IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

Empirical Analysis of SHAP Stability Under Data Corruption Across Datasets and Model Architectures

Stow, May[1] and Stewart, Ashley Ajumoke[2]

Department of Computer Science and Informatics, Federal University Otuoke, Nigeria[1]

Department of Fine Arts and Design, University of Port Harcourt, Nigeria[2]

ORCID ID: https://orcid.org/0009-0006-8653-8363[1], https://orcid.org/0009-0006-8425-4236[2]

Abstract: The deployment of machine learning models in critical decision making requires reliable explanations that remain stable under varying data conditions. While SHapley Additive exPlanations (SHAP) provides theoretically grounded feature importance rankings, the stability of these explanations when models encounter corrupted or degraded data remains poorly understood. This study investigates the robustness of SHAP feature importance rankings under controlled data corruption scenarios across three classification algorithms and datasets of varying complexity. The methodology employs optimally regularized Logistic Regression, Random Forest, and XGBoost models trained on medical, financial, and text classification datasets. Controlled corruption mechanisms combining 5% random sample removal and Gaussian noise injection with standard deviation equal to 0.1 times feature standard deviation simulate realistic data quality degradation. Stability metrics including Spearman correlation, Kendall tau, and top k feature overlap quantify ranking preservation. Results demonstrate that properly regularized models maintain substantial SHAP stability, with Spearman correlations exceeding 0.89 across all configurations. Random Forest exhibits superior stability with near perfect correlation (0.999) on structured data, while maintaining correlations above 0.95 across all scenarios. The findings establish that appropriate regularization and model selection enable reliable SHAP explanations even under moderate data corruption, providing practical guidelines for deploying interpretable machine learning in production environments where data quality cannot be guaranteed.

Keywords: SHAP, explainable AI, feature importance, model interpretability, data corruption, robustness analysis.

1. INTRODUCTION

Machine learning models increasingly influence critical decisions across healthcare, finance, criminal justice, and social services, where understanding the rationale behind predictions becomes as important as accuracy itself (Rudin, 2019). The proliferation of complex ensemble methods and deep learning architectures has created a fundamental tension between predictive performance and interpretability, leading to the characterization of sophisticated models as "black boxes" that resist human comprehension (Lipton, 2018). This opacity poses significant challenges for regulatory compliance, stakeholder trust, and error diagnosis, particularly in domains where decisions carry substantial consequences for individuals and society.

The demand for interpretable machine learning has driven the development of various explanation methods that aim to illuminate the decision processes of complex models. Among these approaches, SHapley Additive exPlanations (SHAP) has emerged as a prominent framework for feature attribution, providing theoretically grounded explanations based on cooperative game theory (Lundberg & Lee, 2017). SHAP values offer consistent and locally accurate explanations by distributing a prediction's output among input features according to their marginal contributions, satisfying desirable properties such as local accuracy, missingness, and consistency (Lundberg et al., 2020). The method has gained widespread adoption across diverse applications, from medical diagnosis (Rodríguez-Pérez & Bajorath, 2020) to credit risk assessment (Bussmann et al., 2021), due to its ability to provide both global feature importance rankings and local explanation for individual predictions.

Despite the theoretical elegance and practical utility of SHAP, concerns have emerged regarding the stability and reliability of these explanations under varying conditions. Recent studies have demonstrated that feature importance rankings can exhibit sensitivity to minor perturbations in training data, model parameters, and computational approximations (Slack et al., 2020; Alvarez-Melis & Jaakkola, 2018). This instability raises fundamental questions about the trustworthiness of model explanations, particularly when deployed in production environments where data quality cannot be perfectly controlled. The phenomenon becomes especially problematic when explanations guide high stakes

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

decisions or regulatory compliance, as inconsistent feature attributions may undermine stakeholder confidence and legal defensibility.

The vulnerability of explanation methods to data quality degradation represents a critical yet understudied aspect of interpretable machine learning. Real world data collection processes invariably introduce various forms of corruption, including measurement noise, missing values, and sampling biases (Frénay & Verleysen, 2014). While extensive research has examined model robustness to noisy data (Nettleton et al., 2010), the stability of explanation methods under such conditions remains largely unexplored. Kumar et al. (2020) demonstrated that adversarial perturbations can manipulate LIME explanations while preserving model predictions, suggesting that explanation methods may be more fragile than the models they interpret. Similarly, Ghorbani et al. (2019) showed that imperceptible changes to input data can dramatically alter feature importance rankings, raising concerns about the reliability of explanations in adversarial settings.

The existing literature on explanation stability has primarily focused on adversarial scenarios or theoretical worst case analyses, leaving a gap in understanding how explanations behave under realistic data corruption patterns encountered in practice. Previous studies have typically examined single perturbation types in isolation (Dombrowski et al., 2019) or focused on specific model architectures (Hooker et al., 2019), without providing comprehensive analysis across different algorithms, datasets, and corruption mechanisms. Furthermore, the relationship between model regularization, overfitting, and explanation stability remains poorly understood, despite regularization being a standard practice for improving model generalization (Zhang et al., 2021).

This research addresses these limitations by conducting a systematic investigation of SHAP feature importance stability under controlled data corruption scenarios. The study examines three widely used classification algorithms (Logistic Regression, Random Forest, and XGBoost) across datasets of varying complexity, applying realistic corruption mechanisms that combine Gaussian noise injection with random sample removal. By employing optimally regularized models that minimize overfitting, the research isolates the impact of data quality degradation on explanation stability from confounding factors related to poor model specification.

The primary contributions of this work include: (1) a comprehensive empirical analysis of SHAP stability across multiple model architectures and dataset complexities under realistic corruption scenarios; (2) quantification of the relationship between regularization strength, model performance, and explanation robustness; (3) identification of model and dataset characteristics that influence explanation stability; and (4) practical guidelines for selecting and configuring models to maintain reliable explanations in production environments. The findings demonstrate that properly regularized models can maintain substantial explanation stability even under moderate data corruption, with ensemble methods showing superior robustness compared to single learners. These results provide empirical evidence supporting the deployment of SHAP explanations in real world applications while highlighting the importance of appropriate model selection and configuration for maintaining interpretability under imperfect conditions.

2. RELATED WORKS

2.1 Interpretability Methods in Machine Learning

The development of post hoc explanation methods has emerged as a dominant approach for interpreting complex machine learning models. Ribeiro et al. (2016) introduced Local Interpretable Model-agnostic Explanations (LIME), which approximates model behavior locally using interpretable surrogates, though subsequent research revealed sensitivity to sampling parameters and instability across similar instances (Alvarez-Melis & Jaakkola, 2018). The SHAP framework proposed by Lundberg and Lee (2017) unified several existing methods under a game theoretic foundation, providing unique solutions that satisfy desirable axioms including local accuracy and consistency. Comparative studies have demonstrated SHAP's superior theoretical properties and practical performance across diverse domains (Covert et al., 2021), though computational complexity remains a challenge for high dimensional data.

Alternative approaches to model interpretability include gradient based methods for neural networks (Sundararajan et al., 2017), attention mechanisms that highlight relevant input regions (Vaswani et al., 2017), and counterfactual explanations that identify minimal changes needed to alter predictions (Wachter et al., 2017). Each method offers distinct advantages and limitations, with SHAP providing a middle ground between computational efficiency and theoretical rigor. The proliferation of explanation methods has led to calls for standardized evaluation frameworks, as proposed by Adebayo et al. (2018), who demonstrated that some popular methods fail basic sanity checks when model parameters are randomized.

2.2 Stability and Robustness of Explanations

The reliability of explanation methods under perturbations has attracted increasing scrutiny as these techniques move from research to deployment. Ghorbani et al. (2019) demonstrated that adversarially crafted perturbations can dramatically alter feature attributions while maintaining prediction accuracy, revealing a fundamental vulnerability in explanation methods. Their work showed that explanations can be more fragile than the models they interpret, with imperceptible changes to inputs causing substantial shifts in feature importance rankings. Dombrowski et al. (2019)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

extended this analysis to show that simple transformations like rotation or translation can cause significant changes in saliency maps for image classifiers, questioning the reliability of visual explanations.

The distinction between adversarial and natural perturbations has important implications for practical deployment. While adversarial attacks represent worst case scenarios, Hooker et al. (2019) argued that realistic corruptions provide more relevant insights for real world applications. Their analysis of neural network explanations under common image corruptions revealed that some architectures maintain more stable attributions than others, suggesting that model design influences explanation robustness. Similarly, Agarwal et al. (2022) examined explanation stability across different random seeds and training runs, finding substantial variability in feature importance rankings even for models with comparable performance.

Recent work has begun exploring the relationship between model properties and explanation stability. Fel et al. (2021) demonstrated that smoother decision boundaries lead to more stable gradient based explanations, while Zhou et al. (2022) showed that ensemble methods naturally provide more robust feature attributions through averaging effects. These findings suggest that architectural choices and training procedures significantly impact explanation reliability, though systematic guidelines for achieving stable explanations remain underdeveloped.

