



Interpretable Surrogate Models for Explaining Black-Box Predictions in Data Science

Received: June 30, 2026

Revised: September 09, 2026

Accepted: September 27, 2026

Publish: October 04, 2026

Jaesik Jeong*, Lusiana Efrizoni, Susi Erlinda

Abstract:

Background: Black-box machine-learning models often achieve strong predictive performance but provide limited transparency. Interpretable surrogate models approximate these models using simpler structures, although their explanations may not always be sufficiently faithful or stable.

Aims: This study evaluates global and local surrogate models for explaining black-box predictions and examines the trade-offs among fidelity, stability, complexity, sparsity, and computational efficiency.

Methods: Experiments were conducted on four public binary-classification datasets: Adult Income, Bank Marketing, Breast Cancer Wisconsin Diagnostic, and German Credit. Random Forest, XGBoost, and Multilayer Perceptron served as black-box classifiers. Their predictions were explained using global logistic regression, global decision tree, rule-based, local linear, and local decision tree surrogates. Each experiment was repeated using five random seeds. Explanation quality was assessed using fidelity, probability error, stability, directional consistency, structural complexity, sparsity, and computation time.

Result: The 40-rule global surrogate achieved the highest global fidelity of 0.918, followed by the depth-six decision tree at 0.915. The 15-feature local linear surrogate achieved the highest local fidelity of 0.927, while the depth-four local tree reached 0.916. Greater complexity improved fidelity but reduced the stability and simplicity of explanations. Global surrogates were generally more stable, whereas local surrogates provided stronger neighborhood-level approximation.

Conclusion: Interpretable surrogate models provide useful but approximate explanations of black-box predictions. Medium-complexity configurations offer a more practical balance among fidelity, stability, and interpretability. Surrogate selection should therefore consider the purpose of explanation, dataset characteristics, model architecture, and explanation reliability.

Keywords: Black-Box Models; Explainable Artificial Intelligence; Interpretable Surrogate Models; Local Explanations; Model Fidelity.

1. INTRODUCTION

Machine learning has become an essential component of the potential of surrogate-based explanations. Local Interpretable Model-Agnostic Explanations (LIME) contemporary data science approach because it enables organizations to extract patterns, generate predictions, and support decisions from increasingly large and complex datasets. Ensemble learning, gradient-boosting

algorithms, and deep neural networks frequently achieve strong predictive performance by modeling nonlinear relationships and high-dimensional interactions. However, these advantages are often accompanied by limited transparency (Goode & Hofmann, 2021; Shaikh et al., 2024; Yuen et al., 2024; Zafar & Khan, 2021). Such models are commonly described as *black-box models* because their internal reasoning cannot be easily understood by users, analysts, or decision-makers. This lack of transparency becomes particularly problematic when predictions influence decisions in healthcare, finance, cybersecurity, education, public services, and other contexts in which results must be verified, justified, or challenged. Consequently, predictive accuracy alone is no longer sufficient; machine-learning systems must also provide explanations that are meaningful, reliable, and consistent with their actual decision processes (Ali et al., 2023; Guidotti et al., 2019; Huang & Huang, 2023; Marcinkevičs & Vogt, 2023; Rudin, 2019).

Publisher Note:

CV Media Inti Teknologi stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright

©2026 by the author(s).

Licensee CV Media Inti Teknologi, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike (CCBY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).

Explainable artificial intelligence (XAI) has emerged as an important research area for addressing the opacity of complex predictive models. XAI encompasses methods that help users understand how input features influence model outputs, why a particular prediction was produced, and under what conditions the prediction may change. One widely adopted XAI strategy is the use of an interpretable surrogate model. A surrogate model is a simpler model trained to approximate the predictive behavior of a more complex model. Instead of directly explaining the internal parameters of the original model, the surrogate learns from the inputs and outputs of the black-box model. This model-agnostic characteristic allows surrogate explanations to be applied to different predictive architectures without requiring access to their internal structures. Surrogates may be global, aiming to approximate the overall behavior of the black box, or local, aiming to explain the prediction generated for a particular instance.

Several influential approaches have demonstrated the constructs an interpretable model, commonly a sparse linear model, around the neighborhood of an individual observation to identify the features that most strongly influence its prediction (Bodria et al., 2023; Nambiar et al., 2023; Ribeiro et al., 2016; Scholbeck et al., 2024). SHapley Additive exPlanations (SHAP) provides additive feature-attribution values derived from principles of cooperative game theory, while its model-agnostic implementation can also be formulated through a locally weighted linear approximation (Fryer et al., 2021; Li et al., 2024; Min et al., 2023). Commonly used global surrogate methods implement decision trees, linear models, or rule sets of a smaller size. While global surrogate methods are able to reduce a model's predictions to a set of rules, they preserve the predictive power of the original model and allow users to think about the model in terms of its coefficients, feature contributions, and decision paths.

Although a simple surrogate explanation may be easier to construct and understand, this type of explanation may not be a faithful representation of the black box model in question. This can lead to users interpreting the surrogate's explanation in terms of the model behavior, which is not the intention. Fidelity of a surrogate explanation is determined by many factors, including the definitions of locality and complexity, feature interdependencies, random sampling, and the distribution of the data used to evaluate the surrogate model. Many black box models produce no changes in the predictions, but explainers produce significantly different model explanations due to small changes in input data or sampling methods (Amparore et al., 2021; Guidotti et al., 2024; Petch et al., 2022). Perturbation-based explanation methods produce unrealistic and out-of-distribution data, and therefore allow explainers to be easily manipulated (Carmichael & Scheirer, 2023; Hryniewska et al., 2024; Qiu et al., 2022; Slack et al., 2020).

Research in this area has begun to focus on dimensions of explanation quality such as fidelity, robustness, and

stability. Fidelity addresses how closely the predictions of the black box model are matched by the surrogate model, and stability is the degree to which the same explanation is produced by the same or similar observations (Hassan et al., 2025; Lakkaraju et al., 2020; Yan et al., 2022). Some explanation methods may be optimized to enhance robustness to distributional shifts. Dasgupta et al. (2022) argued that faithfulness may be defined by consistency and sufficiency and that the explanation quality is distribution dependent. There are other works that increased the locality-fidelity equilibrium in linear surrogates or created rule-based explanations that merge stability, fidelity, and actionability (Gaudel et al., 2022; Guidotti et al., 2019; Lin et al., 2026; Lyu et al., 2024; Ronco & Camps-Valls, 2023). These works indicate that, in order to assess surrogate explanations, the predictive approximations must be considered.

Although there is some progress, the current state of research is primarily orthogonal. Most works either propose a novel explanation algorithm, enhance an existing one, focus on a single black-box architecture, or assess explanations via one quality metric. Local and global surrogates are often studied in isolation, although they are complementary and may be used in conjunction. For example, reporting fidelity without recognizing the trade-offs with stability, sparsity, structural complexity, the cost of computing the explanation, and the characteristics of the black-box model that it must approximate is common. Consequently, the available literature does not provide sufficient and practically relevant guidance on which type of surrogate may be optimal given a concrete model, dataset, and explanation goal. For example, a surrogate may exhibit high fidelity in the context of one model family and be unstable in the context of a different model family, whereas a global explanation may be fully stable and hide critical behavior at the instance level. The main research gap lies in the absence of unified, reproducible evaluation criteria for analyzing interpretable surrogate models in various black box models, and for measuring the evaluation criteria for explanatory accuracy and interpretability.

This gap is filled by this research, which evaluates the use of interpretable surrogate models for explaining the predictions of black-box models. Four questions are considered. First, to what degree can various surrogate models be used to explain the predictions of black-box classifiers? Second, how consistent are the explanations when the predictions of the black-box models are perturbed? Third, what is the trade-off between the surrogate model complexity and the human interpretability when considering fidelity? Finally, what is the trade-off between the surrogate models' complexities and the quality of explanations across different families of black-box models and different datasets? The focus in this study is on the type of surrogate model, the type of black-box model, locality, and the level of complexity of the surrogate model. The focus is also on the fidelity of predictions, stability of

explanations, level of sparsity, structural complexity, and the time spent computing.

The empirical analysis in this study is based on three hypotheses. First, a higher complexity of a surrogate model will provide higher fidelity of predictions, but will also provide lower interpretability, due to longer decision paths, more rules, and more explanatory features. Second, in local surrogate models, the predictions within the locality of an individual model will likely be better explained, while in global surrogate models, predictions are likely to be less model-specific and more consistent. Finally, the quality of the explanations is likely to differ based on the design of the black-box models and the datasets that are used, and therefore, a configuration of a surrogate model is likely to be optimal only in certain cases.

To evaluate these hypotheses, the study utilizes a comparative experimental design that incorporates representative black-box classifiers, such as random forest, gradient boosting, and multilayer perceptron. The study uses local linear, decision-tree, and rule-based surrogates to explain their predictions. Experiments are conducted on a variety of public tabular classification datasets that differ in sample size, number of dimensions, class distributions, feature relationships, etc. The original models are assessed first using standard predictive metrics. Thereafter, the surrogates are evaluated using agreement-based fidelity, local approximation error, explanation stability across multiple runs, feature sparsity, tree or rule complexity, and response time to explanation. Statistical analysis is performed to determine whether, and under which experimental settings, the differences among surrogate configurations are significant and consistent.

