International Journal of Artificial Intelligence, Machine Learning and Data Analytics

www.ijaimda.org

ISSN: 2839-5557 | **Open Access** | **DOI:** https://doi.org/10.68033/ijaimda.1.2.28-42

Explainable Artificial Intelligence and Machine Learning for Responsible Decision-Making: A Comparative Study of Model Accuracy, Transparency, and Fairness

Nawaz Ali Hamdulay[1], Shazia Sameer Khan[2]

[1]Department of Engineering & Management, Infomania IT and Management Academy LLP, Mumbai, India
Email id: nawaz.hamdulay@gmail.com, ORCID iD: https://orcid.org/0009-0007-7377-9990
[2]Department of Mathematics & Science, Noor Al Khaleej International School, Doha, Qatar
Email id: shazia.k@faculty.nooralkhaleej.net, ORCID iD: https://orcid.org/0009-0008-7030-4313

Corresponding Author: Nawaz Ali Hamdulay, nawaz.hamdulay@gmail.com

ARTICLE INFO

Article History:

Received 07/09/2026
Revised 21/10/2026
Accepted 12/11/2026

CITE AN ARTICLE:

Hamdulay N. A., Khan S. S., (2026). ***Explainable Artificial Intelligence and Machine Learning for Responsible Decision-Making: A Comparative Study of Model Accuracy, Transparency, and Fairness****. International Journal of Artificial Intelligence, Machine Learning and Data Analytics, 1(2), 28-42*
https://doi.org/10.68033/ijaimda.1.2.28-42

ABSTRACT

Explainable Artificial Intelligence (XAI) has emerged as an important research area in developing transparent, accountable, and trustworthy machine learning systems. Although machine learning models can achieve high predictive accuracy, their complex decision-making processes may limit interpretability and create challenges concerning fairness, accountability, and responsible implementation. This study aims to comparatively evaluate artificial intelligence and machine learning models based on three critical dimensions: predictive accuracy, transparency, and fairness. The proposed research examines interpretable algorithms, including Logistic Regression and Decision Trees, alongside complex models such as Random Forest, Gradient Boosting, and other ensemble learning techniques. Explainability methods, particularly SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), are considered for interpreting model predictions and identifying influential features. Model performance will be assessed using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve, while transparency and fairness will be evaluated through explanation quality, interpretability, and appropriate group fairness metrics. The study also investigates potential trade-offs between predictive performance, model complexity, and equitable outcomes. Drawing on established research in explainable AI, explanation evaluation, and algorithmic fairness, the research proposes a comparative framework for identifying models suitable for responsible decision-making. The findings are expected to support researchers, developers, and organizations in selecting machine learning approaches that balance predictive effectiveness with transparency, fairness, and accountability.

Keywords **-** Explainable Artificial Intelligence; Machine Learning; Responsible Decision-Making; Model Accuracy; Transparency; Algorithmic Fairness; Model Interpretability; SHAP; LIME; Ethical Artificial Intelligence.

International Journal of Artificial Intelligence, Machine Learning and Data Analytics is licensed under a Creative Commons Attribution 4.0 International License

1. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have become influential technologies for analyzing complex datasets, identifying patterns, predicting future outcomes, and supporting data-driven decision-making. These technologies are increasingly applied in banking, healthcare, education, recruitment, financial services, customer analytics, and public administration. Machine learning algorithms can improve operational efficiency and prediction quality by learning relationships from historical data. However, the increasing complexity of advanced AI systems has raised concerns about the transparency, fairness, reliability, and accountability of their decisions. Consequently, the development of responsible AI systems requires more than predictive accuracy; it also requires an understanding of how models generate predictions and whether their outcomes create unjustified disadvantages for particular groups.

Explainable Artificial Intelligence (XAI) refers to methods and techniques designed to make AI model predictions more understandable to humans. Arrieta et al. (2020) examined the concepts, taxonomies, opportunities, and challenges associated with explainable AI, highlighting its relevance to responsible AI development. Explainability can help researchers and decision-makers understand the features influencing predictions, identify unexpected model behaviour, and evaluate whether a system's outputs are consistent with the intended decision-making objectives. Nevertheless, providing an explanation does not automatically guarantee that a model is accurate, unbiased, or ethically appropriate. Therefore, explainability should be evaluated alongside other performance and governance considerations.

Several techniques have been developed to explain machine learning predictions. Local Interpretable Model-Agnostic Explanations (LIME) provides locally interpretable approximations of complex models, while SHapley Additive exPlanations (SHAP) use Shapley values to estimate the contribution of individual features to predictions. These techniques can assist in interpreting models that would otherwise be difficult to understand directly. However, the usefulness of explanations depends on their consistency, fidelity, clarity, and relevance to the decision-making context. Research on evaluating machine learning explanations emphasizes the importance of assessing explanation quality through appropriate quantitative and human-centred evaluation methods (Zhou et al., 2021).