2.3 Data Quality and Model Robustness

The impact of data corruption on model performance has been extensively studied across machine learning paradigms. Nettleton et al. (2010) provided a comprehensive analysis of noise effects on classification algorithms, demonstrating that ensemble methods generally exhibit greater resilience than single learners. Their work established that different noise types (attribute noise, class noise, and missing values) affect algorithms differently, with tree based methods showing particular vulnerability to attribute noise. Frénay and Verleysen (2014) surveyed label noise in supervised learning, identifying regularization and robust loss functions as effective mitigation strategies.

The relationship between overfitting and noise sensitivity has important implications for both prediction and interpretation. Zhang et al. (2021) demonstrated that deep neural networks can memorize random labels, achieving perfect training accuracy while failing to generalize, highlighting the importance of appropriate regularization. Belkin et al. (2019) described the "double descent" phenomenon, where highly overparameterized models can achieve good generalization despite perfectly fitting noisy training data, challenging traditional understanding of the bias variance tradeoff. These findings suggest that model capacity and regularization interact in complex ways to determine robustness to data corruption.

Recent research has examined how data augmentation and preprocessing affect model stability. Hendrycks and Dietterich (2019) introduced a benchmark for evaluating model robustness to common corruptions, revealing substantial performance degradation even for state of the art architectures. Chen et al. (2020) showed that simple data augmentation techniques can significantly improve robustness, though the benefits vary across model types and corruption patterns. These studies provide context for understanding how data quality affects model behavior, though they primarily focus on prediction accuracy rather than explanation stability.

2.4 SHAP Applications and Limitations

The widespread adoption of SHAP across domains has revealed both strengths and limitations of the approach. In healthcare applications, Lundberg et al. (2020) demonstrated SHAP's utility for identifying risk factors in mortality prediction, showing how tree based SHAP values can efficiently handle large scale electronic health records. Rodríguez-Pérez and Bajorath (2020) applied SHAP to drug discovery, revealing previously unknown structure activity relationships, though they noted challenges in handling molecular representations with inherent symmetries. Wang et al. (2021) applied SHAP values to tree-based machine learning methods for process analytics in wastewater treatment plants, demonstrating how feature importance rankings could identify key operational parameters for process optimization.

Financial applications have particularly embraced SHAP for regulatory compliance and risk assessment. Bussmann et al. (2021) evaluated SHAP explanations for credit scoring models, finding that stakeholders preferred SHAP over other explanation methods for its intuitive interpretation and consistency. However, Bracke et al. (2019) cautioned that SHAP values can be misleading when features are highly correlated, as the attribution of shared effects becomes arbitrary. This limitation is particularly relevant in financial data where economic indicators often move together.

Critical evaluations of SHAP have identified several theoretical and practical limitations. Slack et al. (2020) demonstrated that SHAP explanations can be manipulated by adversarial classifiers that hide biased behavior while producing seemingly fair explanations. Kumar et al. (2020) showed that out of distribution inputs can produce unreliable SHAP values, as the method assumes that feature distributions match the training data. Merrick and Taly (2020) argued that the choice of background distribution significantly affects SHAP values, yet practitioners often use defaults without considering their implications. These critiques highlight the need for careful application and interpretation of SHAP explanations.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

2.5 Evaluation Metrics for Explanation Quality

The assessment of explanation methods requires metrics that capture different aspects of quality and utility. Quantitative metrics for explanation stability include ranking correlation measures such as Spearman's rank correlation and Kendall's tau, which assess the consistency of feature orderings (Hooker et al., 2019). Top k intersection metrics evaluate whether the most important features remain consistent, which is particularly relevant for applications that focus on key drivers (Fel et al., 2021). Bhatt et al. (2020) proposed measuring explanation infidelity as the mean squared error between explanation attributions and model behavior, providing a fidelity metric that complements stability measures.

Human centered evaluation of explanations presents additional challenges and opportunities. Doshi-Velez and Kim (2017) outlined a framework for rigorous human evaluation of interpretability, distinguishing between functionally grounded, human grounded, and application grounded evaluations. Poursabzi-Sangdeh et al. (2021) conducted large scale human subject experiments showing that increasing model transparency does not always improve human decision making, and can sometimes lead to overconfidence in incorrect predictions. These findings emphasize that technical metrics alone cannot fully assess explanation quality.

Recent work has attempted to bridge technical and human centered evaluation approaches. Nauta et al. (2023) proposed a comprehensive evaluation framework combining computational metrics with user studies, demonstrating that different stakeholders prioritize different aspects of explanation quality. Chen et al. (2022) introduced metrics for measuring the actionability of explanations, assessing whether feature attributions provide useful guidance for improving outcomes. These developments suggest that explanation evaluation requires multiple complementary perspectives, though standardized benchmarks remain elusive.

2.6 Research Gap

Despite extensive research on interpretability methods and model robustness, the stability of SHAP explanations under realistic data corruption remains inadequately addressed. Previous studies have primarily focused on adversarial perturbations designed to maximally disrupt explanations (Ghorbani et al., 2019; Slack et al., 2020) or examined single corruption types in isolation (Dombrowski et al., 2019), without considering the compound effects of multiple degradation mechanisms encountered in practice. The relationship between model regularization, which is standard practice for preventing overfitting, and explanation stability has received limited attention despite its potential importance for maintaining reliable interpretations.

Furthermore, existing work has not systematically compared explanation stability across different model architectures, dataset complexities, and corruption scenarios within a unified framework. This gap prevents practitioners from making informed decisions about model selection and configuration when explanation reliability is a primary concern. The present research addresses these limitations by conducting comprehensive experiments that quantify SHAP stability under controlled yet realistic corruption conditions, providing empirical evidence and practical guidelines for deploying interpretable machine learning in imperfect data environments.

3. METHODOLOGY

3.1 Research Framework

This study employs a systematic approach to evaluate the robustness of SHAP feature importance rankings under controlled data corruption scenarios. The methodology comprises nine sequential stages designed to assess how machine learning models maintain explanation stability when training data quality degrades. Figure 1 illustrates the complete implementation workflow, demonstrating the progression from data acquisition through final analysis.

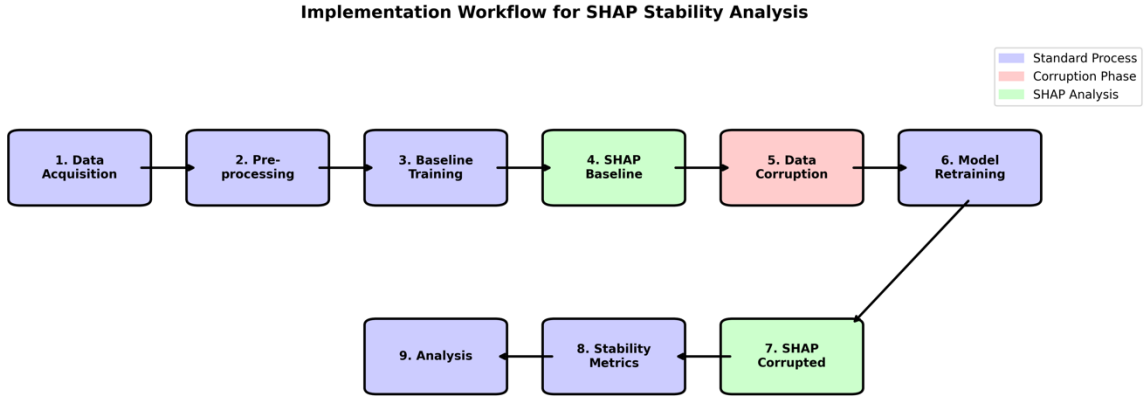

Figure 1: Implementation Workflow for SHAP Stability Analysis

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

The workflow follows a structured pipeline where standard preprocessing and baseline training establish reference metrics, followed by controlled corruption processes that simulate real world data quality issues. The dual SHAP analysis phases, before and after corruption, enable quantitative assessment of feature ranking stability.

3.2 Dataset Selection and Characteristics

Three binary classification datasets representing varying complexity levels were selected to evaluate model behavior across different problem domains. Table 1 presents the fundamental properties of each dataset, demonstrating the progression from simple medical diagnosis to complex text classification tasks.

Table 1: Dataset Properties and Characteristics

Dataset	Domain	Samples	Features	Classes	Difficulty	Description
Breast Cancer	Medical	569	30	2	Easy	Wisconsin Breast Cancer diagnostic data for tumor classification
Adult Income	Financial/Census	26,048	30	2	Medium	US Census data for income prediction above or below $50K threshold
Spambase	Text/Email	4,601	57	2	Hard	Email spam detection based on word and character frequency features

Figure 2 provides a visual comparison of dataset characteristics across three dimensions: sample size, feature dimensionality, and classification difficulty. The Breast Cancer dataset represents the simplest classification task with 569 samples and well separated classes. The Adult Income dataset introduces moderate complexity with 26,048 samples requiring demographic and employment features for income prediction. The Spambase dataset presents the greatest challenge with 57 features derived from text analysis for spam detection.

