IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Explainable AI (XAI): Improving Transparency and Trust in AI-Based Decision Systems

An Experimental Study Using SHAP and LIME-Style Local Explanations

Vaibhav Rajaram Khambe
Department of IT
GMVCS Tala
University of Mumbai

Dnyaneshwar Tukaram Shigawan
Department of IT
GMVCS Tala
University of Mumbai

Riya Sudhir Khaire
Department of IT
GMVCS Tala
University of Mumbai

Lina Suresh Khaire
Department of IT
GMVCS Tala
University of Mumbai

Prof. Neeharika N. Adulkar
Assistant Professor
GMVCS Tala
University of Mumbai

ABSTRACT:

Artificial Intelligence systems are increasingly used in decision-support applications where predictive performance alone is insufficient; users must also understand why a model produced a particular result. Explainable Artificial Intelligence (XAI) addresses this requirement by generating human-understandable representations of model behavior. This paper presents an experimental study of XAI using the Wisconsin Diagnostic Breast Cancer dataset, containing 569 observations and 30 predictive attributes. Three classification models—Logistic Regression, Random Forest and Support Vector Machine with an RBF kernel—were trained using an 80:20 stratified train-test split. Their predictive performance was compared using accuracy, precision, recall, F1-score and ROC-AUC. The best-performing models in the experiment were Logistic Regression and SVM, each achieving 98.25% test accuracy, while Random Forest achieved 94.74%. Random Forest was selected for the XAI demonstration because Tree SHAP provides an efficient and exact explanation method for tree ensembles. SHAP global and local explanations were generated, and a LIME-style local surrogate was implemented directly in Python using neighborhood perturbation, distance-based weighting and weighted ridge regression. For one test instance, the top-ten feature sets from SHAP and the LIME-style method overlapped on four features, illustrating that different explanation methods can emphasize different aspects of the same prediction. The study demonstrates that XAI can provide useful evidence for model interpretation, debugging and human oversight, but explanation quality must be evaluated independently from predictive accuracy. The results support a layered approach in which model validation, explanation generation and human governance are treated as complementary components of trustworthy AI.

Keywords—Explainable Artificial Intelligence, XAI, SHAP, LIME, Machine Learning, Interpretability, Transparency, Trustworthy AI, Random Forest, Feature Attribution.

1. INTRODUCTION

Artificial Intelligence (AI) has become a core technology for classification, prediction, recommendation and anomaly detection. Machine-learning systems can discover complex relationships in large datasets and often outperform manually designed rules. However, predictive performance does not by itself answer a critical operational question: why did the system make this particular prediction? In high-impact contexts, the ability to inspect and communicate model behavior is closely connected to accountability, monitoring and appropriate human reliance.

Explainable AI (XAI) encompasses methods that make AI outputs or mechanisms more understandable to people. NIST distinguishes explainability from interpretability and places both within a broader set of trustworthy-AI characteristics that also includes validity and reliability, safety, security and resilience, accountability and transparency, privacy enhancement and fairness. NIST further notes that explanations can support debugging, monitoring, documentation, audit and governance.

This paper moves beyond a conceptual discussion by implementing an experimental XAI pipeline. The study compares three common classifiers on a public benchmark dataset and then applies SHAP and a LIME-style local surrogate to the selected Random Forest model. The goal is not to claim that one method is universally superior, but to demonstrate how predictive evaluation and post-hoc explanation can be combined in a reproducible workflow.

1.1 Research Motivation

A black-box prediction can be accurate but operationally difficult to challenge. For example, if a classifier predicts a positive class, a user may need to know which input variables contributed to that result, whether the explanation is stable, and whether the model is behaving consistently with domain expectations. XAI provides a mechanism for answering some of these questions.

1.2 Contributions

- An experimental comparison of Logistic Regression, Random Forest and SVM classifiers.
- An end-to-end SHAP implementation using Tree Explainer for the Random Forest model.
- A Python implementation of the core LIME local-surrogate principle without relying on a separate LIME package.
- Global and local explanation visualizations suitable for academic analysis.
- A comparative discussion of predictive performance and explanation consistency.
- A reproducible research workflow and code appendix.