Another important consideration is algorithmic fairness. Machine learning systems learn from historical and contemporary data, which may contain sampling biases, incomplete information, or patterns reflecting existing social inequalities. When such patterns influence model training, predictions may disadvantage particular demographic groups. Mehrabi et al. (2021) reviewed the sources of bias and fairness challenges in machine learning and discussed approaches to understanding and mitigating algorithmic discrimination. Therefore, responsible AI development requires the examination of model outcomes across relevant population groups rather than relying exclusively on aggregate performance measures.

An additional challenge concerns the relationship between predictive accuracy and fairness. A model that performs well overall may produce substantially different error rates across demographic groups. Conversely, reducing disparities between groups may require changes to model design, training data, decision thresholds, or optimization objectives. Valdivia et al. (2021) investigated the boundaries between accuracy and fairness and demonstrated the importance of evaluating their relationship when developing machine learning systems. Such findings suggest that model selection should consider multiple performance dimensions instead of assuming that the algorithm with the highest accuracy is necessarily the most appropriate.

The effectiveness of explainable AI also depends on whether explanations genuinely improve human understanding and decision-making. Wang and Yin (2022) compared explanation methods in AI-assisted decision-making and examined their ability to support model understanding, recognition of uncertainty, and appropriately calibrated trust. Their research highlights why the presence of an explanation alone is insufficient to establish its practical value. An explanation should help decision-makers understand the basis of a prediction without creating unjustified confidence in the system.

In this context, the present study, titled **"Explainable Artificial Intelligence and Machine Learning for Responsible Decision-Making: A Comparative Study of Model Accuracy, Transparency, and Fairness,"** proposes a systematic comparison of interpretable and complex machine learning models. The research will assess predictive performance using standard classification metrics, examine transparency through explainability techniques, and evaluate fairness through suitable group-based measures. Logistic Regression and Decision Trees will be compared with Random Forest and Gradient Boosting models to investigate differences between inherently interpretable algorithms and more complex predictive approaches.

The study is intended to contribute to AI and machine learning research by developing a comparative evaluation framework that considers accuracy, transparency, and fairness together. Rather than treating these dimensions as independent objectives, the research will examine their relationships and potential trade-offs. Its expected contribution

is a more comprehensive basis for selecting machine learning models for responsible decision-making in settings where prediction quality, interpretability, and equitable treatment are important.

2. LITERATURE REVIEW

Salih et al. [2025], in their study on explainable artificial intelligence methods, examine the significance of SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) in improving machine learning interpretability. The study highlights how these techniques help explain model predictions by identifying influential features and clarifying the contribution of individual variables. SHAP provides consistent feature-attribution explanations, whereas LIME approximates model behaviour locally to explain individual predictions. These methods support transparency, trust, and accountability in AI-assisted decision-making. This literature is relevant to comparing model accuracy, interpretability, and fairness within responsible machine learning frameworks (Salih et al., 2025).

Pujari et al. [2026], review developments in explainable artificial intelligence (XAI) evaluation and ethics published between 2023 and 2025. Their study presents a classification framework distinguishing intrinsic and post-hoc explainability, as well as local and global explanations. The authors emphasize evaluating XAI through technical measures, including fidelity, robustness, and complexity, alongside human-centred considerations such as decision impact and interpretability. The review also discusses fairness, privacy, scalability, regulatory compliance, and ethical challenges across healthcare, finance, cybersecurity, and law. This research supports developing responsible machine learning frameworks that integrate predictive accuracy, transparency, fairness, and accountability (Pujari et al., 2026).

Perdana [2026], examines the relationship between algorithmic fairness, procedural justice, and explainable artificial intelligence (XAI). The study highlights that satisfying statistical fairness metrics does not necessarily ensure that individuals perceive algorithmic decisions as fair or legitimate. It identifies transparency, contestability, and correctability as important procedural safeguards for responsible AI systems. The proposed fairness-by-design framework emphasizes stakeholder participation, explainable decisions, and systematic documentation throughout the AI development lifecycle. The study also explains how SHAP and LIME can support decision transparency when appropriately implemented. This research provides a conceptual foundation for integrating accuracy, interpretability, fairness, and accountability into responsible machine learning model selection (Perdana, 2026).

3. OBJECTIVES

The study will pursue the following four objectives:

1. **To compare the predictive accuracy and performance** of selected machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting, using appropriate classification evaluation metrics.
2. **To evaluate the transparency and interpretability** of selected machine learning models using explainable AI techniques, particularly SHAP and LIME.
3. **To assess the fairness of machine learning models** by examining group-based performance differences, potential algorithmic bias, and appropriate fairness metrics across relevant demographic groups.
4. **To investigate the relationship between accuracy, transparency, and fairness** and propose a comparative framework for selecting machine learning models that support responsible decision-making.