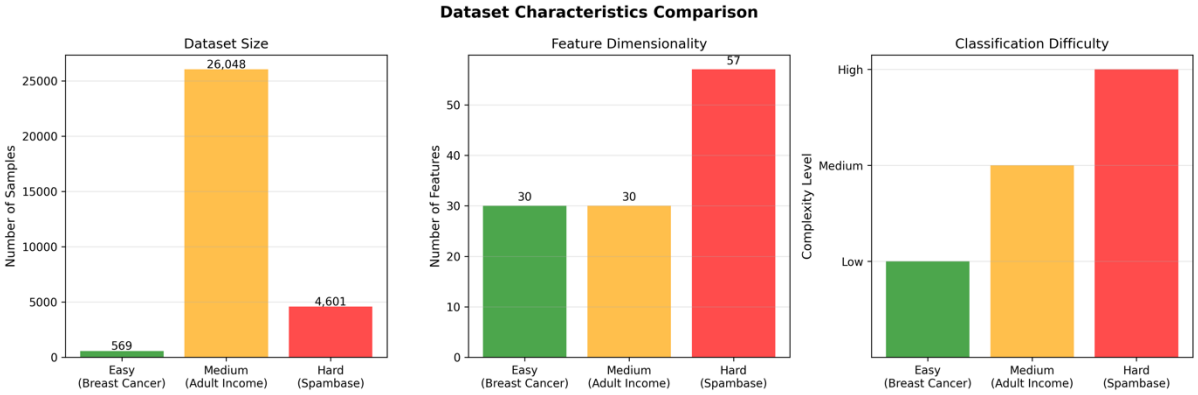

Figure 2: Dataset Characteristics Comparison

3.3 Data Preprocessing Pipeline

The preprocessing pipeline ensures consistent data quality and comparability across all experiments. Each dataset undergoes five standardized preprocessing steps:

1. **Missing Value Imputation**: Numerical features containing missing values are imputed using median values calculated from the training set to maintain statistical properties without introducing bias.
2. **Duplicate Removal**: Identical samples are identified and removed to prevent artificial inflation of model performance metrics.
3. **Train Test Splitting**: Data partitioning follows an 80/20 split with stratification to preserve class distributions. The random seed remains fixed at 42 throughout all experiments to ensure reproducibility.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

4. **Feature Standardization**: All features undergo standardization using the StandardScaler transformation, centering features at zero mean with unit variance. The scaler parameters are fitted exclusively on training data and applied to test data to prevent information leakage.
5. **Class Balance Assessment**: The ratio between minority and majority classes is calculated to identify potential imbalance issues. Datasets with ratios below 0.3 trigger balanced class weight adjustments in model training.

3.4 Model Configuration and Optimization

Three classification algorithms were selected to represent different modeling paradigms: linear models through Logistic Regression, ensemble methods via Random Forest, and gradient boosting through XGBoost. Table 2 details the optimal hyperparameters determined through preliminary gap analysis to minimize overfitting while maintaining predictive performance.

Table 2: Optimal Model Hyperparameters

Parameter	Logistic Regression	XGBoost	Random Forest
C (Regularization)	0.05	-	-
max_iter	1000	-	-
solver	liblinear	-	-
n_estimators	-	100	100
max_depth	-	1	1
learning_rate	-	0.1	-
reg_alpha	-	1.0	-
reg_lambda	-	2.0	-
min_samples_split	-	-	20
min_samples_leaf	-	-	10
class_weight	balanced	-	balanced

The hyperparameter selection prioritizes generalization over training accuracy. Logistic Regression employs strong L2 regularization with C=0.05 to prevent coefficient explosion. XGBoost utilizes stumps with max_depth=1 combined with L1 and L2 regularization to control model complexity. Random Forest similarly restricts tree depth to single splits while requiring minimum samples of 20 for node splitting and 10 for leaf nodes.

Figure 3 presents the learning curves for the Spambase dataset, demonstrating minimal gaps between training and cross validation scores across all models. The gaps remain below 0.005 for all three algorithms, confirming successful regularization without underfitting.

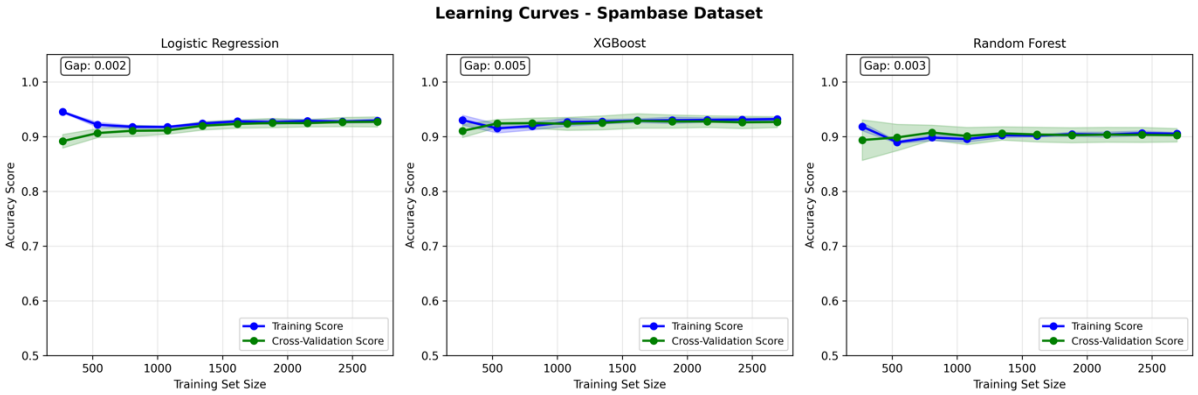

Figure 3: Learning Curves - Spambase Dataset

3.5 Data Corruption Methodology

The corruption process simulates two common data quality degradation scenarios encountered in production environments. Figure 4 illustrates the two stage corruption pipeline applied to training data.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

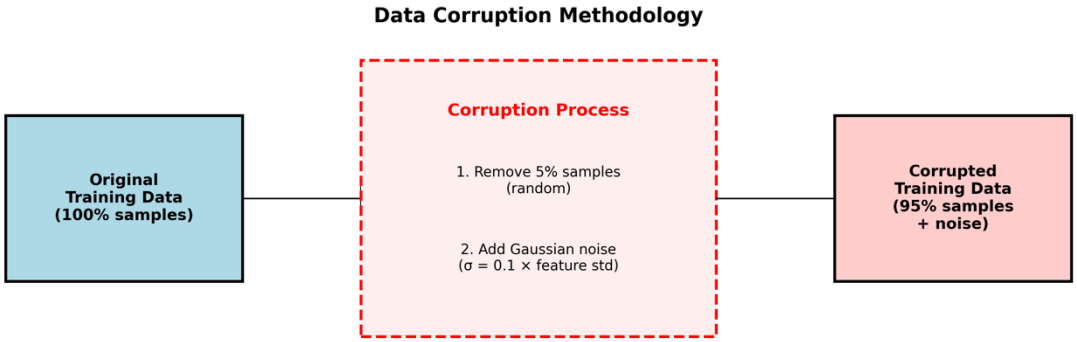

Figure 4: Data Corruption Methodology

The corruption methodology implements:

1. **Random Sample Removal**: Five percent of training samples are randomly removed to simulate incomplete data collection or sample loss. The removal process maintains temporal consistency by using a fixed random seed.
2. **Gaussian Noise Injection**: Zero mean Gaussian noise with standard deviation equal to 0.1 times each feature's standard deviation is added to all remaining samples. This feature specific scaling ensures proportional corruption across different measurement scales.

The mathematical formulation for noise injection follows:

$$X_{corrupted}[i,j] = X_{original}[i,j] + \varepsilon \quad (1)$$

where $\varepsilon \sim N(0, 0.1 * \sigma_j)$

Here, σ_j represents the standard deviation of feature j calculated from the original training data. This approach preserves relative feature importance while introducing controlled perturbations.

Standardization precedes corruption. Let $\sigma_{raw,j}$ denote the standard deviation of feature j on the raw training data. Gaussian noise is injected in standardized space as $\varepsilon_z \sim N(0, 0.1)$, which corresponds to $\varepsilon \sim N(0, 0.1 \cdot \sigma_{raw,j})$ in raw space. Corruption is applied only to the training set; the test set remains clean.

This corruption process is applied exclusively to the training data. The test set remains unmodified to ensure consistent evaluation conditions and isolate the impact of training data corruption on model explanations.

3.6 SHAP Value Computation

SHAP values are computed using model specific explainers to ensure computational efficiency and theoretical consistency. Linear models employ the LinearExplainer, which leverages the additive nature of linear predictions. Tree based models utilize the TreeExplainer, implementing an exact algorithm for computing SHAP values in polynomial time.