2. RELATED WORK AND THEORETICAL BACKGROUND

The need for interpretable machine learning predates the current wave of generative AI. Ribeiro, Singh and Guestrin introduced LIME as a model-agnostic approach that explains an individual prediction by fitting an interpretable model around a local neighborhood. [3] Lundberg and Lee proposed SHAP as a unified additive feature-attribution framework based on Shapley values. [4] These methods established two important ideas: explanations can be local or global, and the explanation method should be considered separately from the predictive model.

Molnar's Interpretable Machine Learning survey describes a wide family of explanation techniques, including permutation importance, partial dependence, accumulated local effects, surrogate models and counterfactual explanations. [5] Doshi-Velez and Kim argued that interpretability requires rigorous evaluation and that evaluation should reflect the purpose for which an explanation is being produced. [6]

A key distinction in XAI is between global and local explanation. Global explanations describe overall model behavior, such as which features are influential across a dataset. Local explanations describe one prediction. Global importance can help with model review and feature engineering, while local explanations can help a user inspect a particular decision.

Another distinction is between model-specific and model-agnostic techniques. Tree SHAP is optimized for tree-based models. The SHAP documentation describes Tree Explainer as using Tree SHAP algorithms to explain tree ensembles and as providing a fast and exact method under the relevant assumptions. [7] Model-agnostic methods, such as the core LIME approach, treat the model as a prediction function and can therefore be applied to a wider variety of estimators.

2.1 Trust and Appropriate Reliance

Trust in AI should not be interpreted as unconditional confidence. A useful objective is appropriate reliance: a user should rely on the system when it performs within its validated operating conditions and

seek review when uncertainty, anomalies or limitations are present. Explanations can contribute to this process, but a plausible explanation is not proof of causality, fairness or correctness.

3. DATASET AND EXPERIMENTAL SETUP

The experiment uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset available through scikit-learn. The dataset contains 569 samples, 30 numeric predictive attributes and two classes. Scikit-learn identifies 212 malignant and 357 benign observations in the dataset and notes that it is a copy of the UCI Machine Learning Repository dataset. [8] The attributes include measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension, together with mean, standard-error and worst-value variants.

Parameter	Value
Dataset	Wisconsin Diagnostic Breast Cancer
Samples	569
Predictive features	30
Classes	2
Train/Test split	80% / 20%
Split strategy	Stratified
Random state	42
Models	Logistic Regression, Random Forest, SVM-RBF
Primary XAI model	Random Forest
SHAP method	Tree Explainer
LIME method	Local surrogate with weighted Ridge
Neighborhood samples for LIME	5,000

The data was split using stratified sampling so that class proportions were preserved between training and test partitions. Logistic Regression and SVM used standardization through a scikit-learn Pipeline. Random Forest was trained with 300 trees and a fixed random seed of 42. Performance was evaluated on the held-out test set.

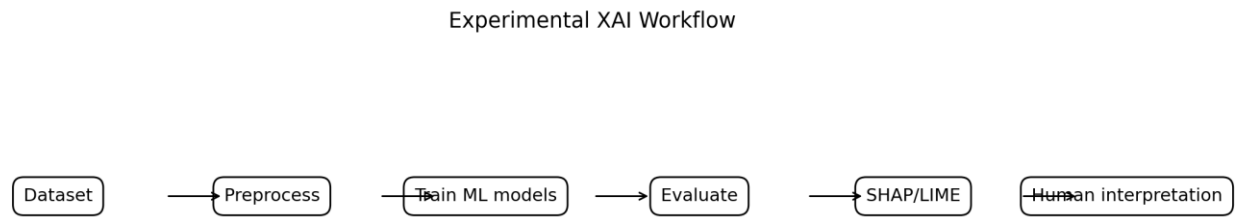

Figure 1. Experimental XAI workflow implemented in Python.