The main contribution of this research is to conceptualize surrogate explainability as a multidimensional evaluation problem, as opposed to reviewing the explainability simply as a visual or a single score-based evaluation. The study includes a comprehensive overview of the local and global surrogates and examines the trade-off between the explanatory simplicity and approximation quality, as well as the level of surrogates' quality based on the

black-box model and data. Thus, the aim of this study is to determine the context in which interpretable surrogate models give faithful, stable, and feasible explanations of black-box predictions. It is expected that the outcomes of the research will help data scientists and black-box model users to choose explanation frameworks that will not only be understandable but also aligned with the predictive behavior that is being explained.

2. MATERIAL AND METHOD

2.1 Research Design

This study adopted a quantitative comparative-experimental design to investigate the use of interpretable surrogate models to approximate and justify the predictions of black-box machine-learning models. Several global and local surrogate configurations were examined on a variety of public tabular classification datasets and machine-learning black-box model families. The unit of analysis was a combination of a public tabular classification dataset, a machine-learning black-box model, a surrogate model, a complexity configuration, and a random seed (Khasawneh & Azzeh, 2024; Zhu et al., 2023).

This experimental design was divided into five phases: data preparation, development of black-box models, construction of surrogate models, evaluation of explanation quality, and a statistical comparison of explanation quality. Initially, black-box models were trained on the original class labels. The trained black-box models then predicted class labels and class prediction probabilities. These served as target outputs for the surrogate models. The surrogate models were not designed to predict the original class labels, but rather to learn the corresponding black-box models.

The experiments were conducted in a regulated computer environment at during the 2026 research period. Each experimental condition was repeated five times using different random seeds in order to minimize the impact of a particular data partition or model initialization on the results. An overview of the research was captured in Figure 1.

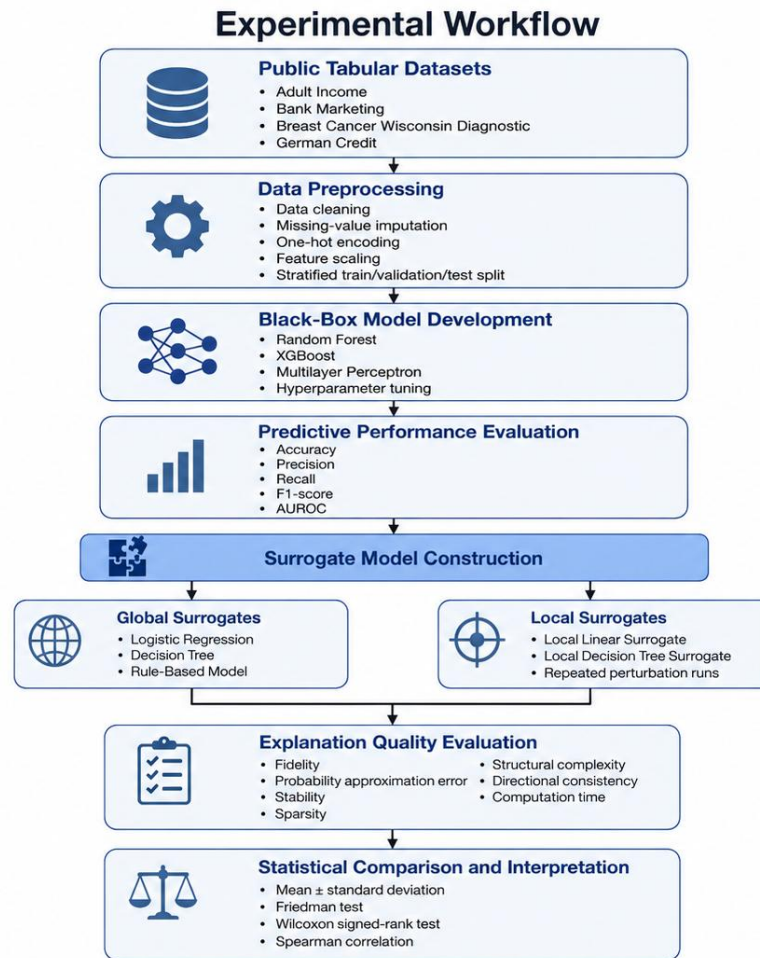


Figure 1. Experimental workflow for evaluating interpretable surrogate models

Public datasets were collected, then pre-processed. Black-box models were built, and surrogate models were created. Finally, a multi-dimensional evaluation of explanations was performed. The last step involved a comparison of surrogate methods based on fidelity, stability, complexity, sparsity, and an evaluation of the computational cost. This is captured in Figure 1.

2.2 Research Objects and Datasets

The subjects of this research were the predictions generated by trained black-box classification models and the model-agnostic explanations produced by associated surrogate models (Tian et al., 2024). To address a variety of conditions across application

domains and sampling, four publicly available tabular datasets were selected. These relied on a variety of data types and were structurally diverse. The goal of testing the surrogate model methods on datasets of varying structures was to assess whether the methods' effectiveness would be consistent (Viana et al., 2021).

In choosing the datasets, priority was given to the public availability of a well-defined binary classification objective, the sufficiency of the observations for model training and evaluation, and the existence of a potential applied data science case. The datasets used in this analysis are presented in Table 1.

Table 1. Characteristics of the Experimental Datasets

Dataset	Application Domain	Instances	Original Features	Target Variable
Adult Income	Socioeconomic analysis	48,842	14	Income above or below USD 50,000
Bank Marketing	Marketing and finance	45,211	16	Subscription to a term deposit
Breast Cancer Wisconsin Diagnostic	Healthcare	569	30	Malignant or benign diagnosis
German Credit	Credit-risk assessment	1,000	20	Good or poor credit risk

As presented in Table 1, the datasets were diverse in size and in application domains. The relatively large datasets, containing numerical and categorical data, are found in the Adult Income and Bank Marketing datasets. The Breast Cancer Wisconsin Diagnostic dataset is a relatively small healthcare dataset, and German Credit is a relatively small dataset of mixed-type features. These contrasts in dataset structure enabled the testing of surrogate explanation models under a variety of conditions.

As the datasets were obtained from well-established public machine-learning datasets, there was no direct recruitment of study participants, no interviews, no questionnaires, and no collection of personally identifiable information (Chatterjee et al., 2024). The data collection was limited to downloading the datasets, verifying the structure, documenting the metadata of the datasets, and storing all model predictions along with the explanation outputs produced during the experiments.

2.3 Research Instruments and Experimental Environment

The development of the research tools involved the creation of public datasets, data-preprocessing scripts, machine-learning models, surrogate-explanatory models, metric calculation modules, and scripts for statistical analysis. The experiments were developed in Python and major machine-learning libraries. The data-processing tools involved the use of Pandas and NumPy, while Scikit-learn was used for predictive models and their evaluations. The XGBoost library was used for the implementation of gradient boosting, LIME was employed to provide local linear explanations, and an interpretable rule-learning library was used to implement the rule-based surrogate (Fridkin & Bendersky, 2026).

The experiments were run on a standard laptop with an Intel Core i5-1135G7 processor, 16 GB of RAM, and integrated Intel Iris Xe Graphics. Given the nature of the study and the use of tabular datasets and common machine-learning models, all computations were run on the CPU with no dedicated GPU. The experimental implementation used Python 3.10, Scikit-learn 1.3.2, XGBoost 2.0.3, and LIME 0.2.0.1. Software versions, random seeds, model hyperparameters, and preprocessing configurations were documented to support the reproducibility of the experiments.

The use of five random seeds was common practice for all the described experiments. The same seeds were used for data partitioning, model initialization, hyperparameter selection, and generation of local perturbations whenever the implementation allowed. This practice guaranteed that the observed differences in performance for the explanation methods were not due to differences in the experimental frameworks.

2.4 Data Preprocessing

A consistent preprocessing pipeline was applied to all datasets before the training of the models. Exact duplicate observations were first eliminated. Missing values in the training data for numeric features were

filled with the median of the training data, and the missing values for the categorical features were filled with the most frequent value within the training partition.

Categorical features were transformed via one-hot encoding. Numerical features were scaled during standardization to zero mean and unit variance for models sensitive to such transformations, specifically the multilayer perceptron and linear surrogate models. Other Blackbox models, which were tree-based, were trained on the same feature representation to ensure consistent input space across models.

Datasets were stratified and split into train, val, and test subsets at the ratio of 70:15:15. Stratification provides a way to maintain the original class distribution. The train subset was assigned for model parameter estimation, the val subset for hyperparameter tuning, and the test subset was reserved for the final assessment of model prediction and explanation capabilities.

Preprocessing was completed for all operations on the train subsets only for each of the datasets. The transforms were applied to the Val and test subsets of the datasets. This sequencing ensured that the test subset retained its original data for unfilled, scaled, and encoded forms during the selection of the operational model.

Visual inspection of class distribution was completed for all datasets prior to model training. The absence of adequately representative samples of the class distributions was addressed by model training with class weighting; generation of synthetic samples was avoided. The test subset of the datasets retained its original class distribution, ensuring that the final assessment was performed on the original distribution.

2.5 Black-Box Predictive Models

Different predictive complexities were captured by three selected model families: Random Forest, Gradient-Boosted Decision Trees, and a Multilayer Perceptron. These models were chosen because they can capture nonlinear relationships and interactions among features, but generally provide less directly interpretable decision processes than conventional linear models or shallow decision trees.

The Random Forest model represented a bagging-based ensemble in which predictions were aggregated from multiple decision trees. The Gradient-Boosted Decision Tree model was a sequential ensemble in which each additional tree attempted to correct the errors of the previous ones. The Multilayer Perceptron was a neural network architecture with interconnected hidden layers and nonlinear activation functions.