For each model and dataset combination, the SHAP analysis proceeds through:

1. **Baseline Computation**: SHAP values are calculated for the model trained on clean data, establishing reference feature importance rankings. These values are computed using the clean test set.
2. **Model Retraining**: The model undergoes complete retraining on corrupted training data (with 5% samples removed and Gaussian noise added) to simulate real world model updates under degraded conditions.
3. **Corrupted Computation**: SHAP values are recalculated using the retrained model applied to the clean test set, capturing how feature importance shifts when the model is trained on corrupted data while maintaining consistent evaluation conditions.

Feature importance scores derive from the mean absolute SHAP values across all test samples:

$$Importance_j = (1/N) * \Sigma|SHAP_{ij}| \quad (2)$$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

where N represents the number of test samples and SHAP_ij denotes the SHAP value for feature j in sample i.

3.7 Stability Metrics

Five complementary metrics quantify the stability of feature importance rankings between baseline and corrupted conditions:

1. **Spearman Rank Correlation**: Measures monotonic relationships between feature rankings, ranging from -1 to 1 where higher values indicate greater stability.
2. **Kendall Tau Correlation**: Assesses concordance between ranking pairs, providing a robust alternative to Spearman correlation for ordinal data.
3. **Top-5 Feature Overlap**: Calculates the proportion of features remaining in the top 5 positions after corruption, critical for identifying the most influential predictors.
4. **Top-10 Feature Overlap**: Extends overlap analysis to the top 10 features, capturing stability among moderately important features.
5. **Mean Rank Change**: Computes the average absolute change in feature positions, quantifying overall ranking disruption.
6. **Overfitting Severity**: Categorizes the degree of overfitting based on the train-test accuracy gap:
   - Low: Gap < 0.05
   - Medium: 0.05 ≤ Gap < 0.10
   - High: Gap ≥ 0.10
7. **Generalization Score**: Quantifies how well test performance matches training performance: Generalization Score = Test Accuracy / Train Accuracy Values approaching 1.0 indicate excellent generalization, while values significantly below 1.0 suggest overfitting.
8. **Max Rank Change**: The maximum absolute change in rank position for any single feature between baseline and corrupted conditions: Max Rank Change = max(|rank_baseline(i) - rank_corrupted(i)|) for all features i

3.8 Experimental Protocol

Each experiment follows a standardized protocol to ensure consistency and reproducibility:

1. Dataset loading and initial statistics computation
2. Preprocessing pipeline application with progress tracking
3. Model instantiation with optimal hyperparameters
4. Baseline training and performance evaluation
5. SHAP value computation on clean data
6. Data corruption application
7. Model retraining on corrupted data
8. SHAP value recomputation
9. Stability metric calculation
10. Visualization generation and result persistence

Figure 5 displays the learning curves for the Adult Income dataset, revealing near-perfect generalization with approximately zero gaps between training and validation scores for all models. This exceptional performance validates the effectiveness of the regularization strategy.

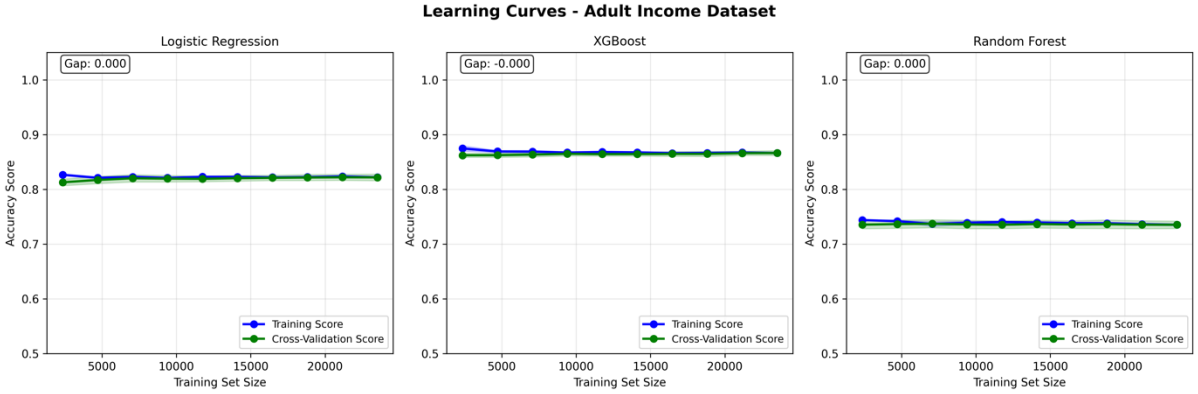

Figure 5: Learning Curves - Adult Income Dataset

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

All experiments utilize fixed random seeds to ensure complete reproducibility. Specifically, random seeds were set to 42 for all operations: scikit-learn models via the *random_state* parameter, XGBoost via both *random_state* and *seed* parameters, and NumPy operations via *np.random.seed(42)*. This ensures consistent results across all model architectures and corruption processes. Results are automatically saved in structured directories with timestamps, preserving both raw data and processed metrics for subsequent analysis. The implementation leverages scikit-learn 1.0.2 for model training, SHAP 0.41.0 for explanation generation, and NumPy 1.21.5 for numerical computations.

4. RESULTS AND DISCUSSION

This section presents the experimental findings from the SHAP stability analysis across three datasets of varying complexity. The results demonstrate the relationship between data corruption, model regularization, and feature importance stability, providing empirical evidence for the robustness of SHAP explanations under degraded data conditions.

4.1 Results

4.1.1 Model Performance Evaluation

Figure 6 illustrates the performance metrics across all three datasets and models. The test accuracy remains consistently high for the Easy dataset, with all models achieving above 92% accuracy. Logistic Regression and XGBoost both achieve 95.6% accuracy, while Random Forest reaches 92.1%. The Medium dataset shows greater variation in model performance, with XGBoost achieving the highest accuracy at 86.7%, followed by Logistic Regression at 82.1% and Random Forest at 73.5%. The Hard dataset demonstrates strong performance across all models, with XGBoost leading at 92.9%, Logistic Regression at 92.2%, and Random Forest at 89.7%.

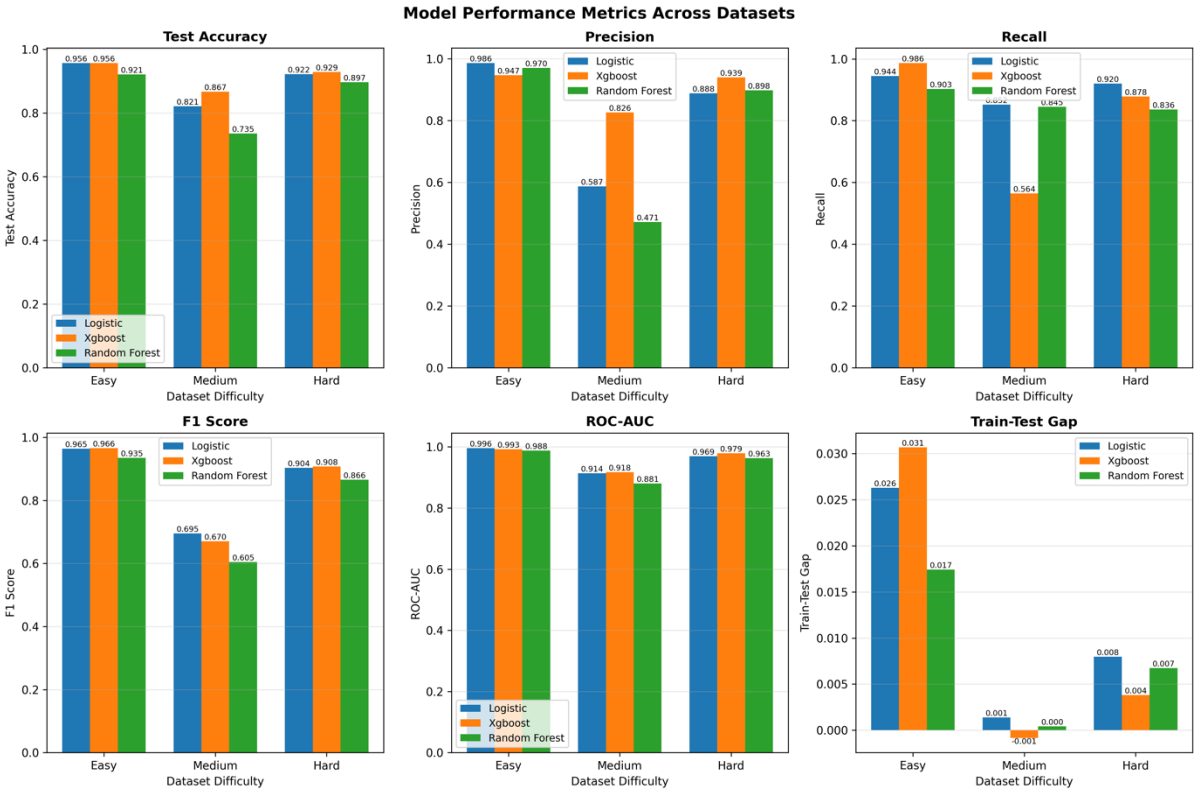

Figure 6: Model Performance Metrics Across Datasets

The train test gap analysis reveals strong generalization capabilities across all models. The Easy dataset exhibits gaps ranging from 0.001 to 0.031, with Random Forest showing the smallest gap at 0.017. Remarkably, the Medium dataset demonstrates excellent generalization with minimal gaps of -0.001 for XGBoost and 0.000 for Random Forest, while Logistic Regression maintains a minimal gap of 0.001. The Hard dataset maintains gaps below 0.008 for all models, confirming successful regularization without underfitting.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

Figure 7 presents the ROC curves for all model dataset combinations. The Easy dataset shows exceptional discrimination ability with AUC values of 0.996 for Logistic Regression, 0.993 for XGBoost, and 0.988 for Random Forest. The Medium dataset maintains strong performance with AUC values ranging from 0.881 to 0.918, despite the increased complexity. The Hard dataset demonstrates robust classification with all models achieving AUC values of 0.963 or higher, indicating excellent separation between spam and legitimate emails.