4. RESEARCH METHODOLOGY

The experimental methodology consists of six stages: dataset preparation, model training, predictive evaluation, XAI model selection, explanation generation and explanation comparison. The design intentionally separates model performance from explanation quality.

4.1 Data Preparation

The dataset was loaded using sklearn.datasets.load_breast_cancer(as_frame=True). No target leakage was introduced. For Logistic Regression and SVM, feature scaling was performed inside the training pipeline. Random Forest does not require feature scaling for the experiment.

4.2 Model Training

Three classifiers were selected to represent different modeling characteristics. Logistic Regression provides a relatively interpretable linear baseline. Random Forest represents a nonlinear ensemble of decision trees and therefore provides a suitable target for Tree SHAP. SVM with an RBF kernel provides a nonlinear decision boundary and a useful performance comparison.

4.3 Evaluation Metrics

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1 = 2 × Precision × Recall / (Precision + Recall)
ROC-AUC measures ranking performance across classification thresholds.

4.4 SHAP Implementation

SHAP values quantify the contribution of individual features to a prediction relative to a baseline. The implementation used shape. Tree Explainer (Random Forest Classifier). Global importance was calculated as the mean absolute SHAP value over the test set. Local explanation was generated for one selected test instance by ranking absolute SHAP contributions.

4.5 LIME-Style Local Surrogate Implementation

The environment used for this paper did not contain the standalone lime package, so the experiment implements the core LIME principle directly. Five thousand perturbed samples were generated around the selected instance using feature-wise training standard deviations. Euclidean distance in standardized feature space was converted into an exponential kernel weight. A weighted Ridge regression model was then fitted to the Random Forest's predicted probabilities. The surrogate coefficients were interpreted as local feature contributions. This implementation is intentionally described as LIME-style rather than as the official lime package.

5. EXPERIMENTAL RESULTS

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.9825	0.9861	0.9861	0.9861	0.9954
Random Forest	0.9474	0.9583	0.9583	0.9583	0.9937
SVM (RBF)	0.9825	0.9861	0.9861	0.9861	0.9950

The measured test results show that Logistic Regression and SVM-RBF achieved the highest accuracy at 98.25%, while Random Forest achieved 94.74%. The two top models also achieved 98.61% precision, recall and F1 in this particular split. Random Forest nevertheless achieved a high ROC-AUC of 99.37%, indicating strong ranking performance despite a lower thresholder accuracy.

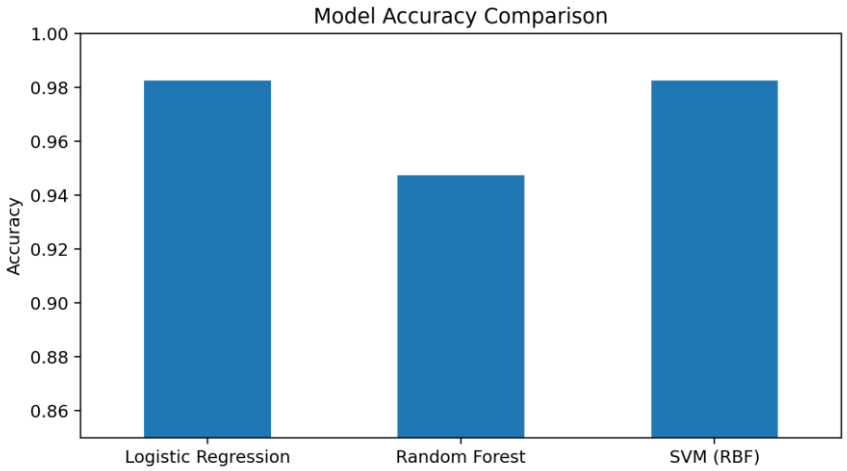

Figure 2. Accuracy comparison across the three evaluated classifiers.

