important class of local methods, providing actionable insights by identifying minimal changes required to achieve different outcomes [19]. Wachter et al. demonstrated the legal and practical significance of counterfactual explanations in automated decision-making contexts [20]. These explanations are particularly valuable in applications such as loan approval and medical diagnosis, where understanding alternative scenarios is crucial.

Global explanation methods aim to understand overall model behavior and have evolved from traditional statistical techniques. Permutation importance, rooted in random forest literature, provides a model-agnostic approach to feature importance assessment [21]. This method has been extended and refined by various researchers to address challenges such as feature correlation and computational efficiency [22].

Partial Dependence Plots (PDPs), introduced by Friedman, visualize the marginal effect of features on predictions while averaging over other variables [23]. However, researchers identified limitations of PDPs in the presence of feature interactions, leading to the development of Accumulated Local Effects (ALE) plots by Apley and Zhu [24]. ALE plots address the correlation problem by considering conditional rather than marginal effects.

Morris sensitivity analysis, adapted from uncertainty quantification literature, provides a systematic approach to assess feature importance across different regions of the input space [25]. This method offers insights into model behavior under various conditions and helps identify features with consistent versus variable importance.

The development of inherently interpretable models represents an alternative approach to post-hoc explanations. Letham et al. introduced Bayesian Rule Lists, which generate probabilistic if-then rules with associated confidence measures [26]. These models maintain competitive accuracy while providing inherent interpretability through their rule-based structure.

Explainable Boosting Machines (EBMs), developed by Lou et al., demonstrate that additive models can achieve performance comparable to complex ensemble methods while maintaining full interpretability [27]. EBMs represent each feature's contribution through univariate functions, allowing for detailed analysis of individual feature effects and interactions.

Tree surrogate models provide another approach to achieving interpretability by training simple decision trees to approximate complex model behavior [28]. This technique enables understanding of global model patterns through the interpretable structure of decision trees while maintaining the performance benefits of complex models.

## III. Methods

Explainable Artificial Intelligence (XAI) methodologies are systematically categorized into three principal classes based on their scope and inherent approach to transparency: Local Explanation Methods, Global Explanation Methods, and Inherently Interpretable Models. This structured classification is fundamental to navigating the diverse landscape of XAI techniques and aligning them with specific interpretability requirements.

### *A. Local Explanation Methods*

SHAP (SHapley Additive exPlanations) is arguably one of the most theoretically grounded and widely adopted local explanation methods. Its foundation lies in cooperative game

theory, where each feature in a model's input is treated as a "player" in a game, and the model's prediction is the "payout." SHAP values quantify the marginal contribution of each feature to the prediction for a specific instance, considering all possible permutations and combinations of features. This sophisticated calculation ensures that the sum of the SHAP values for all features, plus a baseline (often the average prediction), equals the actual prediction. Key properties satisfied by SHAP include efficiency (contributions sum up correctly), symmetry (features with identical marginal contributions receive equal SHAP values), and dummy feature identification (features that do not affect the prediction receive zero SHAP values). This theoretical rigor makes SHAP highly reliable for attributing importance and provides a unified framework for interpreting various model types, from simple linear models to complex neural networks, allowing for consistent comparisons across different models and predictions. SHAP can be used for both local explanations (for a single prediction) and aggregated to provide global insights.

LIME (Local Interpretable Model-agnostic Explanations) is a groundbreaking model-agnostic technique designed to explain the predictions of any "black-box" machine learning model. For a particular instance whose prediction needs explaining, LIME generates a new, localized dataset. This is achieved by systematically perturbing the original input instance (e.g., adding noise to image pixels or removing words from a text document) and observing the black-box model's predictions on these perturbed samples. LIME then weights these perturbed samples by their proximity to the original instance and trains a simpler, inherently interpretable model (such as a linear regression model or a shallow decision tree) on this weighted, localized dataset. This interpretable model acts as a "local surrogate," effectively explaining the complex model's behavior only within the immediate vicinity of the specific prediction. The simplicity of the surrogate model makes its internal workings transparent, providing human-understandable feature importances or rules for that individual prediction. While LIME provides intuitive explanations, its robustness can sometimes depend on the choice of perturbation and the fidelity of the local surrogate.