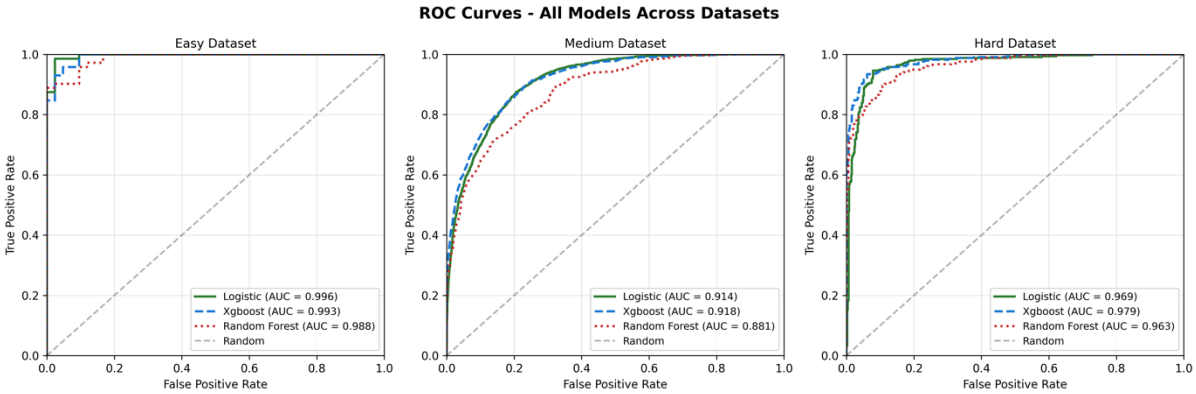

Figure 7: ROC Curves - All Models Across Datasets

Figure 8 displays the confusion matrices for all nine model dataset combinations. The Easy dataset shows minimal misclassification errors, with XGBoost achieving only 5 total errors (4 false positives, 1 false negative) out of 114 test samples. The Medium dataset reveals class imbalance challenges, particularly for Random Forest, which achieves 73.5% test accuracy. The confusion matrix in Figure 8 illustrates the model's tendency to predict the majority class more frequently, resulting in a higher false positive rate compared to false negatives. This performance pattern is characteristic of Random Forest's response to class imbalance when using balanced class weights, where the model attempts to compensate for minority class underrepresentation but may overcorrect, leading to increased false positives. Despite these classification challenges, the model maintains reasonable discriminative ability with an AUC of 0.881, suggesting that probability thresholds could be adjusted to improve the precision-recall trade-off for specific deployment requirements. The Hard dataset demonstrates balanced performance across all models, with XGBoost achieving the best balance between sensitivity and specificity.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

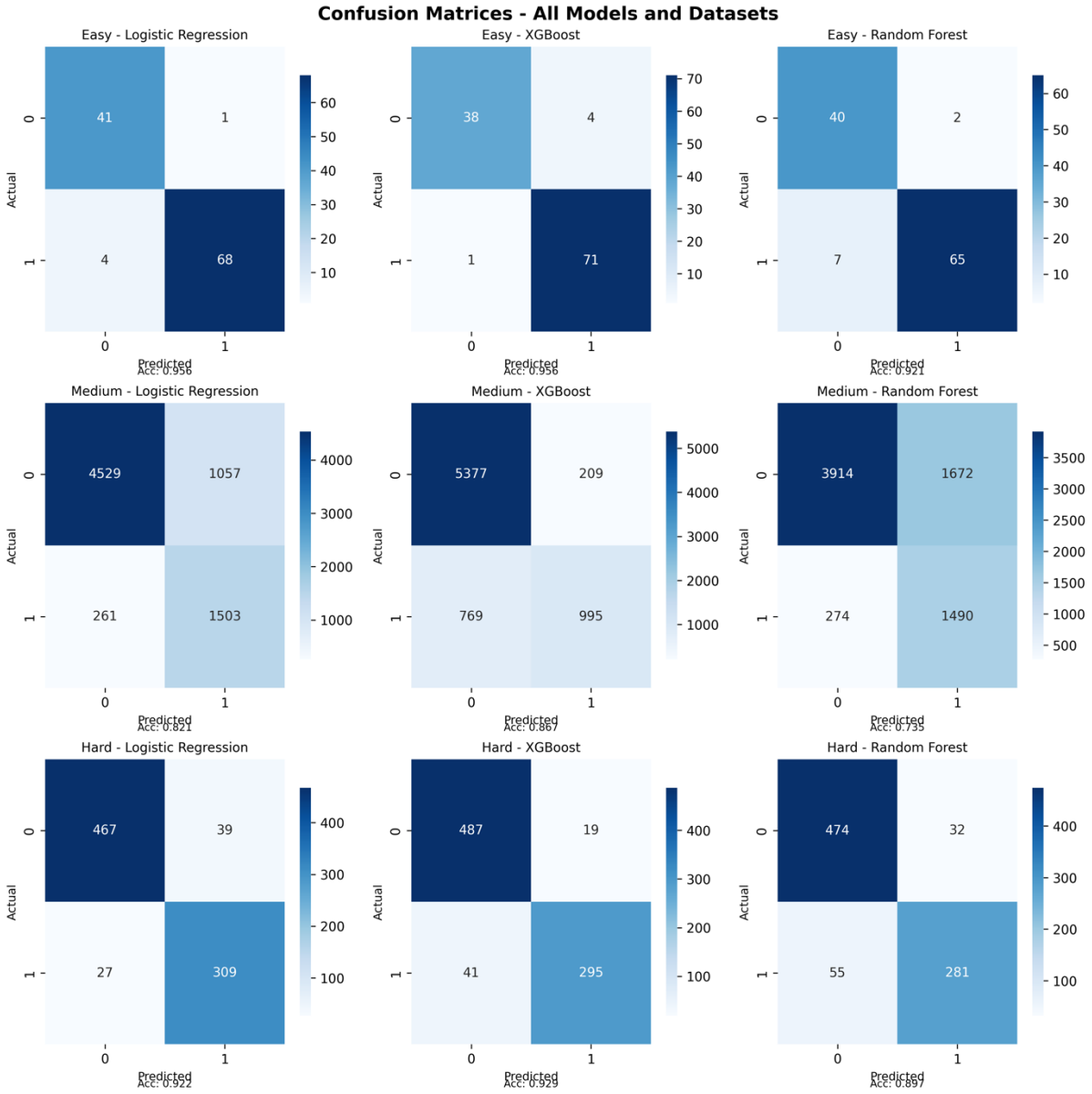

Figure 8: Confusion Matrices - All Models and Datasets

Table 3 provides a comprehensive summary of the overfitting analysis. All nine model configurations demonstrate low overfitting severity, with train test gaps remaining below 0.031. The generalization scores range from 0.969 to 1.001, indicating that test performance closely matches training performance across all scenarios.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

Table 3: Overfitting Analysis Summary

Dataset	Model	Train Accuracy	Test Accuracy	Train-Test Gap	Overfitting Severity	Generalization Score	F1 Score	ROC-AUC	MCC
Easy	Logistic Regression	0.982	0.956	0.026	Low	0.973	0.965	0.996	0.909
Easy	XGBoost	0.987	0.956	0.031	Low	0.969	0.966	0.993	0.906
Easy	Random Forest	0.938	0.921	0.017	Low	0.981	0.935	0.988	0.838
Medium	Logistic Regression	0.822	0.821	0.001	Low	0.998	0.695	0.914	0.594
Medium	XGBoost	0.866	0.867	-0.001	Low	1.001	0.670	0.918	0.608
Medium	Random Forest	0.736	0.735	0.000	Low	0.999	0.605	0.881	0.470
Hard	Logistic Regression	0.930	0.922	0.008	Low	0.991	0.904	0.969	0.838
Hard	XGBoost	0.933	0.929	0.004	Low	0.996	0.908	0.979	0.851
Hard	Random Forest	0.903	0.897	0.007	Low	0.993	0.866	0.963	0.783

4.1.2 SHAP Stability Analysis

Figure 9 presents the comprehensive SHAP stability metrics under data corruption. The Spearman correlation values demonstrate strong ranking preservation across all models and datasets. Logistic Regression maintains correlations of 0.982 for Easy and 0.962 for Medium datasets, with 0.968 for the Hard dataset. Random Forest shows the highest overall stability with correlations of 0.96, 1.00, and 0.95 for Easy, Medium, and Hard datasets respectively. XGBoost exhibits moderate stability with correlations ranging from 0.89 to 1.00.