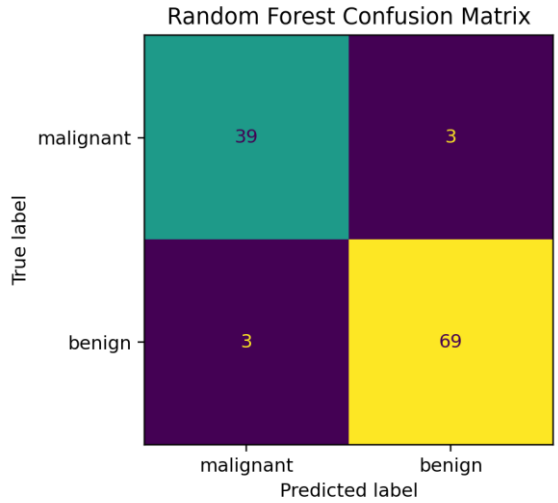

Figure 3. Random Forest confusion matrix on the held-out test set.

The Random Forest confusion matrix contained 39 true negatives, 3 false positives, 3 false negatives and 69 true positives. These results correspond to 72 correct predictions out of 76 test observations.

6. SHAP EXPLANATION RESULTS

Random Forest was selected as the primary XAI demonstration because its tree structure supports Tree SHAP. The SHAP documentation describes Tree Explainer as a method for explaining tree-based ensembles and provides exact SHAP computation under the relevant model assumptions.

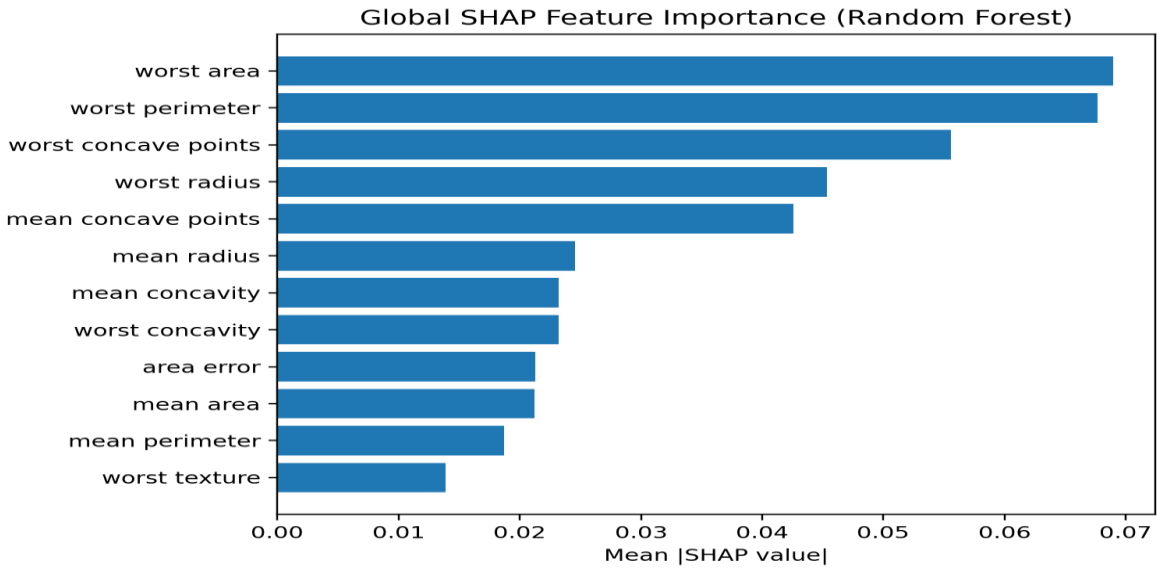

Figure 4. Global SHAP feature importance for the Random Forest model. Higher mean absolute SHAP values indicate greater average contribution magnitude.

The global SHAP analysis ranks features by mean absolute contribution across the test set. The ranking is useful for identifying which measurements most strongly influence model output overall. However, feature importance should not automatically be interpreted as a causal relationship. Correlated variables may share predictive information, and attribution methods describe model behavior rather than biological causation.