Anchors provide a different kind of local explanation: high-precision "if-then" rules that guarantee a specific prediction with high confidence. Instead of attributing importance to features, Anchors identify a minimal set of conditions (features and their corresponding values) that, when present, will almost certainly lead to the same model prediction, regardless of other feature values. These rules are called "anchors" because they "anchor" the prediction; if the conditions of an anchor are met, the prediction is highly stable. Anchors are particularly valuable for providing actionable insights because they clearly state the minimal sufficient conditions for a decision, making it straightforward for users to understand what combinations of inputs consistently lead to a specific outcome. This rule-based clarity is often preferred in scenarios requiring clear, unambiguous guidance.

The Contrastive Explanation Method (CEM) takes a nuanced approach by identifying "pertinent positive" and "pertinent negative" features. Pertinent positives are features that must be present for a specific prediction to occur, highlighting the evidence for the decision. Conversely, pertinent negatives indicate features that must be absent or altered for a different outcome, alongside identifying features that are irrelevant to the decision. This method provides a more comprehensive understanding of the model's logic by showing both supporting and differentiating factors.

Counterfactual Instances focus on providing actionable insights by generating the minimal modifications to input features that would result in a different prediction. This method answers the question: "What is the smallest change I could make to my input for the model to predict something else?" These explanations are particularly valuable as they directly show users how to achieve a desired outcome or avoid an undesirable one.

Integrated Gradients is a method primarily used for deep learning models. It attributes importance scores by accumulating gradients along a linear path from a "baseline" input (e.g., an all-zeros image) to the actual input instance. This approach ensures that the attribution scores satisfy two crucial axioms: sensitivity (features affecting the output should have non-zero attribution) and implementation invariance (attribution should be consistent regardless of the model's specific implementation if it produces the same function).

Protodash provides explanations for individual predictions by identifying a small set of representative prototype examples from the training data. These prototypes are chosen because they are most similar to the instance being explained and best exemplify the model's decision for that particular case. This method leverages the intuitive human tendency to understand concepts and decisions through concrete examples.

B. *Global Explanation Methods*

Global explanation methods aim to provide a holistic understanding of a machine learning model's overall behavior across the entire dataset. These methods tackle questions about the model's general functioning, such as which features are universally significant or how features typically influence predictions. They are crucial for model validation, identifying potential biases, and ensuring the model's logic aligns with domain expertise.

Permutation Importance is a robust and widely used model-agnostic technique for assessing global feature importance. Its mechanism is straightforward: to determine the importance of a specific feature, its values are randomly shuffled (permuted) across the validation or test dataset, while all other features remain intact. The model then makes predictions on this modified dataset, and the resulting drop in performance (e.g., accuracy, R-squared, F1-score) compared to the original performance is measured. A significant decline in performance indicates that the permuted feature was highly important to the model's overall predictions. This method is intuitive, easy to implement, and can be applied to any black-box model. However, it can be misleading when features are highly correlated, as shuffling one correlated feature might inadvertently affect the perceived importance of another.

Partial Dependence Plots (PDPs) visually represent the marginal effect of one or two features on the predicted outcome of a model. To construct a PDP for a given feature, the method averages the model's predictions as the value of that feature

varies across its range, while the values of all other features are held constant (or marginalized over their observed distribution). The resulting plot shows how the average prediction changes in response to variations in the specific feature(s) of interest. PDPs are excellent for understanding the general directional relationship between a feature and the target variable, revealing whether the relationship is linear, monotonic, or more complex. A key limitation, however, is that PDPs assume feature independence, which can lead to misinterpretations if features are strongly correlated, as they may average over unrealistic or impossible data points.