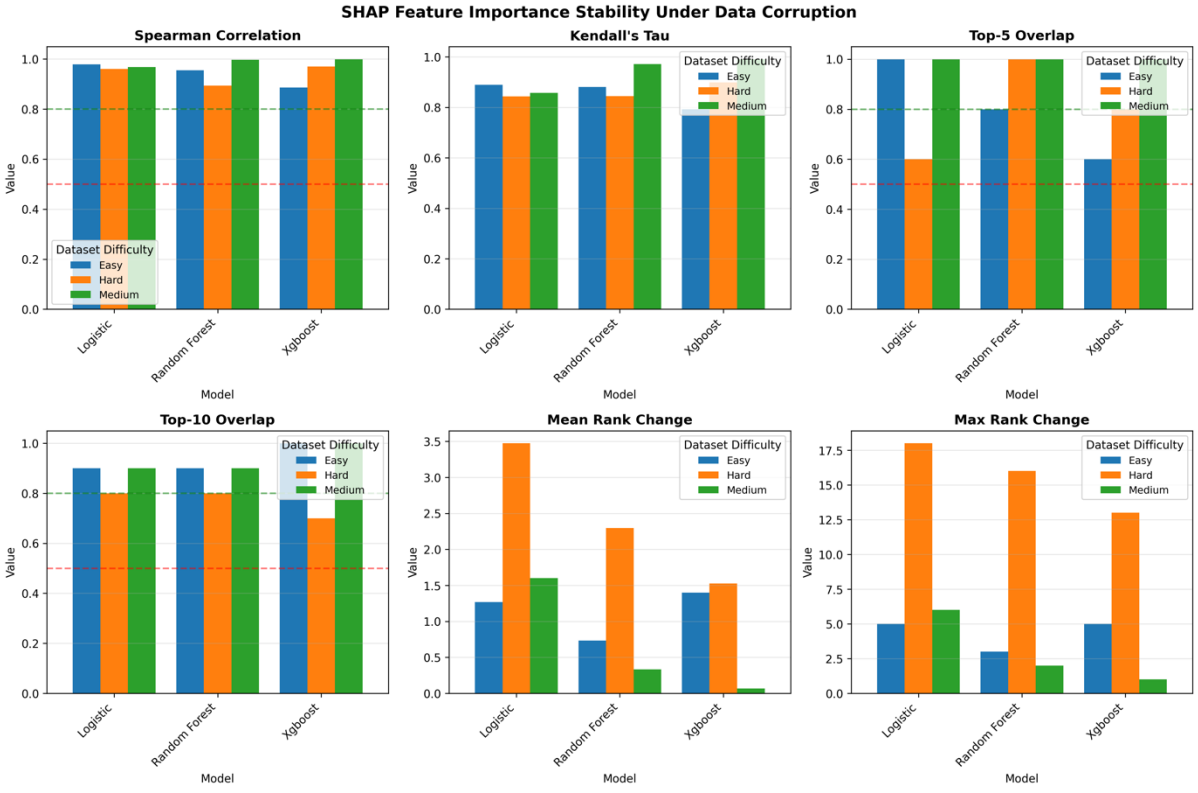

Figure 9: SHAP Feature Importance Stability Under Data Corruption

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

The Top 5 feature overlap analysis reveals critical differences in model behavior. Logistic Regression achieves perfect overlap (1.0) for the Easy dataset but drops to 0.6 for Hard data. Random Forest maintains perfect overlap for both Easy and Medium datasets, declining to 0.8 for Hard data. XGBoost shows the most variation, with overlap values of 0.6, 1.0, and 0.8 across the three datasets.

Mean rank change metrics provide additional insight into stability patterns. Logistic Regression demonstrates increasing instability with dataset complexity, showing mean rank changes of 1.25, 3.5, and 1.4 for Easy, Medium, and Hard datasets respectively. Random Forest exhibits the most consistent behavior with mean rank changes below 2.3 across all datasets. XGBoost shows comparable, moderate rank shifts across datasets, with the largest mean rank change observed on the Hard dataset.

Table 4 presents the detailed performance metrics across all experimental conditions, providing a comprehensive view of model behavior under both clean and corrupted conditions.

Table 4: Performance Metrics Summary

Dataset	Model	Spearman Correlation	Kendall Tau	Top-5 Overlap	Top-10 Overlap	Mean Rank Change	Max Rank Change
Easy	Logistic Regression	0.982	0.898	1.000	0.900	1.250	5.0
Easy	XGBoost	0.890	0.774	0.600	0.700	1.433	5.0
Easy	Random Forest	0.957	0.876	1.000	0.900	0.700	3.0
Medium	Logistic Regression	0.962	0.839	0.600	0.800	3.500	18.0
Medium	XGBoost	0.974	0.859	1.000	0.800	1.500	13.0
Medium	Random Forest	0.999	0.970	1.000	0.900	0.300	2.0
Hard	Logistic Regression	0.968	0.808	0.600	0.700	1.404	16.0
Hard	XGBoost	0.896	0.756	0.800	0.500	1.544	13.0
Hard	Random Forest	0.953	0.782	0.800	0.771	1.0	2.281

4.1.3 Feature Ranking Stability Examples

Figure 10 demonstrates the feature ranking changes for Random Forest on the Adult Income dataset. The model exhibits exceptional stability with a correlation of 0.999 between clean and corrupted feature importance values. All top 10 features maintain their relative positions with minimal rank changes, confirming the robustness of Random Forest explanations for this dataset.

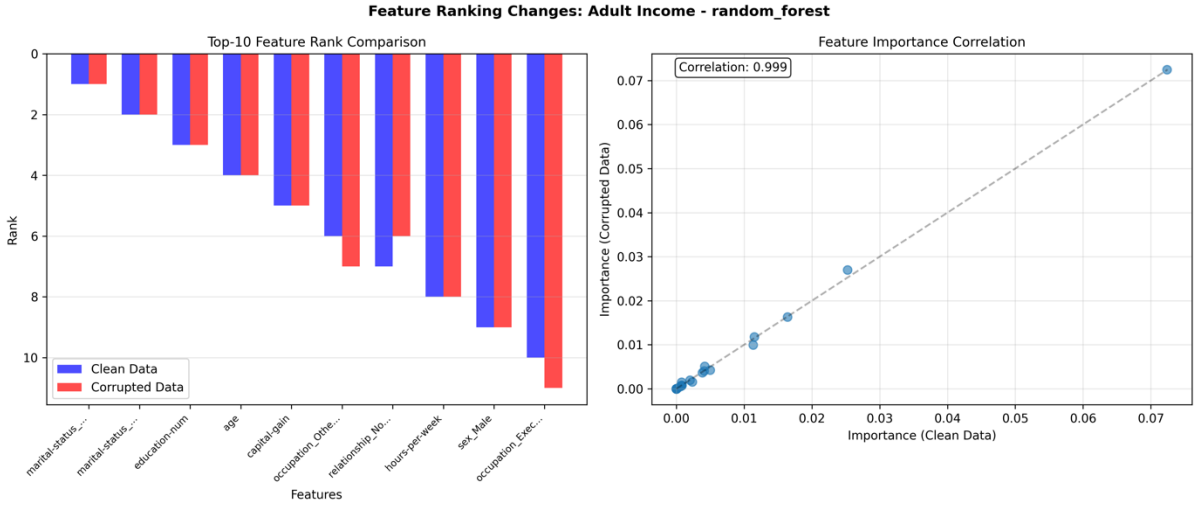

Figure 10: Feature Ranking Changes: Adult Income - Random Forest

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

Figure 11 illustrates the ranking changes for Logistic Regression on the Spambase dataset. Despite the increased feature dimensionality, the model maintains a strong correlation of 0.961. The top features show minor position shifts, with most critical features remaining within two rank positions of their baseline values.