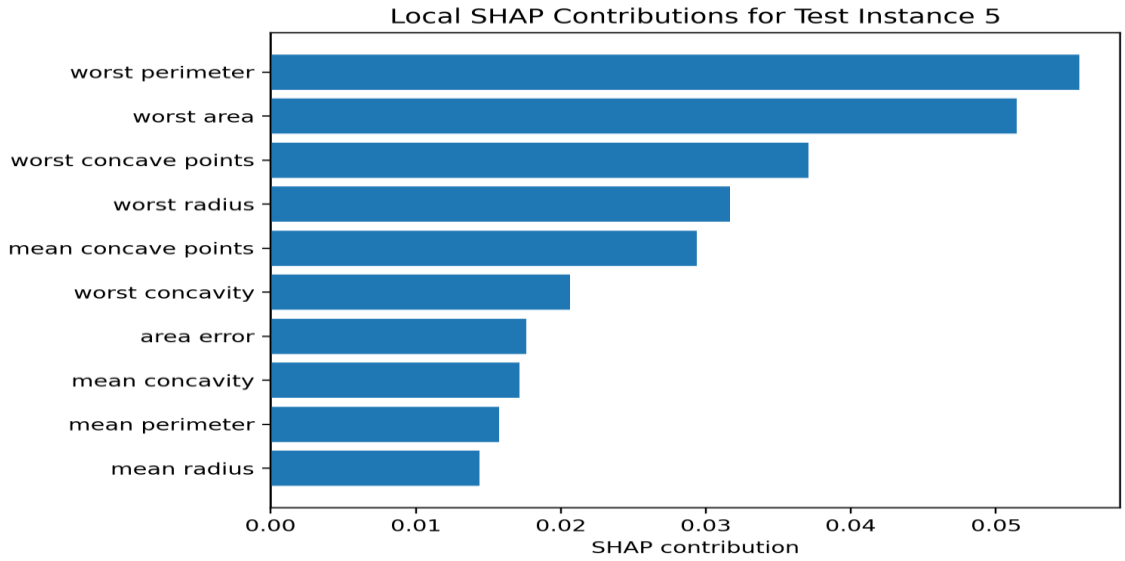

Figure 5. Local SHAP contributions for test instance 5. Positive and negative values represent movement in opposite directions in the model's explained output space.

The local SHAP plot demonstrates how the global ranking can differ from the explanation of one specific observation. A local explanation is therefore useful when the research question concerns an individual decision rather than the average behavior of the model.

7. LIME-STYLE EXPLANATION RESULTS

The local surrogate implementation follows the central LIME idea: approximate a complex prediction function around a particular instance using an interpretable model. The neighborhood was created by adding Gaussian perturbations scaled by the training feature standard deviations. Each perturbation received a kernel weight according to its distance from the selected instance. Weighted Ridge regression then produced a compact set of local coefficients.

