Richa Sharma, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Comparative Analysis of Explainable AI Techniques for Interpreting Machine Learning Models

Richa Sharma[1], Archita Kumari[2], Archana Mongia[3]
[1]Department of Data Science and Artificial Intelligence,
HMRITM (of Guru Gobind Singh Indraprastha University), New Delhi, India
[2]Dept. of Electronics and Communication Engineering (Specialisation: AI),
Indira Gandhi Delhi Technical University for Women, New Delhi, India
[3]HMRITM (of Guru Gobind Singh Indraprastha University), New Delhi, India

Abstract- Breast cancer diagnostic machine learning models need to be not only accurate but also interpretable to be deployed clinically with any reasonable degree of confidence. Although Random Forest, XGBoost and Neural Networks generate accurate predictions, their respective "black box" nature of operation can undermine the clinical community's willingness to trust the output. Explainable AI supplement methods such as SHAP and LIME have been developed in part to address this issue. Herein, we compare the SHAP & LIME measured explainability consistency of three different classifying algorithms (Random Forest, XGBoost, and Neural Network) as they were trained on the Breast Cancer Wisconsin dataset. Each algorithm's accuracy, F1, and Receiver Operating Characteristic Area Under Curve (ROC-AUC) were calculated to evaluate model classifying performance. Spearman rank correlation was calculated to evaluate explainability consistency as assessed by SHAP & LIME. Neural Networks exhibited the highest classifying accuracy (96.49%) and F1-score (0.9718); however, XGBoost provided the greatest consistency between SHAP & LIME (ρ = 0.9534), followed by Neural Network (ρ = 0.9049) and Random Forest (ρ = 0.8889). XGBoost provides the best overall balance of predictive accuracy with explainability consistency.

Keywords: Neural Network, SHAP, LIME.

I. INTRODUCTION

Machine learning (ML) has grown dramatically over the last few years and continues to grow rapidly. As a result, there has been considerable progress in using machine learning in healthcare, especially for the early detection and classification of diseases that can result in death, such as breast cancer. Breast cancer is one of the most common cancers in the world, and early and accurate diagnosis provides the best chance of improving the survival of patients. The existence of structured clinical datasets (e.g., the Breast Cancer Wisconsin (Diagnostic) dataset) has built a great foundation to develop automated classification systems to distinguish between malignant and benign tumors accurately.

Several different machine learning classification algorithms, such as Random Forest [1], XGBoost [3], and deep neural networks [2], have been successfully used to predict the classification of tumours in these datasets with extraordinarily high predictive ability. Despite their accuracy, all of these machine learning classification models suffer from the "black box" problem, which makes it difficult to use these models in actual clinical settings that require interpretation of results.

With the growth of machine learning in healthcare, Explainable Artificial Intelligence (XAI) has emerged as an important field of study. LIME (Local Interpretable Model-Agnostic Explanations) [4] and SHAP (SHapley Additive exPlanations) [5] are two commonly used XAI techniques that provide locally interpretable approximations instead of large complex models. Beyond the application of SHAP and LIME, this research measures the alignment of these interpretability techniques through quantitative means by employing Spearman's rank correlation among different models, thus enhancing trustworthiness in AI-based healthcare systems. However, existing research focuses on the performance of the model or the interpretability of results, but lacks a quantitative approach that analyses the agreement between SHAP and LIME. Such an analysis is essential to evaluate the reliability of models in important areas such as healthcare. The main contributions of this paper include:

- Comparative study of Random Forest, XGBoost, and Neural Network models based on performance measures.
- Quantitative assessment of the consistency in SHAP and LIME interpretations through Spearman rank correlation.
- Examination of variations in interpretation consistency across different models.
- Discussion of the implications of consistent interpretation for reliable AI-based medical diagnostics.

II. LITERATURE REVIEW

The use of machine learning systems is greatly improving the ability to diagnose breast cancer due to their ability to accurately classify tumours as either benign or malignant. Unfortunately, due to a lack of understanding of how the more complex machine learning systems function, it has limited the successful implementation of these systems into clinical settings. As a result, a growing amount of research has been conducted on XAI (explainable artificial intelligence), where researchers attempt to provide an understanding of how these more complex systems arrive at specific decisions.

Breiman introduced the random forest technique in (1), which consists of combining several decision trees together to improve prediction accuracy while also providing an understanding of the importance of each feature. Although random forest does have some level of interpretability, it still may have issues with its feature importance being biased when there are features related to each other. LeCun et al [2] reported significant predictive capabilities for deep learning models while applying them to various applications. However, due to the black-box nature of deep learning models, it often makes it impossible for medical researchers to understand how the models produce their results.