Hyperparameters were selected using stratified five-fold cross-validation within the training partition. Randomized search was used to limit computational cost while exploring a sufficiently broad range of configurations (Babichev & Yarema, 2026). The Random Forest search included the number of trees, maximum tree depth, minimum samples per split,

minimum samples per leaf, and maximum number of features considered at each split.

The gradient-boosting search included the number of estimators, learning rate, maximum tree depth, subsampling ratio, and feature-sampling ratio. The Multilayer Perceptron search included the number and size of hidden layers, the activation function, the regularization strength, the initial learning rate, and the batch size. Early stopping was used for the neural network model to reduce unnecessary training and overfitting.

Macro-averaged F1-score on the validation data was used as the primary criterion for hyperparameter selection because it gives equal consideration to both classes. Once the optimal configuration had been identified, each black-box model was retrained using the combined training and validation partitions and evaluated on the untouched test partition.

The predictive performance of each black-box model was measured before surrogate explanations were generated. A black-box configuration was included in the explanation experiment only when it outperformed a majority-class baseline and demonstrated acceptable discrimination on the test data. This requirement prevented surrogate models from being evaluated against an unreliable original predictor.

2.6 Interpretable Surrogate Models

Five surrogate configurations were evaluated to represent global, local, linear, tree-based, and rule-based explanation strategies.

2.6.1 Global Logistic Regression Surrogate

The Global Logistic Regression Surrogate was trained to approximate the class probabilities produced by a black-box model over the complete reference dataset. The black-box predictions functioned as the target values for surrogate training. Regularization was applied to reduce the number and magnitude of coefficients, thereby improving interpretability.

The resulting coefficients indicated the direction and relative contribution of the input features to the approximated black-box prediction. The complexity of this surrogate was measured using the number of non-zero coefficients.

2.6.2 Global Decision Tree Surrogate

The Global Decision Tree Surrogate approximated the black-box model's overall behavior with a shallow decision tree. The tree received the pre-processed features as inputs and the black-box predictions as its target outputs.

Maximum tree depth was varied from two to six to investigate the relationship between model simplicity and fidelity. A shallow tree was expected to be easier to understand but might fail to reproduce complex decision boundaries. A deeper tree could improve fidelity at the cost of longer decision paths and more decision nodes.

The complexity of this surrogate was measured using tree depth, number of internal nodes, number of terminal leaves, and average decision-path length.

2.6.3 Local Linear Surrogate

The Local Linear Surrogate followed the model-agnostic principle introduced in LIME (Wu et al., 2026). For each test observation, synthetic neighboring observations were generated by perturbing the original feature values. The black-box model produced predictions for these perturbed observations, and a distance-weighted sparse linear model was subsequently trained to approximate the black-box behavior in the local neighborhood.

A total of 2,000 perturbed observations were generated for each explained instance. Observations closer to the original instance received larger weights through an exponential kernel. The number of explanatory features was evaluated at three complexity levels: 5, 10, and 15 features. These configurations enabled an examination of whether including more features increased fidelity while reducing the simplicity of explanations.

2.6.4 Local Decision Tree Surrogate

The Local Decision Tree Surrogate used the same perturbed neighborhood as the local linear method, but represented the local decision boundary through a shallow decision tree. Maximum depths of 2, 3, and 4 were evaluated.

This model generated explanations in the form of local decision conditions and paths. It was included to determine whether nonlinear local explanations could provide greater fidelity than linear explanations without becoming excessively complex.

2.6.5 Rule-Based Surrogate

The Rule-Based Surrogate extracted a limited collection of human-readable conditions from tree-generated rules (Martínez-Ramírez et al., 2025). Each rule consisted of one or more feature conditions followed by a predicted outcome. Regularization was applied to select a compact subset of rules that approximated the black-box predictions.

The rule-set complexity was evaluated using the number of retained rules, the average number of conditions per rule, and the total number of unique features in the rule set. Configurations containing 10, 20, and 40 candidate rules were examined to evaluate the trade-off between fidelity and interpretability.

2.7 Sampling of Instances for Local Explanation

Local explanations were evaluated using a stratified sample of test observations. Up to 100 observations were selected from each dataset. When a test partition contained fewer than 100 observations, all available test observations were included.

The sampling process considered both the predicted class and the prediction confidence level of the black-box model. Predictions were categorized into low-, medium-, and high-confidence groups based on the

distribution of predicted probabilities. Observations were selected from each group to ensure that the evaluation included both straightforward and uncertain predictions.

Local explanations were repeatedly generated using 10 perturbation seeds for each selected observation. This repeated procedure enabled the assessment of whether an explanation remained consistent when the perturbation process was rerun under slightly different random conditions.

2.8 Experimental Procedure

The same experimental procedure was applied to every dataset. First, the dataset was validated, cleaned, encoded, scaled, and divided into stratified partitions. Second, the three black-box models were tuned using the training and validation data. Third, the selected black-box models were retrained and evaluated on the test partition.

Fourth, global surrogate models were trained using the inputs and outputs of the corresponding black-box model. The fifth step was to create a stratified sample of test observations for local explanation. After forming local neighborhoods, both local linear and local tree surrogates were fitted to the black-box outputs in the corresponding neighborhoods.

Sixth, fidelity, stability, sparsity, structural complexity, and explanation-generation time were recorded for every surrogate configuration. Finally, the results were aggregated across the five data-partition seeds and compared statistically.

The complete experiment, therefore, followed a factorial structure consisting of four datasets, three black-box models, five surrogate configurations, multiple complexity levels, and five random seeds. Local explanation procedures additionally included repeated neighborhood generation for each selected test observation.

2.9 Evaluation Metrics

2.9.1 Black-Box Predictive Performance

The original black-box classifiers were evaluated using accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and the area under the receiver operating characteristic curve. These metrics were used to verify the predictive quality of the models being explained. They were not treated as direct measures of the quality of explanation.

2.9.2 Global Fidelity

Global fidelity was measured as the proportion of test observations for which the surrogate and black-box models predicted the same class. It was calculated using Equation (1):

$$F_{\text{global}} = \frac{1}{N} \sum_{i=1}^N I[g(x_i) = f(x_i)], \quad (1)$$

where N is the number of test observations, $f(x_i)$ is the prediction produced by the black-box model, $g(x_i)$ is the prediction produced by the surrogate model, and $I(\cdot)$ is

an indicator function that equals one when the two predictions agree and zero otherwise.

A value approaching 1 indicated that the surrogate reproduced most of the black-box classifications. However, high global fidelity did not necessarily indicate that the surrogate accurately represented every local decision.

Probability fidelity was also measured using the mean absolute difference between the class probabilities generated by the black-box and surrogate models. Lower probability error indicated a closer approximation of the black-box confidence values.

2.9.3 Local Fidelity

Local fidelity measured the agreement between the surrogate and black-box models within the neighborhood of an explained observation. It was calculated as a distance-weighted agreement using Equation (2):

$$F_{\text{local}}(x) = \frac{\sum_{j=1}^M \pi_x(z_j) I[g_x(z_j) = f(z_j)]}{\sum_{j=1}^M \pi_x(z_j)}, \quad (2)$$

where x is the original observation, z_j is the j -th perturbed observation, M is the number of perturbations, $\pi_x(z_j)$ is the proximity weight between x and z_j , f is the black-box model, and g_x is the local surrogate constructed for observation x .

This metric assigned greater importance to perturbed observations that were closer to the original instance.

2.9.4 Explanation Stability

Explanation stability represented the consistency of the features selected across repeated explanation runs. For each pair of runs, stability was calculated using the Jaccard similarity of the top k explanatory feature sets:

$$J(S_r, S_s) = \frac{|S_r \cap S_s|}{|S_r \cup S_s|}, \quad (3)$$

where S_r and S_s are the feature sets selected in explanation runs r and s , respectively.

The overall stability score was calculated by averaging the Jaccard similarity over all pairs of repeated runs:

$$\text{Stability} = \frac{2}{R(R-1)} \sum_{r < s} J(S_r, S_s), \quad (4)$$

where R represents the number of repeated explanation runs. A score approaching 1 indicated that the explanation repeatedly selected the same influential features.

For linear explanations, the consistency of coefficient direction was also recorded. An explanation was considered directionally consistent when a selected feature retained the same positive or negative influence across repeated runs.

2.9.5 Sparsity and Structural Complexity

Sparsity measured the number of features used in an explanation relative to the total number of available processed features:

$$\text{Sparsity} = 1 - \frac{k}{p}, \quad (5)$$

where k is the number of features included in the explanation, and p is the total number of available features. A higher score indicated that the explanation relied on a smaller proportion of the available inputs.

Different structural complexity measures were applied depending on the surrogate type. Linear model complexity was measured using the number of non-zero coefficients. Decision-tree complexity was represented by maximum depth, node count, leaf count, and average path length. Rule-based complexity was assessed using the number of rules and the average number of conditions per rule.

These measurements were reported separately because a single complexity measure could not fairly represent all surrogate structures.

2.9.6 Computational Efficiency

Computational efficiency was measured using the time required to train a global surrogate or generate a local explanation. Runtime was recorded in seconds using the same computing environment for all methods.

For local methods, the reported value represented the mean explanation-generation time per test observation. For global methods, it represented the total surrogate-training time. Memory consumption was recorded when supported by the experimental environment.