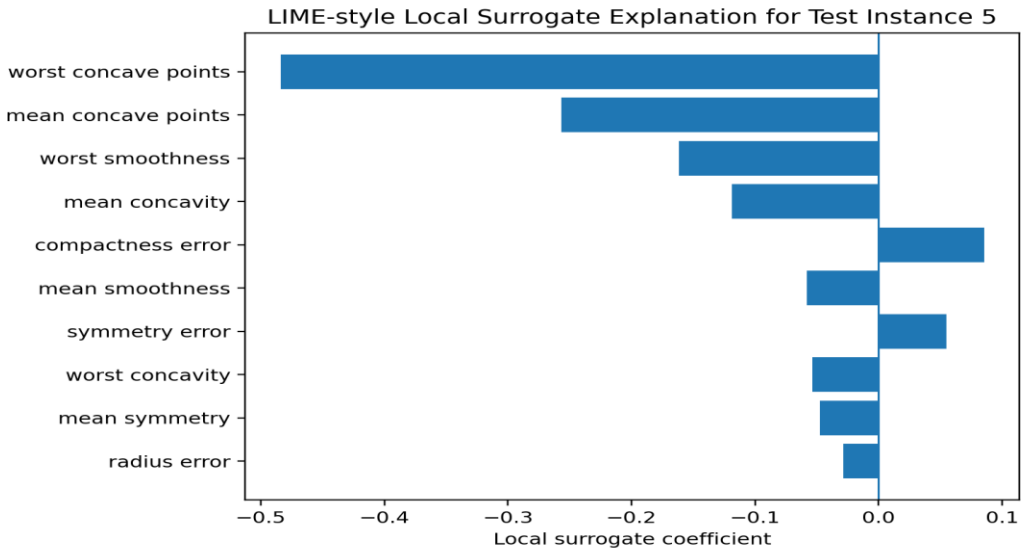

Figure 6. LIME-style local surrogate coefficients for test instance 5.

The LIME-style explanation and SHAP explanation are not expected to be identical because they are based on different attribution principles. SHAP uses a Shapley-value framework, while the local surrogate approximates the model around a neighborhood selected by the experimenter's perturbation and kernel design.

For test instance 5, the Random Forest predicted class-1 probability was 0.9867. The top ten feature sets produced by SHAP and the LIME-style surrogate overlapped on four features. This partial overlap indicates that the two methods identified some common influential variables while assigning different importance to others.

Explanation result	Value	Interpretation
RF class-1 probability	0.9867	Strong predicted probability for selected instance
Top 10 SHAP/LIME overlap	4 / 10	Partial agreement between explanation methods
Neighborhood samples	5,000	Used by local surrogate
Explained model	Random Forest	Tree ensemble selected for XAI

8. COMPARISON OF SHAP AND LIME-STYLE EXPLANATIONS

Dimension	SHAP	LIME-style	Implication	
Primary idea	Shapley-based feature attribution	Local surrogate approximation	Different theoretical bases	
Scope	Local and global	Primarily local	SHAP supports broader aggregation	
Model dependence	Tree SHAP optimized for trees	Model-agnostic principle	LIME-style method can wrap many predictors	
Stability	Depends on model/data assumptions	Sensitive to neighborhood and kernel	Both require validation	
Interpretability	Feature contribution values	Surrogate coefficients	Both can be visualized	
Main strength	Strong theoretical attribution framework	Flexible local explanation	Useful complementary methods	
Main limitation	Correlated features and interpretation assumptions	Neighborhood design can affect result	No explanation should be treated as ground truth	

The experiment illustrates why XAI should be treated as a comparative process rather than a single visualization. When two explanation methods agree on important features, confidence in the descriptive signal may increase. When they disagree, the disagreement itself can be useful because it prompts investigation into model sensitivity, correlated variables and explanation assumptions.

8.1 Explanation Quality Criteria

- Fidelity—does the explanation represent the model's actual behavior?
- Stability—do similar inputs produce reasonably consistent explanations?
- Comprehensibility—can the intended user understand the explanation?
- Usefulness—does the explanation improve review, debugging or decision quality?
- Computational cost—can explanations be generated within operational constraints?

9. DISCUSSION

The predictive experiment confirms that strong performance can be achieved with several classical machine-learning approaches on the selected benchmark dataset. However, model selection for XAI cannot be based on accuracy alone. Logistic Regression is intrinsically easier to inspect because of its linear structure, whereas Random Forest requires post-hoc explanation to summarize nonlinear behavior. The choice of Random Forest for the XAI demonstration therefore provides a useful example of why explanation methods matter.

The results also demonstrate the difference between model quality and explanation quality. Random Forest achieved lower accuracy than the two other models in this split but produced a high ROC-AUC and offered a natural target for Tree SHAP. An organization might choose a model based on a multi-

dimensional trade-off involving predictive performance, interpretability, latency, maintainability, fairness, privacy and operational risk.