XGBoost, proposed by Chen et al [3], is an optimised gradient boosting framework that is capable of producing outstanding performance when used with structured data. Despite the ability of XGBoost to produce predictions with high accuracy, the complexity of the XGBoost framework requires that post hoc explanation methods be applied to provide more transparency into how the model produced its output.

To increase the interpretability of machine learning models, Ribeiro et al [4] introduced local interpretable model-agnostic explanations (LIME). The purpose of LIME is to provide explanations for individual predictions generated by the model by representing the complex model using a simpler local model. Furthermore, LIME is computationally efficient, but the use of different sampling techniques can create variations in the sampling neighbourhoods, which may lead to LIME generating different and possibly inconsistent explanations. Lundberg et al [5] recently proposed the Shapley additive explanation (SHAP) framework to provide greater clarity on how models produce their predictions by providing scientific principles used to model the predicted outcomes produced from various inputs to the model.

III. RELATED WORK

Machine learning has recently advanced rapidly, leading to greatly improved predictive performance in many different fields. The fact that many complex models are not interpretable has increased the interest in explainable AI (XAI). One of the earliest works related to XAI was Leo Breiman's Random Forests in 2001 [1], which provides high-quality predictions while providing basic estimates of feature importance. The introduction of gradient boosting methods, such as XGBoost, has further increased predictive performance but also added more complexity and reduced interpretability. Yann LeCun et al. [2] discuss deep learning models that, by design, achieve extremely high levels of accuracy, but are very much black-box (opaque) models.

However, recent studies have reinforced the position of XAI in improving the interpretability of machine learning approaches. Felix Tempel et al. [20] indicate that the SHAP value approach provides better feature importance in terms of consistency and quantification compared to the visualisation-based Grad-CAM approach. Mubaraqah et al. [21] indicate that although XGBoost provides better predictive

accuracy than Random Forest, SHAP provides better explanations than LIME due to the variability in LIME's results. Ahmet Akgündŏdu et al. [22] emphasize the significance of explainability in clinical decision-making, noting that SHAP provides better trust in prediction results. The comprehensive survey by Yogesh K. Dwivedi et al. [23] indicates that evaluating feature importance consistency is one of the important issues in the context of XAI.

In their papers [11] and [13], comparisons of LIME and SHAP found that SHAP provides more consistent and stable feature importance than LIME, although LIME is considerably more computationally efficient when generating explanations.

Table 1. Comparative Analysis of Existing Studies on Explainable AI Techniques

Study	Dataset Type	Model Used	XAI Method	Eval. Metrics	Key Limitation
Paper 1 [1],[3]	Structured (tabular)	Random Forest, XGBoost	Built-in Feature Importance	Accuracy, ROC-AUC	Feature importance is biased and inconsistent
Paper 2 [2]	Large-scale datasets	Neural Networks	No XAI (Black-box)	Accuracy	Lack of interpretability
Paper 3 [5]	Structured + unstructured	All ML models	SHAP	Accuracy, ROC-AUC	High computational cost
Paper 4 [11],[13]	Financial / Credit dataset	XGBoost, Random Forest	SHAP, LIME	Accuracy, ROC-AUC	LIME inconsistency, SHAP slower
Proposed	Imbalanced (Medical)	RF, XGBoost, Neural Network	SHAP + LIME	Accuracy, ROC-AUC, F1-Score, Spearman ρ	Important consistency, but computational overhead

IV. METHODOLOGY

Overview

Figure 1 displays the overall workflow, which follows a pipeline of 5 steps: (1) data preprocessing, (2) model training, (3) SHAP-based global explanation, (4) LIME-based local explanation, and (5) consistency evaluation using Spearman rank correlation.

Dataset

The experiments were performed using the Breast Cancer Wisconsin (Diagnostics) dataset from Kaggle, containing 569 instances with 30 numerical characteristics of fine needle aspirate (FNA) images of breast masses. The dataset has two classes: Malignant (212) and Benign (357). The numerical characteristics are the mean, standard error, and worst values of cell nuclear characteristics, including radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension. The id and empty Unnamed: 32 columns were removed during data preprocessing. The target variable 'diagnosis' was mapped to M → 0 and B → 1.

Data Preprocessing

All 30 features were standardised to have mean = 0 and variance = 1 to prevent any one feature from dominating due to differences in scale. The data was split into 80% training data (455 samples) and 20% testing data (114 samples), using stratified sampling.