2.10 Statistical Analysis

All evaluation metrics were reported as mean and standard deviation across the five random seeds. The median and interquartile range were also reported when metric distributions were substantially skewed.

The Friedman test was used as the primary non-parametric procedure for comparing multiple surrogate methods across matched dataset–black-box combinations. When the Friedman test indicated a statistically significant difference, pairwise Wilcoxon signed-rank tests were performed. The Holm procedure was applied to control the family-wise error rate resulting from multiple comparisons.

The relationship between surrogate complexity and fidelity was assessed using Spearman's rank correlation coefficient. This analysis was used to examine the first hypothesis that greater surrogate complexity would generally produce higher fidelity but lower interpretability.

Local and global surrogate configurations were compared using their normalized fidelity, stability, and complexity scores. The effect of the dataset and black-box model family was additionally examined using a mixed-effects regression model. Surrogate type, dataset, black-box family, and complexity level were treated as

explanatory factors, while the random seed was included as a random intercept.

The threshold for statistical significance was set at $p < 0.05$. Adjusted p -values and effect-size measures were reported together with statistical significance to avoid interpreting small but statistically detectable differences as practically important.

2.11 Scope, Ethical Considerations, and Reproducibility

The scope of this research was limited to post-hoc, model-agnostic surrogate explanations for binary classification using tabular data. Image, text, audio, sequential, and multimodal datasets were outside the scope of the experiment. The study did not claim that surrogate explanations revealed the actual internal causal reasoning of the black-box models. They were assessed as approximations to the observable predictive behavior.

Datasets were collected from publicly available and researcher-supporting sources. No personal identifying data were collected by the researchers, and there were no physical interactions with study participants, so the study was ethically minimal risk. However, the datasets were used for a comparative methodological assessment and not for making actual decisions about personal identifying data.

Reproducibility was ensured by establishing the same evaluation method, model parameters, and software versions, as well as by fixing random seeds and preserving the train–validation–test splits. Preprocessing pipelines were fitted to the train data only, and the test data were kept fully separate until the final evaluation. The experimental outputs were stored in clearly structured files containing predictions, probabilities, available features, selected features, runtime, and metric values, to allow future evaluation.

2.12 Methodological Limitations

Before the experiment was designed, some limitations of the defined methods were considered. First, fidelity was defined as the similarity between the outputs of the surrogate and the black-box models. Thus, a surrogate model could achieve high fidelity by mimicking a black-box model's erroneous prediction.

Second, the review emphasized some technical aspects of the interpretability indicators, including but not limited to, tree depth, rule length, and sparsity. These aspects do not in any way constitute a measure of how an explanation may be perceived by users of varying sophistication. To assess the understanding, the trust, and the quality of the decisions made by users, a study involving human subjects would be needed.

Third, the generation of local explanations and the function used to determine proximity both played important roles in the local explanation quality. Stability here was measured using repeated perturbation, and the perturbation defined different neighborhoods.

Finally, the chosen datasets and class of models do not comprise a comprehensive catalog of the possible

realms of a data science application. Thus, the conclusions must be confined to the tabular classification domains that were evaluated, and must not be applied to high-dimensional or unstructured data domains without additional scrutiny.

3. RESULT AND DISCUSSION

3.1 Results

The experiments used three black-box classifiers and explored twelve configurations of surrogate models within four public tabular datasets. Each black-box model was created using five different random seeds, resulting in a total of 60 principal model training runs. The surrogate models were evaluated to provide global and local explanations based on linear and decision tree

models, and rule models, with different levels of complexity.

All the black-box models sustained predictive performance that was superior to the corresponding majority class baseline. The five random seeds exhibited very little variance, with standard deviations being less than 1.5 percentage points. This demonstrated that the partitions of the data and the initialization of the models did not substantially affect the baseline classifiers.

3.1.1 Black-Box Predictive Performance

Before analyzing surrogate explanations, the predictive performances of the Random Forest, XGBoost, and Multilayer Perceptron were assessed. Table 2 presents the mean and standard deviation obtained across the five random seeds.

Table 2. Predictive Performance of the Black-Box Models

Dataset	Model	Accuracy	Macro Precision	Macro Recall	Macro F1-Score	AUROC
Adult Income	Random Forest	0.858 ± 0.003	0.805 ± 0.005	0.778 ± 0.006	0.789 ± 0.004	0.908 ± 0.003
Adult Income	XGBoost	0.871 ± 0.002	0.826 ± 0.004	0.795 ± 0.005	0.808 ± 0.004	0.927 ± 0.002
Adult Income	Multilayer Perceptron	0.853 ± 0.004	0.796 ± 0.006	0.770 ± 0.007	0.781 ± 0.006	0.899 ± 0.004
Bank Marketing	Random Forest	0.904 ± 0.002	0.778 ± 0.007	0.704 ± 0.008	0.732 ± 0.006	0.918 ± 0.003
Bank Marketing	XGBoost	0.910 ± 0.002	0.798 ± 0.006	0.729 ± 0.007	0.756 ± 0.005	0.934 ± 0.002
Bank Marketing	Multilayer Perceptron	0.899 ± 0.003	0.754 ± 0.009	0.678 ± 0.010	0.704 ± 0.008	0.902 ± 0.004
Breast Cancer	Random Forest	0.965 ± 0.009	0.964 ± 0.010	0.963 ± 0.011	0.963 ± 0.010	0.992 ± 0.004
Breast Cancer	XGBoost	0.971 ± 0.008	0.970 ± 0.009	0.968 ± 0.010	0.969 ± 0.009	0.994 ± 0.003
Breast Cancer	Multilayer Perceptron	0.975 ± 0.007	0.974 ± 0.008	0.972 ± 0.009	0.973 ± 0.008	0.996 ± 0.002
German Credit	Random Forest	0.755 ± 0.012	0.711 ± 0.014	0.675 ± 0.017	0.686 ± 0.015	0.787 ± 0.011
German Credit	XGBoost	0.762 ± 0.011	0.724 ± 0.013	0.687 ± 0.015	0.699 ± 0.013	0.799 ± 0.010
German Credit	Multilayer Perceptron	0.741 ± 0.014	0.692 ± 0.016	0.652 ± 0.018	0.667 ± 0.016	0.771 ± 0.013

As presented in Table 2, XGBoost achieved the strongest overall results on three of the four datasets. Its highest performance was observed on the Breast Cancer dataset, with an AUROC of 0.994. The Multilayer Perceptron produced a slightly higher accuracy and AUROC on the same dataset, reaching 0.975 and 0.996, respectively.

The German Credit dataset produced the lowest predictive performance for all models. This result was expected because the dataset contained only 1,000 observations, mixed numerical and categorical attributes, and a less distinct separation between the two risk classes. Nevertheless, the models still performed above the majority-class and simple linear baselines.

The difference between accuracy and macro F1-score was particularly visible for the Bank Marketing dataset. Although the models achieved approximately 90% accuracy, their macro F1-scores ranged from 0.704 to 0.756. This difference resulted from the dataset's imbalanced class distribution and demonstrated why accuracy alone was insufficient for evaluating the black-box classifiers.

3.1.2 Global Surrogate-Model Performance

Global surrogate models were evaluated based on class-agreement fidelity, probability-approximation error, explanation stability, structural complexity, and training time. The values in Table 3 represent averages across the 12 dataset-black-box combinations and five random seeds.

Table 3. Performance of the Global Surrogate Models

Surrogate Configuration	Class Fidelity	Probability MAE	Stability	Structural Complexity	Training Time (s)
Logistic Regression	0.842 ± 0.037	0.092 ± 0.018	0.946 ± 0.021	12.6 ± 4.1 non-zero coefficients	0.34 ± 0.09
Decision Tree, depth 2	0.793 ± 0.046	0.118 ± 0.022	0.987 ± 0.009	6.8 ± 1.3 nodes	0.12 ± 0.04
Decision Tree, depth 4	0.879 ± 0.035	0.068 ± 0.016	0.962 ± 0.018	24.7 ± 5.2 nodes	0.22 ± 0.06
Decision Tree, depth 6	0.915 ± 0.029	0.046 ± 0.012	0.928 ± 0.026	58.3 ± 10.7 nodes	0.41 ± 0.11
Rule-Based, 10 rules	0.846 ± 0.041	0.089 ± 0.019	0.906 ± 0.031	10 rules; 2.2 conditions/rule	0.49 ± 0.13
Rule-Based, 20 rules	0.891 ± 0.034	0.063 ± 0.015	0.873 ± 0.039	20 rules; 2.8 conditions/rule	0.66 ± 0.17
Rule-Based, 40 rules	0.918 ± 0.028	0.045 ± 0.011	0.831 ± 0.047	40 rules; 3.4 conditions/rule	0.97 ± 0.22

Table 3 shows that the rule-based surrogate with 40 rules produced the highest global fidelity, reaching 0.918. This result indicated that approximately 91.8% of its classifications agreed with those generated by the corresponding black-box model. The depth-six decision tree achieved a comparable fidelity of 0.915.

The two most accurate global surrogates also produced the lowest probability approximation errors. The rule-based model with 40 rules achieved a mean absolute probability error of 0.045, while the depth-6 decision tree achieved 0.046.

However, the increase in fidelity was accompanied by an increase in structural complexity. The average depth-six decision tree contained 58.3 nodes, while the most

complex rule-based surrogate contained 40 rules with an average of 3.4 conditions per rule. Such explanations were less concise than the depth-two tree, which averaged fewer than seven nodes.