Morris Sensitivity Analysis, adapted from the field of uncertainty quantification, systematically perturbs individual features (one at a time) across various regions of the input space. By observing the impact of these perturbations on the model's output, it assesses the overall influence and consistency of each feature's effect. This method helps identify features that have a consistent impact across the entire input space versus those whose importance varies significantly depending on other feature values.

Accumulated Local Effects (ALE) plots were developed to address the independence assumption limitation inherent in Partial Dependence Plots, especially when dealing with correlated features. Instead of averaging over the marginal distribution of features, ALE plots calculate the effect of a feature by considering its conditional distribution. This means they assess how predictions change when a feature's value varies within its observed local context, considering the values of other features in that specific data point. By focusing on changes in predictions within local neighborhoods, ALE plots provide a more accurate and reliable visualization of a feature's true marginal effect, even in the presence of strong correlations. This makes ALE plots a preferred method for understanding global feature effects in complex, real-world datasets where features are rarely truly independent.

Global Interpretable Rule-based Predictions (GIRP) methods aim to summarize the overall decision logic of complex models into interpretable tree structures. This is often achieved by constructing binary trees from feature contribution matrices, which might be derived from other explanation methods. The goal is to provide a high-level, human-understandable overview of the model's internal workings, distilling complex interactions into a simpler, digestible rule set.

### C. *Inherently Interpretable Models*

Inherently interpretable models represent an alternative paradigm, building transparency directly into their architectural design rather than relying on post-hoc explanation techniques for "black-box" models. These models are transparent "by design," meaning their decision-making process is inherently understandable without additional effort. While they may not always match the peak predictive performance of highly complex models, their interpretability is a core feature, not an afterthought.

Bayesian Rule Lists (BRLs) are a prime example of inherently interpretable models that generate ordered "if-then" rules with associated probabilities. These models learn a concise and accurate list of rules from the data, which are then applied sequentially to make predictions. For instance, a rule list might state: "IF (age > 65 AND smoking_status = 'yes') THEN risk = high (probability 0.8); ELSE IF (cholesterol < 200) THEN risk = low (probability 0.9); ELSE risk = medium." The clear, logical structure of these rules, combined with the probabilistic confidence measures, makes them highly transparent and easy for humans to understand and verify. BRLs prioritize simplicity and accuracy in rule generation, often leading to models that are both performant and deeply understandable, aligning well with human cognitive processes.

Tree Surrogates provide a strategy for achieving interpretability by training a simpler, inherently interpretable model (typically a decision tree) to mimic the behavior of a more complex, opaque "black-box" model. The idea is that while the black-box model (e.g., a complex neural network) makes the ultimate predictions, the decision tree acts as a "surrogate" that approximates its logic. By analyzing the structure and decision paths within this simpler decision tree, one can gain a global understanding of the more complex model's underlying patterns and decision logic, even if the approximation isn't perfect. This approach allows for a valuable balance, leveraging the high predictive performance of complex models while gaining insights into their general decision-making processes through the transparent structure of the surrogate tree. It's particularly useful when full transparency of the complex model is infeasible, but some level of global understanding is required.

Explainable Boosting Machines (EBMs) are a form of generalized additive models (GAMs) that achieve a remarkable balance between high predictive accuracy and complete interpretability. Unlike traditional black-box boosting models, EBMs explicitly model the contribution of each feature independently, and can also account for common pairwise interactions. The model's final prediction is simply the sum of individual functions for each feature (and potentially pairs of features). This additive structure allows for direct and clear visualization of how each feature (or feature interaction) influences the prediction. For example, for a particular feature like "age," an EBM can produce a shape function plot showing precisely how the model's output changes as age increases, independent of other features. This modularity makes EBMs fully transparent and enables detailed analysis of individual feature effects, often matching the performance of complex ensemble methods while retaining full interpretability.

## IV. Case Study: Heart Disease Prediction

### *A. Dataset and Experimental Setup*

To demonstrate the practical application of XAI methodologies, we conducted a comprehensive case study using a heart disease prediction dataset containing 303 patient records with 14 clinical features. The dataset includes demographic information (age, sex), clinical measurements (chest pain type, resting blood pressure, cholesterol levels, blood sugar levels), and diagnostic test results (ECG results, maximum heart rate, exercise-induced angina, ST depression, slope of peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia type).

