Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach

---

Published by: Hossam Meshref

Associate Professor, Computer Science Department

College of Computers and Information Technology

Taif University, Taif, Saudi Arabia in,2019

INTRODUCTION:

The field of machine learning has witnessed remarkable progress, as its techniques have become increasingly popular and accessible. Applications span diverse domains, including face detection, system security, disease diagnosis, drug discovery, and other transformative areas that have significantly impacted modern lifestyles. Unlike traditional programming methods, machine learning models rely on learning patterns from training data rather than being explicitly programmed. These models then use inference to generate valuable predictions.

Prominent machine learning techniques, such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM), have proven to be highly effective for prediction tasks. However, they often face a critical limitation: their "black-box" nature. Once built, these models lack transparency, making it difficult to interpret the reasoning behind their predictions. Most machine learning models are trained on historical data to forecast future scenarios, but the inability to understand the rationale behind their outputs can pose significant challenges.

This lack of interpretability becomes particularly concerning in high-stakes domains, such as healthcare. For instance, a machine learning model may play a critical role in recommending a medical treatment or surgical procedure. In such cases, the recommendation must be not only accurate but also interpretable, as understanding the model's reasoning is crucial to avoid life-threatening consequences. Similarly, in fields like finance, stakeholders rely on prediction models to assess risks or choose investment plans, where transparent reasoning can prevent costly errors.

For machine learning to fully realize its potential in critical decision-making contexts, the ability to interpret and justify model predictions must be emphasized alongside accuracy. This study addresses these concerns by exploring methods to balance predictive performance with interpretability, particularly in applications like heart disease diagnosis.

Methodology:

During model evaluation, the confusion matrix played a crucial role in interpreting the results obtained in this research (see Table II). The True Positive (TP) value represents the number of patients who were correctly predicted to have heart disease. True Negative (TN) indicates the number of patients correctly predicted not to have heart disease. Conversely, False Positive (FP) refers to patients who did not have heart disease but were incorrectly predicted as having it. Similarly, False Negative (FN) represents patients who did have heart disease but were incorrectly predicted as not having it.

TABLE II. CONFUSION MATRIX STRUCTURE

(The confusion matrix structure would be shown here, but ensure it is formatted correctly for better readability.)

Using the definitions of TP, TN, FP, and FN, the accuracy of the models developed in this research was calculated using the following evaluation metric:

Accuracy measures the ratio of correctly classified instances (both positive and negative) to the total number of instances in the dataset. However, accuracy alone is not sufficient for comprehensive model evaluation, as it can be misleading in imbalanced datasets. Consequently, additional metrics such as Precision, Recall, and F1 Score were also used to assess the models:

Precision=TPTP+FP\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}Precision=TP+FPTP
Recall=TPTP+FN\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}Recall=TP+FNTP
F1 Score=2×Precision×RecallPrecision+Recall\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}F1 Score=2×Precision+RecallPrecision×Recall

Precision, also known as confidence, is the ratio of correctly predicted positive instances to the total predicted positive instances. Recall, or sensitivity, is the ratio of correctly predicted positive instances to the total actual positive instances. The F1 Score combines both Precision and Recall into a harmonic mean, providing a balanced evaluation metric that accounts for both false positives and false negatives.

By considering these evaluation metrics alongside accuracy, this study ensures a more robust and comprehensive assessment of the machine learning models.

Conclusion:

In this paper, we demonstrated a thorough analysis and deep understanding of the Cleveland heart disease dataset. Various machine learning classifiers were developed and evaluated to identify the most suitable model for heart disease diagnosis. However, as highlighted in the interpretation section, several critical issues must be considered to fully understand the performance of the machine learning models.

If the selection of the four models—MLP (Multi-Layer Perceptron), NB (Naïve Bayes), SVM (Support Vector Machine), and RF (Random Forest)—were based solely on traditional evaluation metrics such as accuracy, precision, recall, and F1 score, there could have been a risk of choosing an unsuitable model. For instance, the MLP model, which achieved an accuracy of 84.25% using an SVM-wrapping attribute selection method and an 8-feature set, initially seemed promising. However, its feature ranking score (FRC), calculated as 15 based on the 50% threshold used in this research, was three times higher than that of the RF model. This significant discrepancy indicated that selecting the MLP model solely based on accuracy would not have been appropriate for heart disease diagnosis.

The analysis conducted in this research provided a solid foundation for understanding the nature of the heart disease dataset. These efforts were complemented by the introduction of the Feature Ranking Cost (FRC) index, which enhanced the interpretability of the designed models. The FRC index served as an informative metric, enabling clear differentiation between the models based on the importance of their feature sets.

Ultimately, the RF model was chosen as the final diagnostic model. While its accuracy of 79.92% was slightly lower than the MLP model, this compromise was necessary to ensure a balance between transparency and accuracy. The RF model's selection underscores the importance of interpretability and trustworthiness in clinical decision-making, where understanding a model's rationale is as vital as its predictive performance.

The findings of this study contribute meaningfully to the machine learning community by providing a framework for post-hoc interpretability analysis of predictive models. This framework could serve as a foundation for future work involving clinical datasets, ensuring that both accuracy and transparency are adequately addressed.