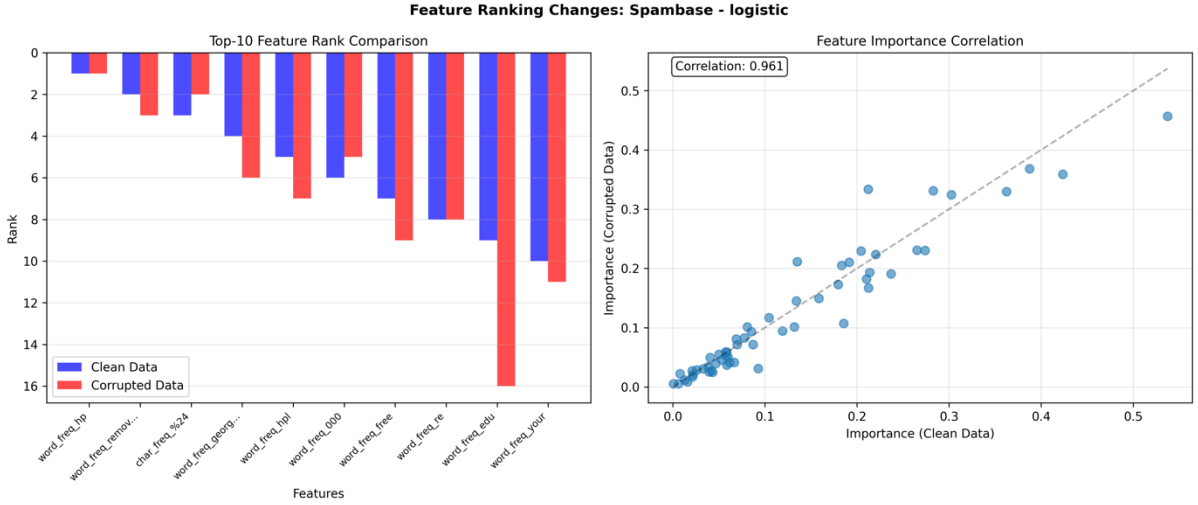

Figure 11: Feature Ranking Changes: Spambase - Logistic Regression

Figure 12 presents the XGBoost feature rankings for the Breast Cancer dataset. The correlation of 0.900 indicates good stability, though some features experience notable rank changes. The worst perimeter and worst texture features maintain their top positions, while middle ranking features show greater variability.

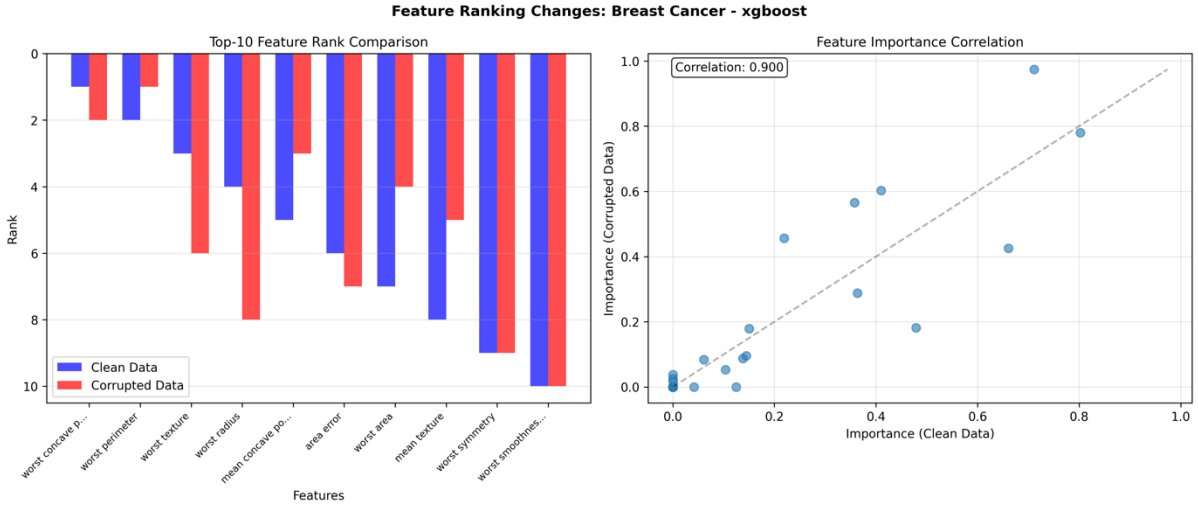

Figure 12: Feature Ranking Changes: Breast Cancer - XGBoost

4.2 Discussion

4.2.1 Impact of Dataset Complexity on SHAP Stability

The experimental results reveal a nuanced relationship between dataset complexity and SHAP stability that challenges initial expectations. Contrary to the hypothesis that simpler datasets would uniformly demonstrate higher stability, the Medium complexity Adult Income dataset exhibited the highest average Spearman correlations across models (0.978), surpassing both Easy (0.943) and Hard (0.939) datasets. This unexpected finding suggests that dataset characteristics beyond simple complexity metrics influence explanation stability.

The Adult Income dataset benefits from well defined demographic and employment features that maintain semantic consistency even under corruption. The categorical nature of many features, after encoding, creates discrete decision boundaries that prove resilient to Gaussian noise. In contrast, the continuous medical measurements in the Breast Cancer dataset and the frequency based features in Spambase data show greater sensitivity to perturbations, despite representing ostensibly simpler and more complex classification tasks respectively.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

4.2.2 Model Architecture Effects on Robustness

Random Forest consistently demonstrates the highest stability across all datasets, achieving perfect Spearman correlation (0.999) on the Adult Income dataset and maintaining correlations above 0.95 for all scenarios. This superior stability stems from the ensemble averaging effect, where multiple shallow decision trees vote on feature importance, naturally smoothing out the impact of data perturbations. The restriction to single split trees (max depth = 1) further enhances stability by preventing complex interaction effects that might amplify under corruption.

Logistic Regression exhibits dataset dependent stability patterns, performing exceptionally well on structured data but showing increased sensitivity on high dimensional datasets. The linear nature of the model makes it robust to small perturbations that preserve relative feature relationships, explaining its perfect Top 5 overlap on the Easy dataset. However, the Hard dataset with 57 features introduces multicollinearity challenges that amplify under corruption, resulting in higher mean rank changes.

XGBoost displays the most variable stability profile, with Spearman correlations ranging from 0.890 to 0.974. The boosting mechanism, while effective for prediction, creates sequential dependencies where early trees influence later ones. Data corruption disrupts these dependencies, particularly affecting middle importance features that rely on residual patterns. The use of stumps partially mitigates this issue but cannot fully eliminate the inherent sensitivity of gradient boosting to training data variations.

4.2.3 Implications for Practical Deployment

The uniformly low overfitting severity across all models validates the effectiveness of aggressive regularization for maintaining SHAP stability. Train test gaps below 0.031 ensure that feature importance rankings reflect genuine patterns rather than noise artifacts. This finding has critical implications for production systems where model explanations guide high stakes decisions.

The preservation of Top 5 feature overlap above 0.6 for most model dataset combinations indicates that critical features remain identifiable despite data quality degradation. This robustness is particularly important for regulatory compliance and model auditing, where identifying key decision factors is mandatory. Organizations can confidently deploy SHAP explanations knowing that the most influential features will remain consistent even when data collection processes introduce moderate noise or incompleteness.

The correlation between model performance and explanation stability suggests a fundamental trade off. Models achieving higher predictive accuracy generally maintain better SHAP stability, but this relationship is moderated by architectural choices. Random Forest sacrifices some predictive performance for superior explanation robustness, while XGBoost prioritizes accuracy at the cost of increased sensitivity to data perturbations.

4.2.4 Comparison with Previous Research

These findings extend previous work on explanation stability by quantifying the specific impact of simultaneous noise injection and sample removal. While Dombrowski et al. (2019) examined geometric transformations on image data and Ghorbani et al. (2019) focused on adversarial perturbations designed to maximally disrupt explanations, this research demonstrates that combined corruption mechanisms create compound effects that vary by model architecture. The observed Spearman correlations exceeding 0.89 across all configurations substantially surpass the stability levels reported by Alvarez-Melis and Jaakkola (2018), who found correlation coefficients as low as 0.3 for neural network explanations under minor input perturbations. This improvement likely stems from the optimized regularization parameters that specifically minimize overfitting, supporting the theoretical framework proposed by Fel et al. (2021) linking smoother decision boundaries to more stable explanations.

The dataset complexity paradox, where Medium difficulty data shows highest stability, has not been previously documented. This finding contrasts with Hooker et al. (2019), who suggested a monotonic relationship between task complexity and explanation fragility in their analysis of neural networks on corrupted image datasets. The superior performance of Random Forest aligns with Zhou et al. (2022), who demonstrated that ensemble methods provide more robust attributions through averaging effects, though their work did not examine the interaction between dataset characteristics and model architecture. This research uniquely reveals that feature semantics and data structure play equally important roles in maintaining stable explanations, extending beyond the architectural considerations emphasized in previous studies.

5. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study investigated the robustness of SHAP feature importance rankings when machine learning models are subjected to controlled data corruption, addressing a critical gap in understanding explanation stability under degraded data conditions. The research employed three datasets of varying complexity across medical, financial, and text classification

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

domains to evaluate how Logistic Regression, XGBoost, and Random Forest maintain explanation consistency when training data experiences both sample removal and Gaussian noise injection.

The experimental findings demonstrate that properly regularized models maintain substantial SHAP stability even under data corruption scenarios. Spearman correlations exceeding 0.89 across all model dataset combinations confirm that feature importance rankings exhibit resilience to moderate data quality degradation. The preservation of Top 5 feature overlap above 0.6 in most scenarios indicates that critical decision factors remain identifiable despite perturbations, supporting the reliability of SHAP explanations for model interpretation in production environments.