The target variable represents the presence (1) or absence (0) of heart disease. After data preprocessing and duplicate removal,

the final dataset contained 302 samples with a balanced distribution: 138 samples (45.7%) without heart disease and 164 samples (54.3%) with heart disease. The dataset was split into training (241 samples) and testing (61 samples) sets using stratified sampling to maintain class balance.

### *B. Model Development*

We implemented and compared four different machine learning models: Random Forest, XGBoost, Neural Network, Logistic Regression, to provide a comprehensive analysis of XAI techniques across various algorithmic approaches.

Additionally, we trained inherently interpretable models for comparison:
Decision Tree: For rule-based interpretability
Simple Logistic Regression: For coefficient-based interpretability

## V. Results and Findings

### *A. Model Performance Analysis*

The experimental results demonstrate varying performance across different machine learning models. All complex models achieved similar test accuracy of approximately 78.7%, with notable differences in other metrics. The Random Forest model showed the highest AUC score (0.878) and cross-validation performance (0.834), indicating superior discriminative ability and generalization. Interestingly, the inherently interpretable models (Decision Tree and Simple Logistic Regression) achieved competitive accuracy (0.803), demonstrating that interpretability does not necessarily compromise predictive performance in this cardiovascular disease prediction domain.

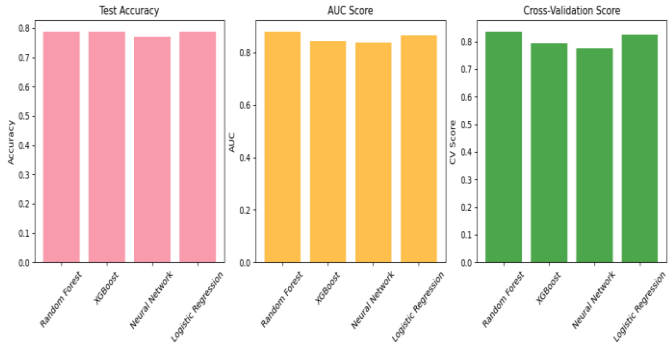


**Figure 1.** Comparative Performance Analysis of Machine Learning Models for Heart Disease Prediction

### *B. SHAP Local Explanations*

The SHAP local explanations revealed how individual features contribute to specific patient predictions (Figure 2). For Patient 0 (correctly predicted as low risk), the analysis showed:

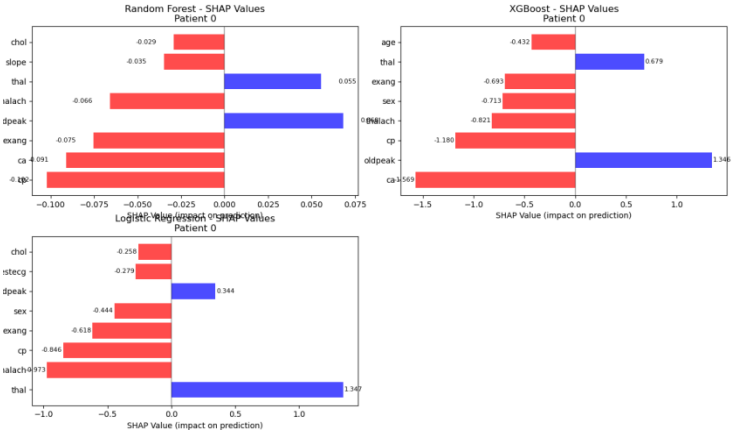


**Figure 2.** Local explainability analysis for a representative patient case across multiple models. (A) Random Forest SHAP values, (B) XGBoost SHAP values, (C) Logistic Regression SHAP values, Red bars indicate protective factors, blue bars indicate risk factors. Feature abbreviations: cp=chest pain type, ca=major vessels, thal=thalassemia, oldpeak=ST depression, exang=exercise angina

SHAP value analysis revealed distinct feature importance patterns across models (Figure 3). XGBoost demonstrated more pronounced feature discrimination with impact values ranging from -1.5 to +1.0, while Random Forest showed more balanced contributions (-0.075 to +0.075). Key features consistently contributed positively or negatively across models, though their relative importance rankings varied between algorithms, indicating different learned patterns from the same dataset.

