Evaluating Explanation Consistency of Explainable Machine Learning Models for Heart Disease Risk Prediction

Kari Sai Vardhan*
AIT-CSE (AIML), Chandigarh University, Mohali, India
saivardhankari@gmail.com
ORCID: 0009-0008-9558-8535

Maram Ramakrishna Reddy
AIT-CSE (AIML), Chandigarh University, Mohali, India
rk0010430@gmail.com
ORCID: 0009-0004-1747-0153

Kore Akhil
AIT-CSE (AIML), Chandigarh University, Mohali, India
akhilkore7@gmail.com
ORCID: 0009-0003-9583-8403

Modugula Pavan Kumar Reddy
AIT-CSE (AIML), Chandigarh University, Mohali, India
modugula.pavan2004@gmail.com
ORCID: 0009-0004-3355-8876

Aaskaran Bishnoi
AIT-CSE (AIML), Chandigarh University, Mohali, India
aaskaran.e15060@cumail.in
ORCID: 0009-0003-1853-2937

***Abstract*—Machine learning models have demonstrated potential in the risk prediction of heart disease but the dependability of the explanation in the clinical setting is not thoroughly examined. This paper compares the prediction of risks of heart diseases using linear and non-linear machine learning models regarding the ability to achieve consistency and explainability. Clinically relevant measures and cross-validation as well as calibration analysis are used in evaluating the Logistic Regression, Random Forest, and Multi-layer Perceptron models. The SHAP and LIME are used to interpret model predictions to produce global and local explanations. The quantitative metric of consistency of the explanation is the overlap of features in SHAP and LIME explanations on a patient level. The experimental findings show significant differences in the degree of explanation, which indicate that there can be a high risk encompassed in the application of one approach to explainability of a study. The results indicate that forecasts predictive performance with a strong emphasis on the examination of explanation stability to support explainable AI that is trustworthy should the application be in the medical industry.**

Keywords*—*Explainable Artificial Intelligence, Heart Disease Risk Prediction, Machine Learning, Explanation Consistency, SHAP, LIME, Healthcare AI, Model Interpretability.

I. INTRODUCTION.

Cardiovascular diseases are still amongst the top causes of deaths in all regions of the globe, with heart disease taking a considerable percentage of the preventable deaths. The identification of high-risk people at an early stage is important in enhancing clinical outcomes due to timely intervention and lifestyle change. Machine learning (ML) practices have grown in popularity in the last few years as cognitive models predictors of heart disease risk because of their capability to capture multidimensional interdependence of clinical variables, and their capacity to perform competitively.

Although numerous researchers have established the promise of machine learning based heart disease prediction models, they are yet to be implemented in the practical clinical environment. The lack of transparency and interpretability of complex models is one of the major barriers because it may reduce clinician trust and may impede decision-making. To overcome this issue, explainable artificial intelligence (XAI) techniques (SHAP (SHapley Additive exPlanations)) and (LIME (Local Interpretable Model-agnostic Explanations)) have become particularly popular to understand the predictions of a model by demonstrating the features (both global and local) that affect its behavior significantly.

Although explainability techniques are increasingly used, the vast majority of obtained works tend to offer explanation images just without properly assessing their validity or integrity. Specifically, little focus has been given on whether various explainability modalities can give consistent explanations of the same patient. This can be key to undermining confidence in AI-assisted decision support systems and can cast uncertainty on their applicability to clinical spots in fields with high stakes like healthcare.

This gap leads to this study that explores the explainable AI risk prediction of heart diseases with particular attention paid to consistency of the explanations. Clinically relevant performance metrics and calibration analysis is used to estimate linear and non-linear machine learning models such as Logistic Regression, Random Forest, and Multi-Layer Perceptron. SHAP and LIME are used to interpret model predictions, and the extent to which an explanation is consistent is a quantitative measure (named feature overlap) of how models explain individual patients. Moreover, case-based explanation framework is proposed to offer reasoning in human readable form by integrating model predictions, key features contributing to such predictions, and similarity-based patient outcome.