Machine Learning Models

Three different classifiers were used in this project:

- **Random Forest:** a group of 100 decision trees trained via bagging, with prediction by majority vote. Provides feature importance for interpretability.
- **XGBoost:** a gradient-boosted model consisting of 100 trees, where each tree corrects the mistakes of the previous tree with respect to the log-loss function. Known to perform well on tabular data.
- **Neural Network (MLP):** a multi-layer perceptron with two hidden layers (64 and 32 neurons), using ReLU activation, trained for 500

epochs. Serves as a black-box comparison model.

Explanation Methods

SHAP (SHapley Additive exPlanations). Features are assigned importance values using Shapley values. TreeExplainer was used for Random Forest and XGBoost, providing exact Shapley values. KernelExplainer was used for the Neural Network (100 background examples). The average absolute SHAP value of the test dataset was used to compute overall feature importance.

LIME (Local Interpretable Model-Agnostic Explanations). For each of 30 randomly chosen test samples, LIME builds a local linear model in the neighborhood of the chosen sample. The top 10 features by absolute weight are recorded per sample. Global feature importance is computed by averaging absolute LIME weights across all 30 samples.

Consistency Evaluation

The Spearman rank correlation coefficient (ρ) of global importance vectors was computed to measure how closely SHAP and LIME agree. The two-tailed p-value was calculated using the t-distribution with n−2 degrees of freedom. A ρ value close to 1 indicates strong agreement; a low ρ value shows divergence.

Sensitivity Analysis

The sensitivity analysis was carried out through changes in LIME and SHAP's critical parameters for the Random Forest model, which offers better interpretability for such analysis. Different sample sizes in the case of LIME and different settings in the case of SHAP were chosen to assess the stability of feature importance rankings using Spearman rank correlation.

Evaluation Metrics

Model performance is evaluated using Accuracy, F1-Score, and ROC-AUC. Accuracy is the fraction of correctly classified instances. F1-Score is the harmonic mean of precision and recall. ROC-AUC is the area under the Receiver Operating Characteristic curve. Explainability is evaluated using the Spearman ρ consistency score between SHAP and LIME feature importance rankings.

V. RESULTS

Model Performance

As shown in Table 2, all three models exhibit good classification capabilities. The Neural Network achieved the highest accuracy (0.9649) and F1-score (0.9718). Random Forest and XGBoost showed similar accuracy (0.9561), with XGBoost performing slightly better in terms of F1-score (0.9660 vs. 0.9655). The Random Forest classifier showed a higher ROC-AUC value (0.9939). These findings suggest that performance metrics differ across models and there is no single universal best model across all metrics.

Table 2. Model Performance Comparison on Breast Cancer Dataset

Model	Accuracy	F1-Score	ROC-AUC
Random Forest	0.9561	0.9655	0.9939
XGBoost	0.9561	0.9660	0.9901
Neural Network	0.9649	0.9718	0.9937

SHAP Analysis

The importance of various attributes on a global level was calculated using SHAP values. Exact Shapley values were computed for Random Forest and XGBoost using TreeExplainer, while Kernel SHAP was used for the Neural Network (dataset size: 100). Across all models, the morphological attributes with the most adverse impact — concave points_worst, perimeter_worst, and radius_worst — consistently played a crucial role, in line with conventional clinical indicators of malignancy.

LIME Analysis

All models were analyzed on 30 randomly chosen test cases using LimeTabularExplainer. Feature weights were computed by averaging local feature weights across samples. The worst-category features were consistently selected by LIME across all models,

with slight differences in local explanations observed for the Neural Network model.

Feature Importance Consistency

Table 3 presents the Spearman rank correlation coefficients between SHAP and LIME feature importance rankings. All three models showed statistically significant agreement ($p < 0.001$). XGBoost showed the highest consistency ($\rho = 0.9534$), followed by Neural Network ($\rho = 0.9049$) and Random Forest ($\rho = 0.8889$).

Table 3. SHAP–LIME Feature Importance Consistency

Model	Spearman ρ	p-value	Agreement
Random Forest	0.8889	< 0.001	Strong
XGBoost	0.9534	< 0.001	Very Strong
Neural Network	0.9049	< 0.001	Strong

Balance of Accuracy versus Interpretability

Table 4 displays the balance of accuracy and interpretability of all three models. The Neural Network achieved the highest accuracy but also the highest computational explanation cost due to Kernel SHAP. XGBoost provided the best balance — high accuracy (95.61%) and very high SHAP-LIME consistency ($\rho = 0.9534$) — at low explanation cost via TreeExplainer, making it the best candidate for clinical adoption.