### *C. LIME Local Explanations*

LIME explanations for the same patient (Figure 3) provided complementary insights through interpretable rules:

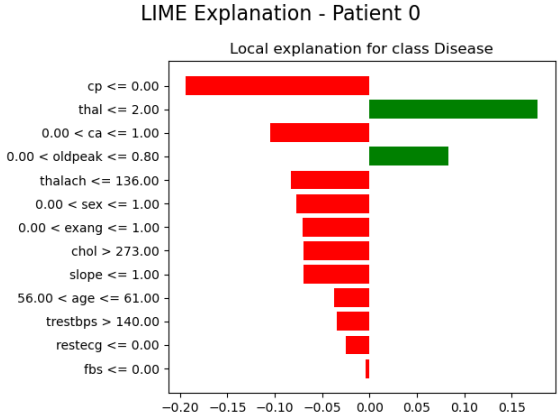


Figure 3. Local explainability analysis for a representative patient case. LIME local explanations. Red bars indicate protective factors, green bars indicate risk factors. Feature abbreviations: cp=chest pain type, ca=major vessels, thal=thalassemia, oldpeak=ST depression, exang=exercise angina.

LIME analysis for Patient 0 identified key features influencing disease classification (Figure 4). The model's prediction was primarily driven by chest pain type (cp <= 0.00) and maximum heart rate (thal <= 2.00) as strong positive predictors (red bars), while calcium score (0.00 < ca <= 1.00) and exercise-induced angina (0.00 < oldpeak <= 0.80) served as negative predictors (green bars). Additional contributing factors included thalassemia type, sex, and various cardiac parameters,

demonstrating the model's reliance on clinically relevant cardiovascular indicators.

### *D. Permutation Importance*

The permutation importance analysis across all models revealed consistent patterns in feature significance (Figure 4). The top five most important features identified were:

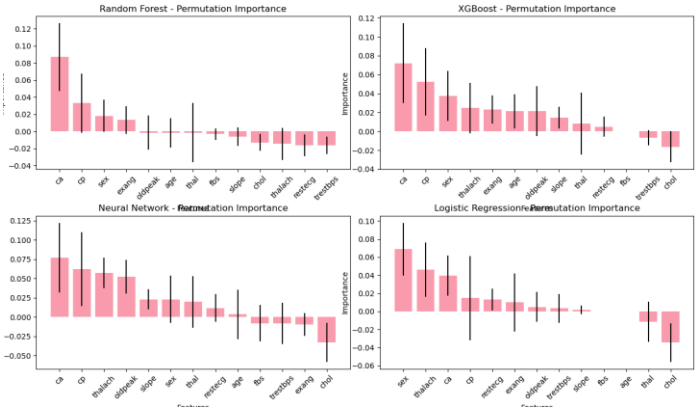


**Figure 4.** Permutation importance analysis across four machine learning models showing consistent feature ranking patterns. Error bars represent standard deviation across cross-validation folds. Pink bars indicate positive importance values.

ca (Number of major vessels): Consistently ranked highest across all models
cp (Chest pain type): Second most important feature universally
sex (Gender): Significant predictor across all models
exang (Exercise-induced angina): Important diagnostic indicator
oldpeak (ST depression): Cardiac stress test parameter

The consistency of these rankings across different model types strengthens confidence in their clinical relevance. Notably, traditional risk factors such as age showed relatively lower importance, suggesting that specific cardiac indicators are more predictive than demographic factors in this dataset.