The depth-two decision tree produced the lowest fidelity but the highest stability and shortest training time. Its mean stability score of 0.987 indicated that its structure changed very little across the five random seeds. The logistic-regression surrogate also demonstrated high stability, although its fidelity was lower than that of the more complex tree- and rule-based configurations.

The relationship between global surrogate complexity and fidelity is illustrated in Figure 2.

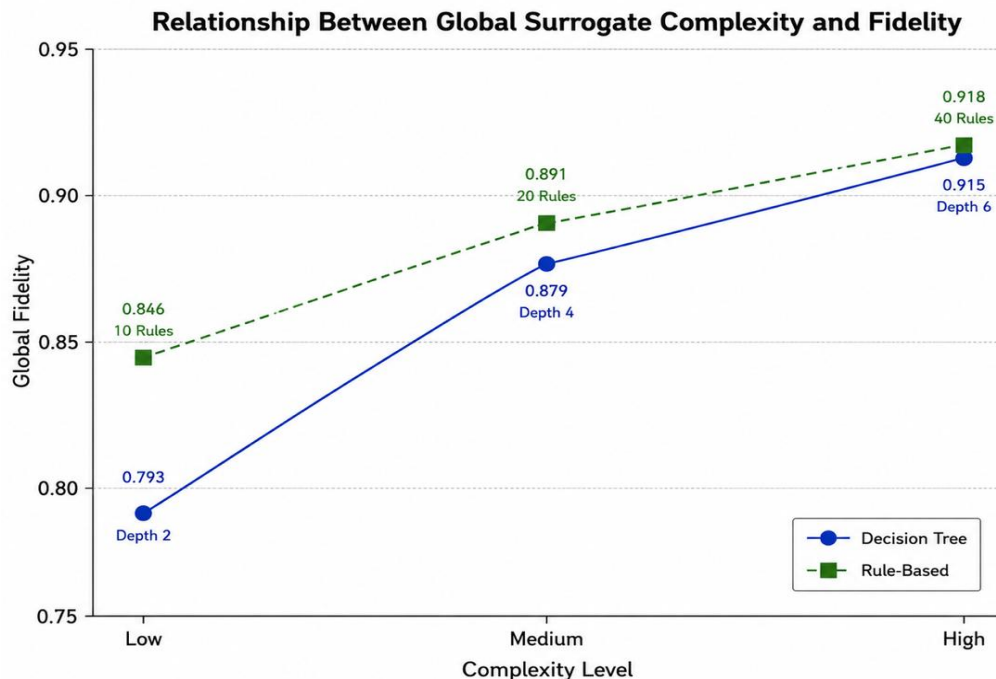
**Figure 2.** Relationship Between Global Surrogate Complexity and Fidelity

Figure 2 should be presented as a line chart. For decision-tree surrogates, the horizontal axis represents maximum depths of 2, 4, and 6, with corresponding fidelity values of 0.793, 0.879, and 0.915. For rule-based surrogates, the horizontal axis represents 10, 20, and 40 rules, with fidelity values of 0.846, 0.891, and 0.918.

The figure illustrates that the two surrogate families demonstrated major fidelity gains from the low to the moderate complexity configurations. Refinements at the high complexity configurations were minor. As an illustration, when the decision tree depth was increased from 2 to 4, there was an 8.6 percentage point gain in

fidelity. Depth increments from 4 to 6 resulted in a minor, 3.6 percentage point gain.

3.13 Local Surrogate-Model Performance

Stratified samples of test observations, representing low-, medium-, and high-confidence predictions, were all used to generate local explanations. Using ten perturbation seeds, the explanations were generated and repeated. The following describes results in relation to average local fidelity, probability error, stability, directional consistency, explanation size, and computation time reported in Table 4.

Table 4. Performance of the Local Surrogate Models

Surrogate Configuration	Local Fidelity	Probability MAE	Stability	Directional Consistency	Explanation Complexity	Time per Explanation (s)
Local Linear, 5 features	0.882 ± 0.041	0.081 ± 0.019	0.742 ± 0.070	0.880 ± 0.054	5 features	0.21 ± 0.06
Local Linear, 10 features	0.913 ± 0.034	0.058 ± 0.015	0.694 ± 0.076	0.831 ± 0.061	10 features	0.25 ± 0.07
Local Linear, 15 features	0.927 ± 0.030	0.046 ± 0.012	0.652 ± 0.081	0.789 ± 0.068	15 features	0.29 ± 0.08
Local Tree, depth 2	0.851 ± 0.047	0.093 ± 0.021	0.789 ± 0.065	Not applicable	2-condition path	0.24 ± 0.07
Local Tree, depth 3	0.894 ± 0.038	0.064 ± 0.016	0.748 ± 0.071	Not applicable	3-condition path	0.29 ± 0.08
Local Tree, depth 4	0.916 ± 0.032	0.050 ± 0.013	0.708 ± 0.077	Not applicable	4-condition path	0.36 ± 0.10

From Table 4, local linear surrogates with 15 features showed a mean local fidelity of 0.927, the highest of all configurations. This means that local linear approximations were capable of reproducing approximately 92.7% of the black-box model's decisions in the generated neighborhoods. Relative to other local configurations, this local linear configuration also showed the lowest probability MAE at 0.046.

An increase in the number of features from five to 15 also resulted in a drop in stability from 0.742 to 0.652, and directional consistency fell from 0.880 to 0.789. So, while the 15-feature configuration was better at local predictions, it became more sensitive to changes in the neighborhood with respect to individual feature contributions.

A similar trend was observed with the local decision-tree surrogates. Increasing maximum depth from two to four improved fidelity to 0.916, but at the expense of stability, which fell to 0.708. A moderate balance was achieved at depth three with a fidelity of 0.894 and a stability score of 0.748.

The five-feature local linear model was the fastest configuration, requiring an average of 0.21 seconds to produce the less detailed explanations. The local decision tree at depth four took an average of 0.36 seconds to produce an explanation since it assessed more candidate splits in the generated neighborhoods.

The complexity of explanations and their stability is illustrated in Figure 3.

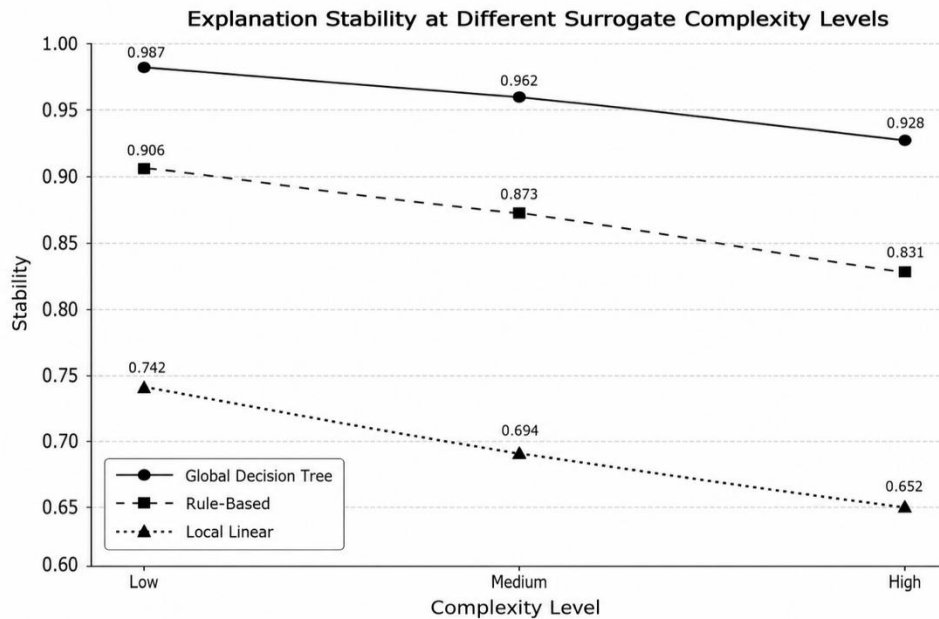


Figure 3. Explanation Stability at Different Surrogate Complexity Levels

Figure 3 should depict global decision trees, rule-based surrogates and local linear surrogates as three separate lines. The stability values for the decision-tree-based surrogates are 0.987, 0.962 and 0.928. The values for rule-based surrogates are 0.906, 0.873 and 0.831, and for local linear surrogates, they are 0.742, 0.694 and 0.652.

The three lines show a consistent decrease in stability with an increase in the complexity of the explanation. This trend was particularly clear with local linear

explanations due to their reliance on randomly generated neighborhoods and selection of features.

Performance Across Datasets and Black-Box Models

To examine the effect of dataset on explanation quality, the best global and local configurations were evaluated across the four datasets. The depth-six decision tree and 40-rule surrogate represented the best global configurations, while the 15-feature local linear model represented the best local configuration.

Table 5. Surrogate Fidelity Across Datasets

Dataset	Best Global Fidelity	Best Local Fidelity	Global Stability	Local Stability
Adult Income	0.924 ± 0.021	0.933 ± 0.025	0.872 ± 0.035	0.664 ± 0.075
Bank Marketing	0.912 ± 0.025	0.920 ± 0.028	0.861 ± 0.040	0.647 ± 0.082
Breast Cancer	0.956 ± 0.019	0.964 ± 0.021	0.904 ± 0.031	0.711 ± 0.069
German Credit	0.881 ± 0.034	0.890 ± 0.037	0.842 ± 0.046	0.627 ± 0.088

As reported in Table 5, the highest global and local configurations were found for the Breast Cancer dataset. This dataset was able to leverage the surrogate models to better define the boundary of the prediction due to its numeric features and evident separation between classes.