This work has threefold contributions, including (i) comparative analysis of interpretable and non-interpretable machine learning models to predict heart disease, (ii) quantitative experimentation of the consistency of explanations at the patient level offered by SHAP and LIME, and (iii) a case-

based approach to explanation whose purpose is to enhance the comprehensibility and reliability of healthcare AI systems. This study gives empirical indications to a more dependable and trusted use of explainable AI in clinical decision support by focusing on both the stability of explanting and predictive performance.

II. LITERATURE REVIEW.

1) **Cleveland et al. (1988)** presented a benchmark dataset of heart disease that has been extensively used in the cardiovascular risk prediction research and showed the importance of clinical factors, in particular age, cholesterol, blood pressure, and electrocardiography.

2) **Detrano et al. (1989)** used the classical statistical methods and logistic regression to predict heart disease. Their results indicated that linear models could deliver results which were clinically interpretable and with a sufficient predictive performance thus the choice of logistic regression is becoming a standard baseline of medical risk prediction tasks.

3) **Karegowda et al. (2011)** examined the use of machine learning classifier which includes: Decision Trees, naive bayes, and support vector machines (SVM) made to predict heart disease. Their research focused on the relevance of feature selection in enhancing the accuracy of classification and minimizing complexity of the model. The authors have concluded that the ensemble and tree based models perform better as compared to the simple linear classifiers in the occurrence of non-linear relationships.

4) **Polat and Güneş (2007)** examined hybrid machine learning methods of the fusion of feature selection and classification methods in the detection of heart disease. They showed that an increase in the number of features which are irrelevant and redundant leads to a significant drop in the quality of the model and its generalization ability.

5) **Reddy et al. (2018)** compared the machine learning models, such as Logistic Regression, Random Forest, and Neural Networks, to predict cardiovascular disease. They have found that the accuracy of the Random Forest models was greater because they are ensemble models, but less interpretable.

6) **Rajkomar et al. (2019)** expressed the increased concern about the interpretability of any of the machine learning models in healthcare usage. They stressed the fact that although black-box models work well, they cannot be adopted in clinical practice because they are not transparent and are not explainable.

7) **Lundberg and Lee (2017)** introduced SHAP (SHapley Additive exPlanations), a game theory method to understand individual predictions of a machine learning model. SHAP is increasingly stable, with local accuracy of attributing features to classifiers and has been popularly applied in healthcare research to understand multicompany classifier models like Random Forest and Gradient Boosting classifiers.

8) Ribeiro et al. (2016) developed LIME which interprets. nearest neighbor approximates of the model using surrogates which were simple and interpretable. LiME framework is model-agnostic i. e. framework can be used on. a heart disease risk is any classifier.

9) **Caruana et al. (2015)** interpretable models with conditions near to the black-box ones can be used to solve healthcare problems, and these models can be explained in terms of clinical meaning. They found that transparency rather than marginal gains in accuracy is the key factor that medical decision support systems should be driven by.

10) **Adebayo et al. (2018)** critically evaluated how explainability methods work by demonstrating that explanation methods can be prone to give unsteady or distorted explanations of certain conditions. Their work sensitized about how people could be blindly trusting in explanatory ways, without considering their soundness.

11) **Molnar (2020)** showcased the drawbacks associated with using visual explanation plots, including the feature importance charts, and claimed that explaining stability and consistency in high stakes environments and applications, to which healthcare belongs, should be determined quantitatively.

12) **Slack et al. (2020)** also noted that varying explainability techniques could provide dissimilar results to the reason behind an identical model forecast. Their paper revealed that systematic testing of the agreement of explanations cannot be achieved by replacing interpretability methods.

13) **Doshi-Velez and Kim (2017)** reiterated the human-centered interpretability to high-stakes fields like healthcare and stated that explanations should be comprehensible and implementable by clinicians. Their efforts contributed to the realization of the disparity between technical forms of explainability techniques and human reasoning that led to the adoption of case-based explanation systems.

Although previous investigations have used SHAP or LIME as prediction frameworks of heart disease, the majority of them did not assess the reliability of their explanation using qualitative data visualizations. There are limited studies that analyze the consistency of interpretations of different explainability methods at the individual level of patients. Conversely, the current study quantitatively assesses the consistency of the explanation of SHAP and LIME based on metric feature

overlap, and uses case-based explanations to optimize the model outputs with the clinical thinking.