Table 4. Accuracy vs. Interpretability Summary

Model	Accuracy	SHAP – LIME ρ	Explanation Cost	Recommendation
Random Forest	0.9561	0.8889	Low	Interpretable baseline
XGBoost	0.9561	0.9534	Low	Best overall balance
Neural Network	0.9649	0.9049	High	Maximum accuracy

Sensitivity Analysis

As shown in Table 5, SHAP offers better stability ($\rho = 0.9835$) compared to LIME ($\rho = 0.85$) under parameter variation. This confirms that SHAP is more reliable for feature explanations when hyperparameter sensitivity is a concern.

Table 5. Sensitivity Analysis Results (Random Forest Model)

Method	Parameter Variation	Spearman ρ	Stability
LIME	1000 vs 5000 perturbation samples	0.8500	Moderate stability
SHAP	2 vs 10 test instances averaged	0.9835	Very stable

Key Findings

- XGBoost has the best balance of accuracy and interpretability for clinical use cases.
- The high correlation (> 0.88) between SHAP and LIME validates the reliability of both methods; SHAP is more stable than LIME under parameter variation.
- The most important features (concave points_worst, perimeter_worst, radius_worst) were consistently identified across all models.
- Neural Network results can be interpreted using SHAP values, showing good correlation with LIME values (> 0.90).
- Tree-based models are faster to explain via TreeExplainer compared to KernelSHAP for neural networks.

VI. DISCUSSION

Clinical Interpretability and Trustworthiness

Results show that in health care, explainability is critical when deploying any machine learning model. While all classifiers have good predictive abilities, what matters most relates to the ability of the clinician to interpret and trust their machine learning model's predictions.

Overall, the neural network model showed the highest performance (accuracy: 96.49%; F1-score: 0.9718), which reflects the advantages deep learning models have in terms of learning capability. On the other hand, because the explanations provided by the neural network are based on KernelSHAP, which is expensive in regard to compute time and slower than TreeExplainer, the neural network's ability to be

used for real-time clinical decision support systems where fast explanations provided will determine how practical this model may be.

Another model, Random Forest, was able to provide comparable classification performance and, when compared to XGBoost, a lower execution time; nevertheless, Random Forest was able to provide comparable interpretability with its prediction output. Consequently, there is good agreement (ρ = 0.8889) between SHAP-LIME correlation distributions when using either explanation method regarding Random Forest. Furthermore, the collaboration between TreeExplainer and Random Forest provides an efficient, practical, and interpretable model baseline.

Finally, for all three classifiers used in this study, XGBoost had the best overall tradeoff between predictive performance and interpretability. While inferior accuracy (95.61%) was found in the XGBoost model when compared to the neural network, the SHAP-LIME correlation (ρ = 0.9534) between XGBoost is higher than that of the neural network and Random Forest; thus, XGBoost has a relatively low-cost value for generating explanations. Hence, this suggests that XGBoost, as a model to be validated and trusted, also provides an explanation of output predictions in a more stable and consistent manner than the neural network.

Limitations & Future Scope

Although the potential of this research study appears to be positive, several limitations exist within the body of work. First, because the Breast Cancer Wisconsin Diagnostic dataset was the only medical dataset used in the experimental setup, this may limit the generalizability of the study's conclusions to other datasets/diseases. In addition, only three machine learning classifier types (Random Forest, Support Vector Machine, and k-Nearest Neighbour) as well as two techniques for explainable artificial intelligence (XAI) (i.e., SHapley Additive exPlanation (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME)) were utilised in this research, while many of the latest XAI techniques [e.g., Integrated Gradients, Gradient Class Activation Mapping (Grad-CAM), Anchors and Counterfactual Explanations] were not included in these experiments. The only classifier to undergo sensitivity analysis was the Random Forest classifier; performing a sensitivity analysis on all classifiers (and their respective hyperparameters) may provide additional information on the robustness of the explanation. Further, the discussion of the computational efficiency aspect of the research was qualitative rather than based on an evaluation of the actual execution time and resource usage.

Future work can evaluate our proposed model on various medical datasets to increase the robustness and generalizability of the results. Other machine learning approaches, including transformer-based architectures, convolutional neural networks, and ensemble deep learning methods, could be considered as well. Including advanced models of explainability, such as Integrated Gradients, Grad-CAM, Counterfactual Explanations, and causal explainability, would allow for a better comparative analysis of explainable AI methods. It will also consider the consistency of explanation in federated and private-based healthcare systems, where data cannot be shared. Evaluating computational performance and stability of explanation during failure, and clinician usage to capture user experiences, will support the practical use of explainable AI. Ultimately, incorporating accurate and credible explainable machine learning models into real-time clinical decision support systems can significantly increase the transparency, reliability, and patient-centeredness of the healthcare system.