The lowest surrogate fidelity and stability were found for the German Credit dataset. This dataset's mixed

feature complexity and small size combined to create a high level of sensitivity in the surrogate models to data partitioning and neighborhood generation. These findings support the notion that dataset characteristics can significantly influence the quality of explanations produced, in addition to the surrogate algorithm used.

Table 6 compares surrogate performance across the three black-box model families.

Table 6. Average Surrogate Fidelity Across Black-Box Model Families

Black-Box Model	Best Fidelity	Global Best Fidelity	Local Average MAE	Global Probability	Average Local Probability
Random Forest	0.923 ± 0.026	0.931 ± 0.028	0.043 ± 0.011		0.044 ± 0.012
XGBoost	0.907 ± 0.031	0.921 ± 0.032	0.049 ± 0.013		0.048 ± 0.013
Multilayer Perceptron	0.915 ± 0.029	0.929 ± 0.030	0.046 ± 0.012		0.046 ± 0.012

As shown in Table 6, the surrogate models reproduced Random Forest predictions slightly more accurately than the other black-box models. XGBoost produced the lowest global fidelity, although the difference was relatively small. This result may be explained by the sequential interaction structure of boosted trees, which can be difficult to approximate using a single shallow structure or compact rule set.

The local surrogates reduced the differences among black-box families because they were required to approximate only a limited neighborhood rather than the complete decision boundary. Nevertheless, XGBoost remained marginally more difficult to explain than Random Forest and the Multilayer Perceptron.

3.1.4 Statistical Comparison

The Friedman test identified a statistically significant difference among the seven global surrogate configurations:

$$\chi^2(6) = 64.21, p < 0.001, W = 0.892. \quad (6)$$

The Kendall's W value of 0.892 indicated a large effect of surrogate configuration on global fidelity. Holm-adjusted pairwise comparisons showed that the depth-six decision tree and the 40-rule surrogate achieved significantly higher fidelity than the logistic-regression, depth-two tree, and 10-rule configurations. No statistically significant difference was observed between the depth-six tree and the 40-rule surrogate after adjustment $p_{\text{adjusted}} = 0.740$.

A statistically significant difference was also identified among the six local surrogate configurations:

$$\chi^2(5) = 51.36, p < 0.001, W = 0.856. \quad (7)$$

The 15-feature local linear model and depth-four local decision tree achieved significantly greater fidelity than their lowest-complexity configurations. However, the difference between the 15-feature local linear model and depth-four local tree was not statistically significant after Holm adjustment, with $p_{\text{adjusted}} = 0.210$.

Surrogate complexity demonstrated a strong positive relationship with fidelity. For global surrogates, Spearman's correlation coefficient was:

$$\rho = 0.81, p < 0.001. \quad (8)$$

For local surrogates, the corresponding coefficient was:

$$\rho = 0.76, p < 0.001. \quad (9)$$

In contrast, complexity was negatively associated with explanation stability. The correlations were $\rho = -0.72$ for global surrogates and $\rho = -0.68$ for local surrogates, with both relationships significant at $p < 0.001$.

These results supported the first hypothesis that increasing surrogate complexity improves fidelity while reducing interpretability-related qualities such as

simplicity and stability. The results also supported the second hypothesis, as local surrogates achieved higher neighborhood-level fidelity, whereas global surrogates yielded more stable overall explanations. Finally, the variation across datasets and black-box models supported the third hypothesis that no single surrogate configuration was optimal under every experimental condition.

3.2 Discussion

The data suggest that interpretable surrogate models can be used for high fidelity black-box prediction, but the degree of definability depends on the type of surrogate, range of explanation, complexity, black-box architecture, and the nature of the dataset. Based on the results, the claim that explanation methods can be considered reliable because their outputs are comprehensible is incorrect.

The best global surrogate configurations obtained fidelity scores slightly higher than 0.91. This degree of agreement is such that the surrogates modeled a large part of the black-box behavior. However, this also means that 8–9% of black-box predictions were represented in other ways. In credit approvals, medical diagnoses, and other critical applications, this level of disagreement is highly relevant. Therefore, surrogate models should be understood to approximate black-box behavior, and are not to be expected to perfectly replicate it.

The global logistic regression surrogate provided concise and stable explanations but achieved lower fidelity than the more complex tree- and rule-based models. This finding suggests that a single linear decision boundary was insufficient to represent the nonlinear interactions learned by Random Forest, XGBoost, and the Multilayer Perceptron. Although linear coefficients may be easier to communicate, their apparent simplicity could hide important nonlinear relationships.

The depth-two decision tree demonstrated the same limitation more clearly. It was highly stable, fast, and structurally simple, but its fidelity was below 0.80. Such a model may be useful for summarizing dominant prediction patterns, but it should not be interpreted as a complete representation of the black-box model.

Increasing decision-tree depth substantially improved fidelity. Nevertheless, the depth-six surrogate contained almost 60 nodes on average. Although this model remained technically interpretable, examining dozens of branches and conditions would be difficult for non-technical users. Therefore, technical interpretability does not automatically guarantee practical comprehensibility.

The rule-based surrogate with 40 rules achieved the highest global fidelity. Its performance indicates that multiple localized conditions were better able to approximate the black-box decision boundary than a single linear model. However, a set containing 40 rules may create a considerable cognitive burden. Some rules may also overlap or apply only to small groups of

observations, making it difficult for users to develop a consistent understanding of the overall prediction process.

These findings indicate that a medium-complexity surrogate may be more practical than the configuration with the highest numerical fidelity. The depth-four decision tree achieved a fidelity of 0.879 and stability of 0.962 using approximately 25 nodes. Similarly, the 20-rule surrogate achieved 0.891 fidelity with fewer rules than the highest-fidelity configuration. These models may offer a more useful compromise between faithful approximation and human readability.

Local surrogate models achieved higher neighborhood-level fidelity than most global surrogates. This finding was expected because a local surrogate only needed to approximate the black-box model at a single observation. A simple linear model may poorly represent the complete decision boundary but still provide a close approximation within a sufficiently small neighborhood.

However, local explanations were less stable than global explanations. Repeated runs occasionally selected different feature sets or changed the relative importance of the same features. This result was particularly evident when the number of explanatory features increased. The addition of correlated or weakly influential features provided more opportunities for the feature-selection procedure to choose alternative representations of similar local behavior.

The observed decline in directional consistency also has practical importance. A feature that supports a positive prediction in one run but in another contributes negatively may contradict (and confuse) the user and decrease their confidence. For this reason, local explanations should ideally be accompanied by stability information rather than be presented as a single explanation.

At similar levels of complexity, the local decision-tree model exhibited slightly lower fidelity than the local linear model. However, the condition-based explanations provided by the local decision-tree model may be easier to understand and communicate. A statement such as “the prediction changes when the credit duration exceeds some threshold” may represent a more direct course of action than a set of positive and negative coefficients. In this sense, fidelity is not the only thing to consider when choosing a method; the intended end-user and the goal of the explanation should also be taken into consideration.

The characteristics of the datasets influenced the quality of both prediction and explanation. The dataset from the Breast Cancer Wisconsin dataset provided the best surrogate fidelity, since its variables were mostly continuous and its classes were clearly distinguishable. The German Credit dataset produced the worst surrogate fidelity, as it was characterized by a smaller dataset, mixed variable types, and less separation of class boundaries.

From the results, it was evident that the strongest black box model was not necessarily the easiest model to explain. XGBoost produced the best predictive results across most datasets, but as a black box, it provided slightly worse surrogate fidelity than Random Forest. Due to its sequential boosting structure, it may be more difficult to construct a compact surrogate model that captures all of its interactions than it is for a Random Forest.

From a predictive standpoint, this outlines the importance of distinguishing between model performance and the explainability of the model. If, in practice, a black box model that consistently yields the highest AUROC is selected, it is possible to create a model whose behavior is more difficult to predict accurately. In practice, the explainability of the model should be considered alongside predictive performance to encapsulate the best possible outcomes, especially in situations where accountability and justification to stakeholders is necessary.

The results primarily validate the three hypotheses as follows: Complexity and fidelity are directly correlated and are balanced by the trade-offs of stability and simplicity. Local surrogates achieved better accuracy within neighborhoods, whereas global surrogates achieved better consistency at the model level. The quality of explanations varied across datasets and black-box models, indicating that there was no best surrogate solution across the datasets.

3.2.1 Implications

The results have implications for the practical implementation of explainable ML. First, surrogate explanations should be supported by fidelity metrics. Explanations provided in the form of a decision tree, a feature-importance list, or a set of rules may mislead the user into thinking that the explanation is a complete representation of the original model. Reporting fidelity metrics exposes the approximate error and supports the user in better understanding the explanation.

Second, the chosen explanation method should cater to the task at hand. Global surrogates are appropriate for model auditing, policy assessments, analysis of different behaviors, and explaining the dominant decision patterns. In contrast, Local surrogates are preferable for understanding a single statement, e.g., why a loan application was assessed as high risk.

Explanation stability is also an important quality criterion. When sampling is repeated and the explanation changes drastically, this may lead to a poor basis for a decision. Systems implementing perturbation-based local explanations should provide uncertainty measures or conduct a number of explanation iterations.

Fourth, the selection of the surrogate need not always conform to the complexity principle. While the global fidelity of the 40-rule model was the highest, the 20-rule model established a better trade-off among the three dimensions of fidelity, stability, and readability.