III. EQUATIONS.

1) Logistic Regression Model:

Logistic Regression predicts the probability of heart disease using a sigmoid function applied to a linear combination of input features:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

Where:

- x= input feature vector
- **w**= weight vector
- b= bias term
- $\sigma(z)$= sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

2) Cross-Entropy Loss Function:

The model is trained by minimizing binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Where:

- N= number of training samples
- y_i= true label
- $\hat{y}_i$= predicted probability

3) Random Forest Prediction:

Random Forest combines predictions from multiple decision trees using majority voting:

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x})\}$$

Where:

- $h_t(\mathbf{x})$= prediction of the t^{th}tree
- T= number of trees

4) SHAP Explanation:

SHAP explains a prediction as a sum of feature contributions:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

Where:

- **ϕ_i= contribution of feature i**
- **M= number of features**
- **ϕ_0= expected model output**

5) Explanation Consistency (SHAP–LIME Overlap):

Agreement between SHAP and LIME explanations is measured using feature overlap:

$$\text{Overlap} = \frac{|F_{SHAP} \cap F_{LIME}|}{|F_{SHAP} \cup F_{LIME}|}$$

Where:

- F_{SHAP}= top-k SHAP features
- F_{LIME}= top-k LIME features

6) Probability Calibration (Brier Score):

The quality of predicted probabilities is evaluated using the Brier Score, which measures the mean squared difference between predicted probabilities and actual outcomes:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2$$

Where:

- N= number of samples
- $\hat{p}_i$= predicted probability for sample i
- y_i= true class label

IV. METHODOLOGY.

1) Dataset Description:

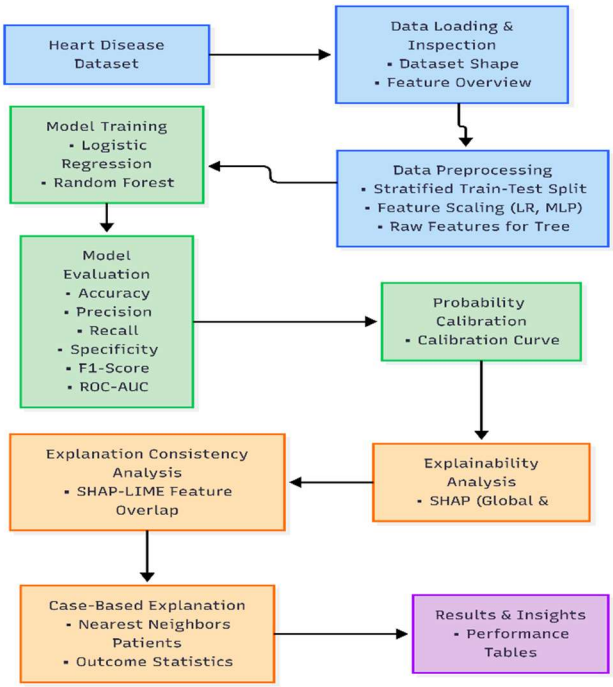
The dataset that is used in the study is publicly available heart disease data, which is composed of structured clinical features and data about the patients who undergo cardiovascular assessment. The information presented in the dataset includes demographic data, physiological data, and diagnostic data like

age, sex, chest pain type, Resting blood pressure, fasting blood sugar, cholesterol level, maximum level in the heart rate, electro-cardiographic results, as well as, maximum, acute level and thalassemia existence and number of great vessels. Dependent variable is a dichotomous variable.

Prior to modeling, the dataset is examined to verify dimensional Before modeling, the data is tested to check the dimensional consistency, the classes, and integrity of data. The figure above indicates that the target distribution is relatively balanced and will eliminate the possibility of egregious bias in the training process without encouraging the application of the recovery-oriented evaluation metrics since the ability to identify positive cases correctly is of clinical significance.

2) Overall Workflow:

The proposed structure takes a pipeline that is structured to create a balance between predictive accuracy and interpretability. The process starts with the preprocessing of data and training the model then continues to systematic evaluation, calibration analysis and explainability assessment. The main factor of the framework is the combination of analysis of consistency of the explanation and the case-based reasoning to increase the confidence in the model predictions. Figure 1 gives the entire workflow of the proposed explainable heart disease risk prediction system.