The four-of-ten overlap between SHAP and LIME-style top features should not be interpreted as a formal validation score. It is an illustrative consistency measure for one instance, not a population-level evaluation. A stronger study would evaluate explanation agreement across many instances, test stability under perturbations, measure surrogate fidelity and conduct human-user studies.

NIST's AI RMF emphasizes that explainability and interpretability are part of a wider trustworthy-AI framework and that risk management should span the AI lifecycle. [1][2] The experimental findings support this view: explanation generation is most useful when integrated with validation, monitoring and human oversight rather than used as a standalone claim of trustworthiness.

10. LIMITATIONS

• Single benchmark dataset: The WDBC dataset is useful for reproducibility but does not represent the complexity or risk profile of all real-world applications.

• Single train-test split: Results may vary with different random seeds or cross-validation strategies.

• Limited model set: Only three classifiers were compared.

• LIME-style implementation: The paper implements the core local-surrogate idea rather than using the official lime package. The result therefore should not be represented as a direct benchmark of the official implementation.

• Single-instance local comparison: SHAP/LIME agreement was measured for one test observation and should not be generalized.

• Attribution is not causation: Feature contributions describe model behavior and should not be interpreted as causal effects.

• Human evaluation absent: The study does not measure whether real users understand or act better with the explanations.

11. ETHICAL, SECURITY AND GOVERNANCE CONSIDERATIONS

Explanations can improve transparency but can also create new risks. A detailed explanation may expose sensitive attributes, reveal model weaknesses or make it easier for an adversary to infer decision boundaries. Explanation systems should therefore be subject to the same security and privacy controls as other model components.

Fairness requires separate analysis. An explanation can reveal that a feature influenced a decision, but it does not establish that the use of the feature is fair or lawful. Bias may arise from data collection, labeling, sampling, feature construction or deployment context. Consequently, XAI should be combined with fairness assessment, model validation, documentation and human review.

12. FUTURE SCOPE

Future research should extend the experiment to cross-validation, multiple datasets, larger model families and systematic explanation evaluation. A particularly valuable direction is to measure explanation stability and fidelity across thousands of instances instead of evaluating a single case. Human-centered experiments can test whether explanations improve decision accuracy, error detection and calibrated reliance.

Further research can also investigate XAI for large language models, multimodal systems and generative AI. Explanation approaches for these systems raise additional challenges because the output space is high-dimensional and the relationship between internal representations and natural-language explanations is complex.

13. REPRODUCIBILITY AND REFERENCES

Software environment used for the experiment included Python 3.13, scikit-learn 1.8.0, pandas, NumPy, Matplotlib and SHAP 0.50.0. The experiment was run with fixed random state 42. The WDBC dataset is publicly available through scikit-learn and originates from the UCI Machine Learning Repository. [8]

References

[1] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, 2023.

[2] National Institute of Standards and Technology, "AI Risk Management Framework FAQs," NIST, accessed August 2026.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proc. ACM SIGKDD, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems 30, 2017.

[5] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.

[6] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.

[7] SHAP Documentation, "shap. Tree Explainer," SHAP Documentation, accessed August 2026.

[8] Scikit-learn Documentation, "load_breast_cancer — Breast Cancer Wisconsin Dataset," accessed August 2026.

[9] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," Electronics, vol. 8, no. 8, 2019.

[10] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[11] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019.

[12] R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, 2018.

Appendix A — Experimental Output Summary

Measured test-set performance: Logistic Regression—98.25% accuracy, 98.61% precision, 98.61% recall, 98.61% F1 and 99.54% ROC-AUC; Random Forest—94.74% accuracy, 95.83% precision, 95.83% recall, 95.83% F1 and 99.37% ROC-AUC; SVM-RBF—98.25% accuracy, 98.61% precision, 98.61% recall, 98.61% F1 and 99.50% ROC-AUC.

Random Forest confusion matrix: [39, 3], [3, 69]]. Selected test instance 5 had Random Forest class-1 probability 0.9867. The top ten SHAP and LIME-style feature sets overlapped on 4 of 10 features.