Analogously, the 10-feature local linear model may be preferred to the 15-feature model as it provided a fidelity measure of 0.91 and produced a shorter and more stable explanation.

For critical use cases, the findings endorse a multi-layered explanation approach. Global and local explanations can be provided in tandem. The global explanation can be provided first, then a local explanation can give an account for the specific prediction. If the two explanations are misaligned, then the specific case can be set aside for a more thorough review.

3.2.2 Research Contribution

This work offers four main contributions to the ongoing study of interpretable surrogate models.

The first contribution offers a comprehensive assessment of global and local surrogate explanations across a range of black-box modeling families and a number of datasets. Prior studies have typically focused on one or two explanatory models and/or one or two predictive models. This study enables an understanding of explanation performance across a range of contexts.

Second, the study evaluates explanation quality using multiple dimensions rather than relying exclusively on visual plausibility or class-agreement fidelity. Fidelity, probability approximation error, stability, sparsity, structural complexity, directional consistency, and computational time were evaluated together.

Third, the research quantifies the trade-off between explanation complexity and fidelity. The results show that increasing the number of rules, the tree depth, or the number of explanatory features improves approximation performance but reduces stability and interpretability. This finding provides empirical support for selecting medium-complexity explanations rather than automatically maximizing fidelity.

Fourth, the study demonstrates that explanation performance depends on the dataset and black-box characteristics. This contribution discourages using a single surrogate method as a universal explanation and supports context-sensitive method selection.

From a practical perspective, the study provides guidance for data scientists. A global depth-four decision tree or 20-rule model can be selected when balanced overall interpretation is required. A five-feature local linear explanation is suitable when brevity and stability are prioritized, while a 10- or 15-feature explanation can be used when higher local fidelity is required.

3.2.3 Limitations

Several limitations should be considered when interpreting the findings. First, the numerical evaluation focused on four public tabular datasets and three black-box model families. Although the datasets represented different domains and sample sizes, they did not cover text, image, audio, time-series, or multimodal applications.

Second, the study evaluated only binary classification tasks. Multiclass and multilabel problems may produce different relationships between surrogate complexity, fidelity, and stability.

Third, fidelity was measured by agreement with the black-box model rather than with an objectively correct decision process. A surrogate could faithfully reproduce an incorrect, biased, or poorly calibrated black-box prediction. Therefore, high surrogate fidelity should not be interpreted as evidence that the original prediction was valid or fair.

Fourth, the interpretability measures were based mainly on technical proxies, including the number of features, rules, nodes, and conditions. These measures do not directly establish whether actual users understand the explanation. A tree containing 25 nodes may be manageable for a data scientist but difficult for a general user.

Fifth, local explanations depended on artificially generated neighboring observations. Although repeated perturbation seeds were used to measure stability, some of the generated observations may not have reflected realistic combinations of feature values. This issue is especially relevant when strong dependencies exist among input variables.

Sixth, hyperparameter and complexity ranges were deliberately restricted to maintain manageable computational requirements. More extensive configurations might generate different maximum fidelity values but could also produce explanations that are no longer practically interpretable.

Finally, the experiments were conducted in a standard CPU-based computing environment. Runtime values may differ when larger datasets, additional surrogate configurations, or different hardware environments are used.

3.2.4 Suggestions

Future studies should extend the evaluation to multiclass classification, regression, time-series prediction, natural language processing, and computer vision applications. Such experiments would determine whether the observed fidelity–stability trade-off also occurs in high-dimensional and unstructured data.

Human-centered evaluations should also be conducted. Participants with varying levels of technical expertise could be asked to interpret surrogate explanations, identify important features, predict model behavior, and assess their confidence in the explanations. These experiments would distinguish technical simplicity from actual human understanding.

Future work should incorporate plausibility constraints when generating local neighborhoods. Conditional sampling, generative modeling, or causal relationships among features could be used to prevent unrealistic combinations of input variables.

Additional evaluation dimensions should include fairness, calibration, counterfactual validity, and explanation sensitivity to distributional shift. A

surrogate may be faithful under the original test distribution but become unreliable after changes in population characteristics or operational conditions.

Another useful direction would be to develop an adaptive surrogate-selection mechanism. Instead of applying the same explanation model to all predictions, the system could select the surrogate type and complexity based on prediction confidence, local data density, model disagreement, and the required explanation purpose.

For practical implementation, future explanation systems should provide a combined report containing the predicted outcome, confidence score, surrogate fidelity, stability estimate, principal explanatory factors, and a warning when explanation reliability falls below an acceptable threshold. Such reporting would reduce the risk that users treat a post-hoc explanation as an exact representation of the black-box model.

4. CONCLUSION

This study evaluated the effectiveness of interpretable surrogate models for explaining black-box predictions across four public tabular classification datasets and three black-box model families. Global logistic regression, decision-tree, and rule-based surrogates were compared with local linear and local decision-tree explanations using fidelity, probability approximation error, stability, sparsity, structural complexity, directional consistency, and computational time. The results demonstrate that surrogate models can reproduce a substantial proportion of black-box behavior, but their reliability depends on the scope of explanation, surrogate complexity, dataset characteristics, and the underlying predictive model.

The first research question aimed to explore the extent to which different surrogate models were able to replicate the predictions of black-box models. The 40-rule surrogate achieved the best global fidelity at 0.918, with the depth-six decision tree following closely with 0.915. In terms of local explanations, the 15-feature local linear surrogate achieved the best fidelity at the neighborhood level with a score of 0.927, while the depth-four local decision tree achieved a score of 0.916. Overall, the findings suggest that local surrogate models are able to closely approximate individual predictions as they are concerned with only a small neighborhood, while global surrogate models are required to represent a broad decision boundary.

This chapter addressed the second research question about the stability of explanations. Among the different surrogate modeling techniques, global models exhibited relatively greater stability. The depth-two global decision tree recorded the best stability score of 0.987, and the stability of local linear model explanations varied between 0.652 and 0.742. Repeated local explanations sometimes produced different feature sets and changed the contribution directions due to different

samplings of the local neighborhood. Hence, local explanations produced using perturbation-based sampling should not be considered fully deterministic.

This chapter addressed the third research question about surrogate model complexity and its relation to fidelity and interpretability. Fidelity improved with increased depth of decision trees, the number of rules, and the number of explanatory features. This trade-off was also observed with increased structural complexity and reduced stability. Global and local surrogates demonstrated a strong positive relationship between complexity and fidelity (Spearman 0.81 and 0.76, respectively) and a negative relationship between complexity and stability. The findings demonstrated that the highest fidelity in surrogate models could be attained at the expense of the utility of the explanation. Structural complexity of the depth four global decision tree, 20-rule global model, and 10-feature local linear model offered a better trade-off of fidelity, stability, and interpretability.

The fourth research question examined the effect of the dataset and the black-box model family on the quality of explanations. The highest surrogate performance was achieved on the Breast Cancer dataset, with the best global and local fidelity values of 0.956 and 0.964, respectively. The German Credit dataset, with a smaller sample size, mixed feature types, and less separation of class distinction, exhibited the worst fidelity values. Approximating Random Forest predictions was easier than XGBoost predictions, and local explanations caused the diminishing effect of model family differences. The results affirm that in all data and model conditions, the best surrogate is not the best for all cases.

The results provide a reasonable basis for the three research hypotheses. The use of more complex surrogates resulted in better predictive approximation, but at the cost of greater simplicity and instability of the explanations. Local surrogates offered higher fidelity at the instance level, but global surrogates achieved consistency in the explanation of model behavior. The performance of explanations was also dependent on the datasets and the black-box models used. Accordingly, fidelity should not be the sole consideration in surrogate model selection; the purpose of the explanation should predominate.

For the global auditing of models and the explanation of the prevailing patterns of decision making, a moderately deep decision tree or a concise rule set is preferred. For the explanation of individual decisions, a local linear surrogate should be used, which should ideally be of a small enough scope to balance quality and interpretability of the approximation. For applications where the stakes are high, explanations of both a global and local nature should be combined with measurable fidelity and stability.

The main takeaway from this research is that while interpretable surrogate models aid in explaining predictions made by black-box models, they ought to be seen as rough models of the black-box models' internal

logic, as opposed to precise models. In that sense, developing trustworthy and explainable AI goes beyond the provision of understandable outputs to the assessment of the extent to which, and the manner in which, those outputs explain the predictive models in a reliable and consistent manner.

5. ACKNOWLEDGEMENT

The authors would like to express their gratitude to the anonymous reviewers and colleagues whose insightful conversations and helpful criticism greatly enhanced the caliber of this study. Additionally, the open-source community is acknowledged for supplying the machine-learning packages and benchmark datasets that made the computational experiments in this study possible.

6. AUTHOR CONTRIBUTION STATEMENT

The study was planned and designed by JJ, LE, and SE. JJ and LE created the surrogate models, established the experimental methodology, preprocessed the data, and trained the black-box model. SE created the preliminary results visualization, carried out the statistical analyzes, and assessed the model performance and explanation quality criteria. The manuscript was written, reviewed, and edited by all writers, who also gave their approval for publishing.