3) Data Preprocessing:

The data set is stratified into a training dataset and testing dataset in order to maintain the original distribution of the classes. The preprocessing of features is done selectively as per the demands of the model. Linear and neural network models have the option of standardizing features with z-score normalization, as an optimization has to be stable, and no feature with larger numeric scales should be given an edge. The training of tree-based models is performed with unscaled values of features, since these values are by current definition monotonic feature transformations.

The advantages of this selective preprocessing approach include that each of the models is educated on the circumstances that the model hypothesizes and therefore a fair comparison can be done between linear, non-linear, and neural methods.

4) Model Selection and Training:

Three machine learning models are picked representing different degrees of model complexity and interpretability. The use of Logistic Regression as a linear baseline can be explained by the fact that it is highly spread in clinical risk models and interprets its coefficients in a transparent way. Random Forest is taken as a non-linear ensemble model which has the ability of modeling complex interactions between clinical features and provides robust performance. Multi-Layer Perceptron is added to demonstrate the learning of the neural network and to test how the learning of the model expressiveness can positively affect the predictive results.

Binary cross-entropy loss is used to test all models. Aiming to reduce the effect of possible imbalance in classes and focusing on classifying cases of heart diseases, the use of class-weighting tactics is used in training. Hyperparameters are used in a non-aggressive way so as to prevent overfitting and to allow extrapolating the results to unobserved data.

5) Model Evaluation Strategy:

Various evaluation metrics, which are clinically significant to heart disease prediction, are used to evaluate model performance. Besides the accuracy, precision, recall (sensitivity), specificity, F1-score, and ROC-AUC, accuracy, precision, and recall (sensitivity) are also computed. The significance of recall and specificity can be attributed to the high cost of false negative and false positive in medical decision-makers. Confusion matrices are adopted to present specific analysis of classification results.

To further check the robustness of the models, the stratified k-fold cross-validation is applied to the main models. The process also lowers the differences in the performance estimates and makes the findings not reliant on one train and test split.

6) Probability Calibration Analysis:

The probability calibration analysis was done based on the percentage statistics of seizure prediction. In addition to the classification accuracy, reliability of predicted probabilities is also measured using calibration analysis. Calibration curves are

obtained to compare observed frequencies of personal results with values of prediction probabilities. The Brier Score is a quantative calculation of probability calibration values with negative Brier Score values implying a high match between predicted risks and the actual outcomes. The analysis is important in the clinical setting where probabilities as predicted could be directly utilized in treatment choice.

7) Explainability Techniques:

Global and local methods of explanation are also used to increase the interpretability. SHAP is employed to obtain the contribution of features according to the cooperative game theory to allow the ranking of global significance opinionably and give explanations to local predictions. LIME is used as a model-agnostic complementary technique, which estimates the model locally through an interpretable surrogate model.

Global explanations give an idea of which clinical factors have the strongest effect on prediction of the entire dataset, whereas local ones explain why this or that patient has to be classified as a high or low risk.

8) Explanation Consistency Analysis:

Another new element of the given study is quantitative assessment of the consistency of explanation. The best performing features by SHAP and LIME are plucked out in each instance of the tests. Measurement of the agreement between these explanations is done by metric of feature overlap to provide an evaluation of the stability of the explanation in dissimilar explainability methods. It is an analysis that takes a step further than the visual aspect and empowers the reliability of the explanations produced through various approaches through empirical means.

9) Case-Based Explanation Framework:

A case-based explanation framework is proposed in order to ensure that model explanations are more aligned with clinical reasoning. The predictions are made relative to similar past cases of a patient through nearest-neighbor analysis in the feature space. Result statistics of comparable patients are summarized and appropriated alongside feature-level clarifications to produce human readable insight. It is an intermediate between the algorithmic explanation and clinician-like comparative reasoning.

10) Implementation and Reproducibility:

All the experiments are done on Python with popular machine learners and their explainability libraries. There is the use of fixed random seeds in the workflow so that it can be reproducible. All the pipeline that consists of preprocessing, training, evaluation, and explanation generation are processed in a controlled environment that promotes clear and reproducible experimental studies.