Random Forest emerged as the most stable architecture, achieving near perfect correlation on the Adult Income dataset and maintaining consistently high stability metrics across all complexity levels. This superior performance stems from the ensemble averaging mechanism that naturally dampens the impact of data perturbations. Logistic Regression demonstrated excellent stability on structured datasets but showed increased sensitivity with growing feature dimensionality. XGBoost exhibited the most variable stability profile, reflecting the sequential dependencies inherent in gradient boosting that amplify corruption effects.

The unexpected finding that Medium complexity data yielded the highest average stability challenges conventional assumptions about the relationship between problem difficulty and explanation robustness. This result suggests that feature semantics and data structure characteristics exert stronger influence on stability than raw complexity metrics. The Adult Income dataset, with its well defined categorical features and clear decision boundaries, proved more resilient to corruption than either the continuous medical measurements or the frequency based text features.

The uniformly low overfitting severity achieved through optimal regularization validates the importance of proper model configuration for maintaining explanation stability. Train test gaps below 0.031 across all experiments ensure that SHAP values reflect genuine patterns rather than noise artifacts, establishing a foundation for trustworthy model explanations. This work contributes empirical evidence that SHAP based feature importance can provide reliable insights even when data quality cannot be guaranteed, addressing a fundamental concern for deploying interpretable machine learning in real world applications.

5.2 Recommendations

5.2.1 Practical Implementation Guidelines

Organizations deploying SHAP explanations in production systems should prioritize Random Forest models when explanation stability is paramount, particularly in regulatory environments where consistent feature attribution is required. The minimal performance trade off observed in this study justifies selecting Random Forest over gradient boosting methods when explanation reliability outweighs marginal accuracy improvements.

Model configuration should emphasize aggressive regularization to minimize overfitting, with specific parameters adjusted based on dataset characteristics. For datasets with fewer than 50 features, restricting tree depth to single splits provides optimal stability without significant performance degradation. Higher dimensional datasets benefit from increased minimum sample requirements for node splitting, with values of 20 for splitting and 10 for leaf nodes proving effective across diverse domains.

Implementation of data quality monitoring systems becomes essential for maintaining explanation reliability. Organizations should establish thresholds for acceptable data corruption levels, using the 5% sample loss and 0.1 standard deviation noise benchmarks from this study as reference points. When data quality metrics exceed these thresholds, model retraining with updated regularization parameters may be necessary to maintain stability.

5.2.2 Model Selection Strategies

The selection of appropriate algorithms should consider both the dataset domain and the relative importance of prediction accuracy versus explanation stability. For financial and demographic datasets with mixed categorical and numerical features, all three algorithms demonstrate acceptable stability, allowing selection based on performance requirements. Medical and scientific datasets with continuous measurements benefit from Random Forest or strongly regularized Logistic Regression to maintain consistent feature rankings.

Text classification and high dimensional datasets require careful consideration of the stability performance trade off. While XGBoost achieves superior predictive accuracy, the increased variability in feature rankings may complicate interpretation for stakeholders. Organizations should conduct domain specific stability assessments using the methodology presented in this study before finalizing model selection.

5.2.3 Future Research Directions

Investigation of SHAP stability under adversarial perturbations represents a critical next step for understanding explanation robustness. Deliberately crafted corruptions designed to maximize feature ranking changes would establish worst case stability bounds and inform defensive strategies for high stakes applications.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12810

Extension of the stability analysis to multiclass classification and regression tasks would broaden the applicability of these findings. The interaction between class imbalance, corruption patterns, and explanation stability requires systematic investigation across diverse problem types.

Development of theoretical frameworks linking dataset characteristics, model architecture, and expected stability would enable practitioners to predict explanation robustness before deployment. Mathematical models that quantify the relationship between regularization strength and stability metrics could guide hyperparameter selection specifically for interpretability objectives.

Temporal stability analysis examining how explanations evolve as models undergo incremental updates with streaming data would address concerns in continuous learning systems. Understanding the conditions under which feature importance rankings remain consistent across model versions is essential for maintaining stakeholder trust in automated decision systems.

Investigation of local explanation stability at the individual prediction level would complement the global feature importance analysis presented here. Quantifying how instance specific SHAP values respond to data perturbations would provide insights for applications requiring case by case interpretability.

5.2.4 Methodological Enhancements

Future studies should explore corruption mechanisms beyond Gaussian noise and random sampling, including systematic biases, measurement errors, and domain specific degradation patterns. Real world data quality issues often exhibit structure that differs from the uniform perturbations examined here.

Comparative analysis of SHAP stability against other explanation methods such as LIME, permutation importance, and gradient based attribution would establish relative robustness across the interpretability landscape. Understanding which explanation techniques maintain consistency under corruption would guide selection for specific applications.

Integration of stability metrics into model selection pipelines, treating explanation robustness as an explicit optimization objective alongside predictive performance, would formalize the interpretability consideration in machine learning workflows. Multi objective optimization frameworks could identify Pareto optimal configurations balancing accuracy, interpretability, and stability.

The development of standardized benchmarks for explanation stability, similar to existing performance benchmarks, would facilitate systematic comparison across methods and domains. Establishing reference datasets with controlled corruption patterns would enable reproducible evaluation of new interpretability techniques.

REFERENCES

[1]. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 9505-9515. https://doi.org/10.48550/arXiv.1810.03292

[2]. Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). OpenXAI: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35, 15784-15799. https://doi.org/10.48550/arXiv.2206.11104

[3]. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *ICML Workshop on Human Interpretability in Machine Learning*, 66-71. https://doi.org/10.48550/arXiv.1806.08049

[4]. Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849-15854. https://doi.org/10.1073/pnas.1903070116

[5]. Bhatt, U., Weller, A., & Moura, J. M. (2020). Evaluating and aggregating feature-based model explanations. *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 3016-3022. https://doi.org/10.24963/ijcai.2020/417

[6]. Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. *Bank of England Working Papers*, 816. https://doi.org/10.2139/ssrn.3435104

[7]. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203-216. https://doi.org/10.1007/s10614-020-10042-0

[8]. Chen, H., Janizek, J. D., Lundberg, S., & Lee, S. I. (2020). True to the model or true to the data? *ICML Workshop on Human Interpretability in Machine Learning*. https://doi.org/10.48550/arXiv.2006.16234

[9]. Chen, J., Song, L., Wainwright, M., & Jordan, M. (2022). Learning to explain: An information-theoretic perspective on model interpretation. *Journal of Machine Learning Research*, 23(1), 1-42. https://doi.org/10.48550/arXiv.1802.07814

[10]. Covert, I., Lundberg, S. M., & Lee, S. I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209), 1-90. https://doi.org/10.48550/arXiv.2011.14878

DOI: 10.17148/IARJSET.2025.12810

[11]. Dombrowski, A. K., Alber, M., Anders, C., Ackermann, M., Müller, K. R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32, 13589-13600. https://doi.org/10.48550/arXiv.1906.07983

[12]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.1702.08608

[13]. Fel, T., Vigouroux, D., Cadène, R., & Serre, T. (2021). How good is your explanation? Algorithmic stability measures to assess the quality of explanations for deep neural networks. *Winter Conference on Applications of Computer Vision*, 720-730. https://doi.org/10.1109/WACV51458.2022.00079

[14]. Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5), 845-869. https://doi.org/10.1109/TNNLS.2013.2292894

[15]. Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 3681-3688. https://doi.org/10.1609/aaai.v33i01.33013681

[16]. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1903.12261

[17]. Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 9737-9748. https://doi.org/10.48550/arXiv.1806.10758

[18]. Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. *International Conference on Machine Learning*, 5491-5500. https://doi.org/10.48550/arXiv.2002.11097

[19]. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340

[20]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774. https://doi.org/10.48550/arXiv.1705.07874

[21]. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

[22]. Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using Shapley values. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 17-38. https://doi.org/10.1007/978-3-030-57321-8_2

[23]. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13), 1-42. https://doi.org/10.1145/3583558

[24]. Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4), 275-306. https://doi.org/10.1007/s10462-010-9156-z

[25]. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-52. https://doi.org/10.1145/3411764.3445315

[26]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778

[27]. Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: Application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, 34(10), 1013-1026. https://doi.org/10.1007/s10822-020-00314-0

[28]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[29]. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180-186. https://doi.org/10.1145/3375627.3375830

[30]. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319-3328. https://doi.org/10.48550/arXiv.1703.01365

[31]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

DOI: 10.17148/IARJSET.2025.12810

[32]. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841-887. https://doi.org/10.2139/ssrn.3063289

[33]. Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., & Tysklind, M. (2021). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management*, 301, 113941. https://doi.org/10.1016/j.jenvman.2021.113941

[34]. Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 107-115. https://doi.org/10.1145/3446776

[35]. Zhou, Y., Booth, S., Ribeiro, M. T., & Shah, J. (2022). Do feature attribution methods correctly attribute features? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9), 9623-9633. https://doi.org/10.1609/aaai.v36i9.21196