### *E. Logistic Regression Coefficients*

The logistic regression coefficient analysis (Figure 6) provided linear relationships between features and disease probability:

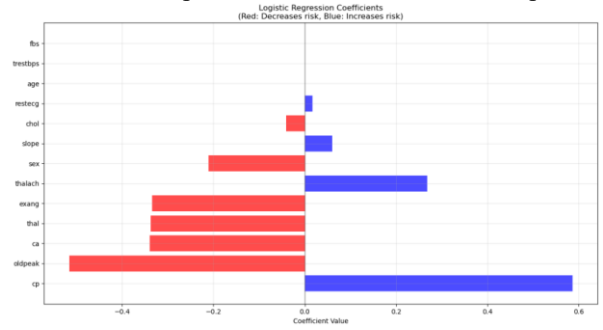


**Figure 6.** Logistic regression coefficient analysis showing linear feature contributions to disease probability. Red bars indicate protective factors (negative coefficients), blue bars indicate risk factors (positive coefficients). Coefficient magnitudes represent log-odds ratios.

**Strongest Positive Coefficients (Increase Disease Risk):**
cp (Chest pain type): +0.586
thalach (Maximum heart rate): +0.268
slope (ST segment slope): +0.060

**Strongest Negative Coefficients (Decrease Disease Risk):**
oldpeak (ST depression): -0.516
ca (Number of major vessels): -0.340
thal (Thalassemia): -0.337
exang (Exercise-induced angina): -0.335

These coefficients align with the SHAP and permutation importance rankings, providing convergent evidence for feature significance.

## VI. Conclusion

This comprehensive review of Explainable Artificial Intelligence methodologies, enhanced by practical application to heart disease prediction, reveals both the maturity and ongoing challenges in the XAI field. The case study demonstrates that XAI techniques can provide clinically meaningful insights while maintaining competitive predictive performance. The results suggest that using multiple explanation techniques provides more robust understanding of model behavior, with different formats suitable for various stakeholders. As AI systems become increasingly prevalent in healthcare and other critical domains, this work demonstrates that XAI techniques are sufficiently mature for practical deployment while providing a template for responsible AI deployment that balances performance, transparency, and clinical utility.

## VII. FUTURE ENHANCEMENTS AND LIMITATIONS

Future research directions include developing standardized metrics for evaluating explanation quality, advancing human-AI interaction studies to understand how clinicians incorporate AI explanations into decision-making, and integrating causal inference methods with current XAI techniques for deeper insights into disease mechanisms. Research into computationally efficient explanation methods for real-time clinical decision support systems remains critical. This study acknowledges several limitations: the dataset size (302 samples) limits generalizability, the focus on a single medical domain may not represent XAI performance across all healthcare applications, and the retrospective analysis does not capture real-world clinical workflow integration challenges. The continued advancement of XAI methodologies is essential for building trustworthy, fair, and accountable AI systems that benefit humanity while maintaining necessary transparency and human oversight for ethical deployment in high-stakes domains.

## References


[1] McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Magazine, 27(4), 12-14.

[2] Castelvecchi, D. (2016). Can we open the black box of AI? Nature News, 538(7623), 20-23.

[3] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[4] Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44-58.

[5] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[6] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

[7] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50-57.

[8] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.

[9] Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23(3-4), 351-379.

[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "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).

[12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

[13] Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647-665.

[14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42.

[15] Molnar, C. (2020). Interpretable machine learning. Lulu.com.

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[17] Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328).

[18] Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711.06104.

[19] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841.

[20] Karimi, A. H., Barthe, G., Balle, B., & Valera, I. (2020, July). Model-agnostic counterfactual explanations for consequential decisions. In International Conference on Artificial Intelligence and Statistics (pp. 895-905).

[21] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[22] Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8(1), 1-21.

[23] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

[24] Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4), 1059-1086.

[25] Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. Technometrics, 33(2), 161-174.

[26] Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 9(3), 1350-1371.

[27] Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013, August). Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 623-631).

[28] Craven, M., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. Advances in Neural Information Processing Systems, 8, 24-30.