V. RESULTS AND DISCUSSION.

1) Model Performance Evaluation:

This Table provides a comparison of Logistic Regression (LR), Random Forest (RF), and Multi-Layer Perceptron (MLP) based on accuracy, precision, recall, specificity, F1-score, and ROC -AUC. RF has the best ability in detecting heart disease cases as indicated by the highest value of recall and ROC-AUC. LR offers a competitive performance and decision boundaries are simpler as compared to MLP which does not offer a big improvement yet is more complex.

	LR	RF	MLP
Accuracy	0.7869	0.8033	0.7869
Precision	0.7632	0.7561	Precision is omitted due to unstable probability estimates
Recall	0.8788	0.9394	0.8182
Specificity	0.6786	0.6429	0.7500
f1	0.8169	0.8378	0.8060
auc	0.8712	0.9042	0.8463

2) Error Analysis and Calibration:

Figure 1 below provides confusion matrix of Random Forest. RF has the lowest false-negative rate which is very important in clinical risks prediction. Figure 2 shows RF and LR calibration curve indicating that there is a reasonable comparison between the expected probabilities and the result. The acceptable calibration of probability is verified by the Brier Score.

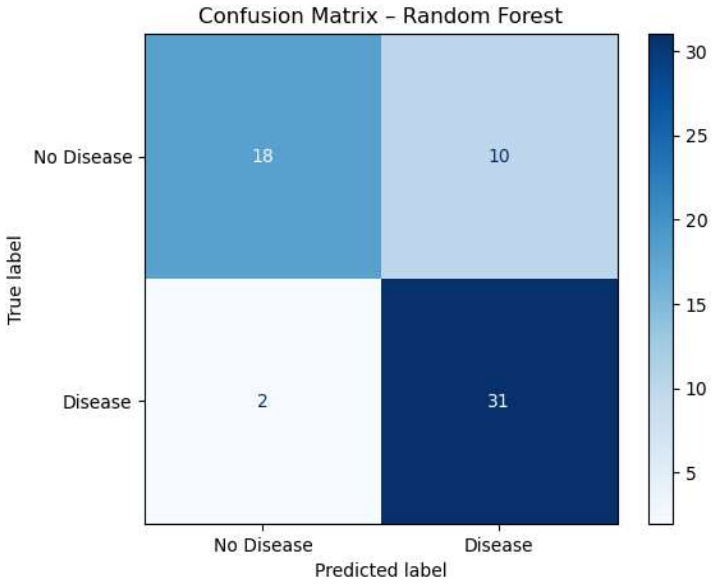


Figure 1

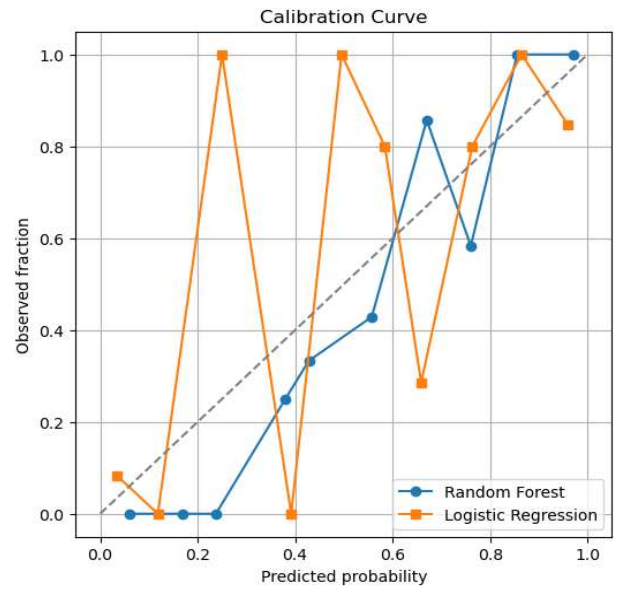


Figure 2

3) Global Explainability Analysis:

SHAP was used to obtain global feature importance as shown in Figure 3. The clinical characteristics of the type of chest pain, number of great vessels, ST depression, thalassemia, and exercise-induced angina make the best contribution to prediction of heart disease. The SHAP beeswarm plot demonstrates the trend and swings in change in the influence of features, which are known medical risk factors.