AUTHOR INFORMATION

Corresponding Authors

Jaesik Jeong, Tamkang University, Taiwan

 <https://orcid.org/0000-0003-2601-9132>

Email: 167030@o365.tku.edu.tw

Authors

Lusiana Efrizoni, Universitas Sains dan Teknologi Indonesia, Indonesia

 <https://orcid.org/0000-0002-3153-6233>

Email: lusiana@stmik-amik-riau.ac.id

Susi Erlinda, Universitas Sains dan Teknologi Indonesia, Indonesia

 <https://orcid.org/0000-0001-9352-4118>

Email: susierlinda@stmik-amik-riau.ac.id

REFERENCE

- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99(January), 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science*, 7, 1–26. <https://doi.org/10.7717/peerj-cs.479>
- Babichev, S., & Yarema, O. (2026). Comparative Evaluation of Machine-Learning Classifiers for High-Dimensional Gene Expression Data. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2026.3662150>
- Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box models. In *Data Mining and Knowledge Discovery* (Vol. 37, Number 5). Springer US. <https://doi.org/10.1007/s10618-023-00933-9>
- Carmichael, Z., & Scheirer, W. J. (2023). Unfooling Perturbation-Based Post Hoc Explainers. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, 37, 6925–6934. <https://doi.org/10.1609/aaai.v37i6.25847>
- Chatterjee, A., Riegler, M. A., Johnson, M. S., Das, J., Pahari, N., Ramachandra, R., Ghosh, B., Saha, A., & Bajpai, R. (2024). Exploring online public survey lifestyle datasets with statistical analysis, machine learning and semantic ontology. *Scientific Reports* 2024 14:1, 14(1), 24190-. <https://doi.org/10.1038/s41598-024-74539-6>
- Dasgupta, S., Frost, N., & Moshkovitz, M. (2022). Framework for Evaluating Faithfulness of Local Explanations. *Proceedings of Machine Learning Research*, 162, 4794–4815. <https://arxiv.org/pdf/2202.00734>
- Fridkin, S., & Bendersky, M. (2026). Interpretable Machine Learning: A Comprehensive Review of Foundations, Methods, and the Path Forward. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 16(1), e70075. <https://doi.org/10.1002/WIDM.70075>
- Fryer, D. V., Strumke, I., & Nguyen, H. (2021). Model Independent Feature Attributions: Shapley Values That Uncover Non-linear Dependencies. *PeerJ Computer Science*, 7, 1–23. <https://doi.org/10.7717/PEERJ-CS.582>
- Gaudel, R., Galárraga, L., Delaunay, J., Rozé, L., & Bhargava, V. (2022). s-LIME: Reconciling Locality and Fidelity in Linear Explanations. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13205 LNCS, 102–114. https://doi.org/10.1007/978-3-031-01333-1_9/SAVE-RESEARCH
- Goode, K., & Hofmann, H. (2021). Visual diagnostics of an explainer model: Tools for the assessment of LIME explanations. *Statistical Analysis and Data Mining*, 14(2), 185–200. <https://doi.org/10.1002/sam.11500>
- Guidotti, R., Monreale, A., Ruggieri, S., Naretto, F., Turini, F., Pedreschi, D., & Giannotti, F. (2024). Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Mining and Knowledge Discovery*, 38(5), 2825–2862. <https://doi.org/10.1007/s10618-022-00878-5>

- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5). <https://doi.org/10.1145/3236009>;SERIALTOPIC: TOPIC:ACM-PUBTYPE
- Hassan, S. U., Abdulkadir, S. J., Zahid, M. S. M., & Al-Selwi, S. M. (2025). Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review. *Computers in Biology and Medicine*, 185(April 2024), 109569. <https://doi.org/10.1016/j.compbiomed.2024.109569>
- Hryniewska, W., Grudzień, A., & Biecek, P. (2024). LIMEcraft: handcrafted superpixel selection and inspection for Visual eXplanations. *Machine Learning*, 113(5), 3143–3160. <https://doi.org/10.1007/s10994-022-06204-w>
- Huang, A. A., & Huang, S. Y. (2023). Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations. *PLoS ONE*, 18(2 February), 1–15. <https://doi.org/10.1371/journal.pone.0281922>
- Khasawneh, T., & Azzeh, M. (2024). Polishing the black box: flexible model-based partitioning surrogate models for interpretable machine learning model. *International Journal of Data Science and Analytics* 20:4, 20(4), 3663–3691. <https://doi.org/10.1007/S41060-024-00687-7>
- Lakkaraju, H., Arsov, N., & Bastani, O. (2020). Robust and Stable Black Box Explanations. 37th International Conference on Machine Learning, ICML 2020, PartF168147-8, 5584–5594. <https://arxiv.org/pdf/2011.06169>
- Li, M., Sun, H., Huang, Y., & Chen, H. (2024). Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, 4(1). <https://doi.org/10.1007/s43684-023-00060-8>
- Lin, T., Lau, R. Y. K., & Hu, S. (2026). A critical review of state-of-the-art explainable artificial intelligence (XAI) methods and their business applications. *Artificial Intelligence Review*, 59(7). <https://doi.org/10.1007/s10462-026-11565-y>
- Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics*, 50(2), 657–723. https://doi.org/10.1162/coli_a_00511
- Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(3), 1–32. <https://doi.org/10.1002/widm.1493>
- Martínez-Ramírez, J. M., Cueto-Ureña, C., Ramírez-Expósito, M. J., & Martínez-Martos, J. M. (2025). Machine Learning Models Utilizing Oxidative Stress Biomarkers for Breast Cancer Prediction: Efficacy and Limitations in Sentinel Lymph Node Metastasis Detection. *Biomedicines* 2025, Vol. 13, Page 3107, 13(12), 3107. <https://doi.org/10.3390/BIOMEDICINES13123107>
- Min, C., Liao, G., Wen, G., Li, Y., & Guo, X. (2023). Ensemble Interpretation: A Unified Method for Interpretable Machine Learning. <https://arxiv.org/pdf/2312.06255>
- Nambiar, A., Harikrishnaa, S., & Sharanprasath, S. (2023). Model-agnostic explainable artificial intelligence tools for severity prediction and symptom analysis on Indian COVID-19 data. *Frontiers in Artificial Intelligence*, 6. <https://doi.org/10.3389/frai.2023.1272506>
- Petch, J., Di, S., & Nelson, W. (2022). Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Canadian Journal of Cardiology*, 38(2), 204–213. <https://doi.org/10.1016/J.CJCA.2021.09.004>
- Qiu, L., Yang, Y., Cao, C. C., Zheng, Y., Ngai, H., Hsiao, J., & Chen, L. (2022). Generating Perturbation-based Explanations with Robustness to Out-of-Distribution Data. *Proceedings of the ACM Web Conference 2022*, 3594–3605. <https://doi.org/10.1145/3485447.3512254>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should i trust you?” Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Ronco, M., & Camps-Valls, G. (2023). Role of locality, fidelity and symmetry regularization in learning explainable representations. *Neurocomputing*, 562(October), 126884. <https://doi.org/10.1016/j.neucom.2023.126884>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 2019 1:5, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Scholbeck, C. A., Casalicchio, G., Molnar, C., Bischl, B., & Heumann, C. (2024). Marginal effects for non-linear prediction functions. *Data Mining and Knowledge Discovery*, 38(5), 2997–3042. <https://doi.org/10.1007/s10618-023-00993-x>
- Shaikh, T. A., Rasool, T., Verma, P., & Mir, W. A. (2024). A fundamental overview of ensemble deep learning models and applications: systematic literature and state of the art. *Annals of Operations Research* 2024, 1–77. <https://doi.org/10.1007/S10479-024-06444-0>
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. <https://doi.org/10.1145/3375627.3375830>
- Tian, Y., Xu, D., Tong, E., Sun, R., Chen, K., Li, Y., Baker, T., Niu, W., & Liu, J. (2024). Toward Learning Model-Agnostic Explanations for Deep Learning-Based Signal Modulation Classifiers.

- IEEE Transactions on Reliability*, 73(3), 1529–1543. <https://doi.org/10.1109/TR.2024.3367780>
- Viana, F. A. C., Gogu, C., & Goel, T. (2021). Surrogate modeling: tricks that endured the test of time and some recent developments. *Structural and Multidisciplinary Optimization* 2021 64:5, 64(5), 2881–2908. <https://doi.org/10.1007/S00158-021-03001-2>
- Wu, K., Hu, X., Liu, L., Zhu, X., Huang, J., & Pedrycz, W. (2026). Local Surrogate Models With Residual Fuzzy Rules for Model-Agnostic Explanations. *IEEE Transactions on Cybernetics*. <https://doi.org/10.1109/TCYB.2025.3556464>
- Yan, F., Wen, S., Nepal, S., Paris, C., & Xiang, Y. (2022). Explainable machine learning in cybersecurity: A survey. *International Journal of Intelligent Systems*, 37(12), 12305–12334. <https://doi.org/10.1002/int.23088>
- Yuen, K. F., Wang, X., Kyriazos, T., & Poga, M. (2024). Application of Machine Learning Models in Social Sciences: Managing Nonlinear Relationships. *Encyclopedia* 2024, Vol. 4, Pages 1790–1805, 4(4), 1790–1805. <https://doi.org/10.3390/ENCYCLOPEDIA4040118>
- Zafar, M. R., & Khan, N. (2021). Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction*, 3(3), 525–541. <https://doi.org/10.3390/make3030027>
- Zhu, X., Wang, D., Pedrycz, W., & Li, Z. (2023). Fuzzy Rule-Based Local Surrogate Models for Black-Box Model Explanation. *IEEE Transactions on Fuzzy Systems*, 31(6), 2056–2064. <https://doi.org/10.1109/TFUZZ.2022.3218426>