VII. CONCLUSION

Using the Breast Cancer Wisconsin Diagnostic dataset, this work compared SHAP and LIME on three machine learning classifiers: Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP) Neural Network. The MLP had the best accuracy (96.49%), F1-score (0.9718), and ROC-AUC (0.9937); whereas Random Forest and XGBoost had the ROC-AUC values of 0.9939 and 0.9901, respectively. XGBoost had the best overall consistency between SHAP–LIME (ρ = 0.9534), followed by MLP (ρ = 0.9049) and Random Forest (ρ = 0.8889). The correlation between SHAP and LIME was the strongest using

XGBoost (ρ = 0.9534). The second-highest correlation was recorded using MLP (ρ = 0.9049) and Random Forest (ρ = 0.8889). All correlations were statistically significant ($p < 0.001$). The morphological worst features — concave points_worst, perimeter_worst, and radius_worst — showed high predictive ability across all classifiers.

This work addresses three critical gaps in the literature. First, by introducing Spearman's rank correlation as a quantitative consistency metric between SHAP and LIME, it provides an objective and reproducible framework for evaluating explainability reliability.

Second, the results empirically demonstrate that high predictive accuracy and strong explainability consistency are not mutually exclusive, with XGBoost emerging as the optimal model for clinical deployment. Third, the inclusion of Neural Networks reveals that even inherently opaque models can achieve high SHAP–LIME consistency (ρ = 0.9049) when appropriate post-hoc explanation methods are applied, though at significantly higher computational cost. Future research will investigate gradient-based XAI methods such as Integrated Gradients, explore attention-based architectures, and extend the consistency evaluation framework to multi-modal and unstructured clinical datasets.

REFERENCES

1. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
2. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, 2015.
3. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
4. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016.
5. S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. 31st Conf. Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017.
6. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. 34th Int. Conf. Machine Learning (ICML), Sydney, Australia, 2017.
7. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017.
8. R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, 2019.
9. C. Molnar, Interpretable Machine Learning. Lulu Press, 2020.
10. X. Zhao et al., "BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations," arXiv preprint arXiv:2012.03058, 2020.
11. A. Gramegna and P. Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," Frontiers in Artificial Intelligence, vol. 4, 2021.
12. B. H. M. van der Velden et al., "Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis," Medical Image Analysis, vol. 79, 2022.
13. A. Salih et al., "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," arXiv preprint arXiv:2305.02012, 2023.
14. S. J. Silva et al., "Explainable Machine Learning Using SHAP for Environmental Modeling," Journal of Advances in Modeling Earth Systems (JAMES), 2022.
15. H. Zhang et al., "Explainable Artificial Intelligence Using XGBoost and SHAP for Risk Prediction," International Journal of Disaster Risk Reduction, 2025.
16. I. Givisis et al., "Comparing Explainable Artificial Intelligence Models: SHAP and LIME," Electronics, vol. 14, no. 23, 2025.
17. M. Panda and S. Mahanta, "Explainable Artificial Intelligence for Healthcare Using Random Forest with SHAP and LIME," arXiv preprint arXiv:2311.05665, 2023.
18. K. Devireddy, "A Comparative Study of Explainable AI Methods," arXiv preprint arXiv:2504.04276, 2025.

19. S. Shreya and H. Pathak, "Explainable AI Credit Risk Assessment Using Machine Learning," arXiv preprint arXiv:2506.19383, 2025.
20. F. Tempel et al., "Comparison of SHAP and Grad-CAM Methods," arXiv preprint arXiv:2412.16003, 2024.
21. N. Mubaraqah et al., "Comparison of Random Forest and XGBoost with SHAP and LIME Interpretation," Journal of Technology and Engineering Research and Applications (JTERA), 2024.
22. A. Akgündoğdu et al., "Explainable AI-Based Brain Tumour Detection Using Machine Learning Models," Applied Sciences, 2025.
23. Y. K. Dwivedi et al., "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," ACM Computing Surveys, vol. 55, no. 9, 2023.
24. I. Kakogeorgiou and K. Karantzalos, "Evaluating Explainable Artificial Intelligence Methods for Multi-Label Deep Learning Classification Tasks in Remote Sensing," International Journal of Applied Earth Observation and Geoinformation, 2021.
25. Zhou et al., "Interpretable Landslide Susceptibility Modelling Using SHAP Values and the XGBoost Algorithm," Geocarto International, 2022.