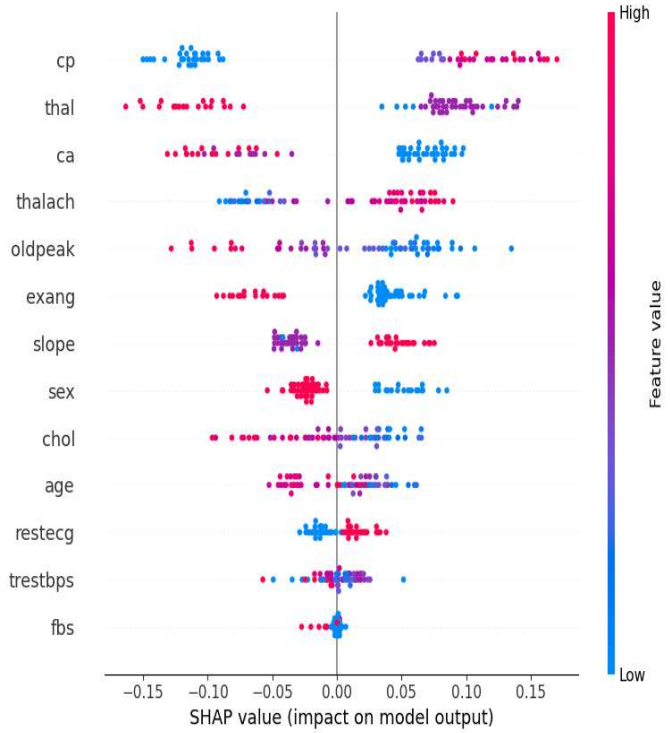


Figure 3

4) Local Explainability and Consistency:

To understand how individual patients can have their predicted risk increased or decreased by given features, Figure 4 is a SHAP force plot which is used to explain the mechanism. The explanation of LIME of the same patient is given in Figure 5. Figure 6 shows SHAP-LIME feature overlap distribution in the test instances, which portrays the moderate agreement, as well as the variability among the explanation method.

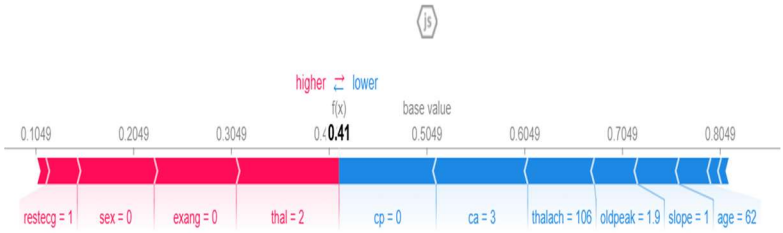


Figure 4

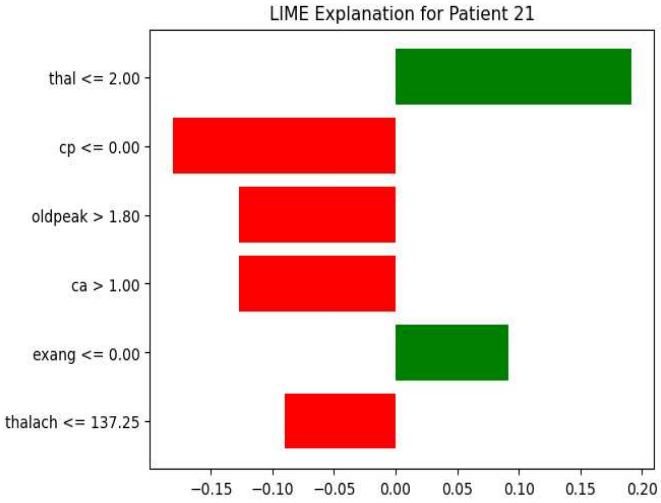


Figure 5

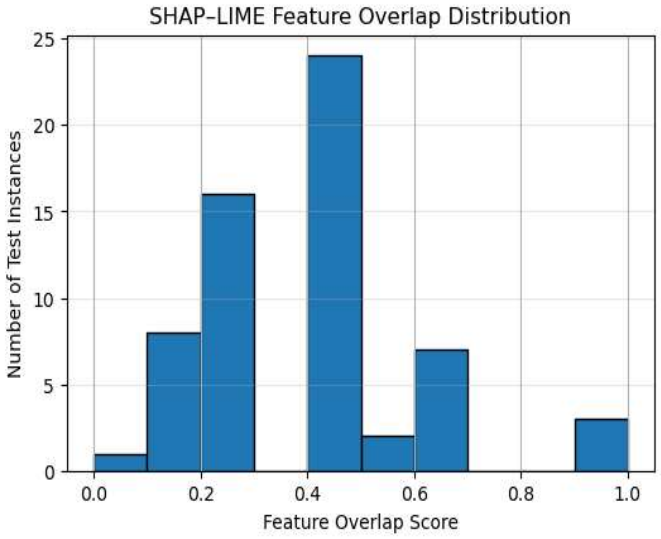


Figure 6

5) Case-Based Interpretation and Discussion:

Figure 7 outlines a case-based explanation, which incorporates SHAP attributes and nearest-neighbor similarity and outcome statistics of similar groups of patients. The method grounds the predictions on historically acquired patient profiles, which agrees with model results and clinician-style reasoning. Altogether, RF represents the most balanced score between the performance and interpretability, and the analysis of the consistency of the explanation proves the necessity of applying several interpretability tools to advance the distrust in the healthcare AI systems.

	Component	Explanation
0	Predicted Risk	Low
1	Top Risk Factors (SHAP)	thal, cp, ca, exang, sex
2	Similarity to Past Patients	-4086.8%
3	Disease Rate in Similar Patients	32.0%
4	SHAP–LIME Agreement	0.0

Figure 7

VI. CONCLUSION.

The present study proposed a explainable risk-prediction framework of heart diseases, balancing predictive performance and explainability by integrating both linear and non-linear machine learning models with explainable AI ways. Clinically relevant measures were used to assess the Logistic Regression, Random Forest, and Multi-layer perceptron models and demonstrated that random forest was the best model in overall performance and in terms of recall and ROC-AUC.

Additionally to predictive integrity, interpretability was also highlighted in the study; SHAP and LIME were used to combine global and local explanations. The consistency analysis of the explanation showed that these two methods had moderate agreement, which indicated that a single method of explainability cannot be reliable to support clinical decision making. Further, the suggested scheme of explanation based on cases made the prediction framework more interpretable as it provided a context by administering comparisons of similar patient profiles.

Future directions will be aimed at testing the proposed structure on a bigger and multi-institutional scale, with time and longitudinal patient information, as well as assessing clinician-in-the-loop assessment to further examine the explainability and reliability of explainable AI in a real clinical context.

VII. REFERENCES.

1. UCI Machine Learning Repository. Heart Disease Dataset. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. Detrano, R., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5), 304–310.
3. Alizadehsani, R., et al. (2013). A data mining approach for diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine*, 111(1), 52–61.
4. Ahmad, T., et al. (2018). Machine learning classification techniques for heart disease prediction. *2018 IEEE International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*.
5. Krittanawong, C., et al. (2017). Machine learning prediction in cardiovascular diseases. *Nature Reviews Cardiology*, 14(7), 418–429.
6. Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930.
7. Weng, S. F., et al. (2017). Can machine-learning improve cardiovascular risk prediction? *PLOS ONE*, 12(4).
8. Rajkomar, A., et al. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380, 1347–1358.
9. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD Conference*, 1135–1144.
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*.
12. Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI. *Nature Machine Intelligence*, 2, 252–258.
13. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
14. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31–57.
15. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions. *Nature Machine Intelligence*, 1(5), 206–215.
16. Molnar, C. (2022). *Interpretable Machine Learning*. 2nd Edition, Leanpub.
17. Guidotti, R., et al. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42.

18. Goldstein, A., et al. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65.
19. Van Calster, B., et al. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230.
20. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of ICML*, 625–632.
21. Biecek, P., & Burzykowski, T. (2021). *Explanatory Model Analysis*. CRC Press.
22. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence. IEEE *Transactions on Neural Networks and Learning Systems*, 32(11), 4793-4813.
23. Samek, W., et al. (2019). Explainable AI: interpreting, explaining and visualizing deep learning. *Springer*.
24. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832.
25. Ahmad, M. A., et al. (2018). Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics*.