4. HYPOTHESIS

The following hypotheses are proposed for empirical testing.

H1: There is a statistically significant difference in predictive performance among the selected machine learning algorithms.

H2: Interpretable machine learning models demonstrate greater inherent transparency than complex ensemble models.

H3: The application of explainable AI techniques significantly improves the interpretability of predictions generated by complex machine learning models.

H4: There is a statistically significant relationship between predictive accuracy, model transparency, and algorithmic fairness.

These hypotheses will be evaluated using the results of model training, performance testing, explanation analysis, and fairness assessment. Statistical significance will be determined using an appropriate significance level, such as $p < .05$, where applicable. H2 concerns inherent model transparency, which will be operationalized using predefined interpretability criteria rather than assumed from predictive performance alone.

5. RESEARCH METHODOLOGY

This study will adopt a quantitative, comparative experimental design using a labelled dataset suitable for classification and containing demographic attributes for fairness assessment. Sample size of 100. After data cleaning, missing-value treatment, encoding, and feature scaling where necessary, the dataset will be divided into training and test sets using stratified sampling. Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting will be trained under consistent validation procedures. Accuracy, precision, recall, F1-score, and ROC-AUC will measure predictive performance. SHAP and LIME will support explanation analysis, while group fairness metrics will assess disparities. Statistical tests will compare model outcomes, subject to data suitability and assumptions.

6. COMPARE THE PREDICTIVE ACCURACY AND PERFORMANCE OF SELECTED MACHINE LEARNING ALGORITHMS, INCLUDING LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST, AND GRADIENT BOOSTING, USING APPROPRIATE CLASSIFICATION EVALUATION METRICS

Sr. No.	Machine Learning Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
1	Logistic Regression	84.5	83.2	81.7	82.4	88.6
2	Decision Tree	82.1	80.4	79.6	80.0	83.5
3	Random Forest	89.3	88.1	87.5	87.8	93.2
4	Gradient Boosting	91.2	90.4	89.7	90.0	95.1

Table 1: Comparative Analysis of Predictive Accuracy and Performance of Machine Learning Algorithms

Source: Values prepared for research framework development

Interpretation of Results

The illustrative results indicate that Gradient Boosting achieves the highest predictive performance, with an accuracy of 91.2%, an F1-score of 90.0%, and a ROC-AUC of 95.1%. Random Forest ranks second, achieving an accuracy of 89.3% and a ROC-AUC of 93.2%.

Logistic Regression achieves an accuracy of 84.5%, while Decision Tree records the lowest accuracy among the four algorithms at 82.1%. These illustrative results suggest that ensemble learning methods may provide stronger predictive performance than the individual interpretable models in this hypothetical comparison.

Based on these illustrative values, Gradient Boosting is the best-performing algorithm in terms of accuracy, precision, recall, F1-score, and ROC-AUC. However, predictive performance alone does not establish that a model is the most suitable for responsible decision-making. Transparency, fairness, and explainability must also be evaluated.

7. THE TRANSPARENCY AND INTERPRETABILITY OF SELECTED MACHINE LEARNING MODELS USING EXPLAINABLE AI TECHNIQUES, PARTICULARLY SHAP AND LIME

Transparency and interpretability are essential components of responsible artificial intelligence because they help researchers understand how machine learning models generate predictions and identify the factors influencing their decisions. Although complex algorithms can achieve strong predictive performance, their internal decision-making processes may be difficult for users to understand. Explainable Artificial Intelligence (XAI) addresses this challenge by providing techniques that make model predictions more understandable and support informed decision-making. Arrieta et al. (2020) emphasized the importance of explainability in developing responsible AI systems that support transparency and accountability.

This study proposes evaluating the transparency and interpretability of four machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Logistic Regression and Decision Tree provide comparatively interpretable decision structures, whereas Random Forest and Gradient Boosting combine multiple predictive components that can make their overall behaviour more difficult to explain. The evaluation will examine how these differences affect the accessibility and clarity of model explanations.

Two explainable AI techniques, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), will be employed. SHAP estimates the contribution of individual features to a model's prediction and can support both local and global interpretation of model behaviour (Lundberg & Lee, 2017). LIME explains individual predictions by approximating the behaviour of a complex model within a local region using an

interpretable surrogate model (Ribeiro et al., 2016). Applying both techniques will enable a comparative examination of feature importance, explanation consistency, and the interpretability of individual predictions.

The proposed evaluation will consider several criteria, including explanation fidelity, consistency, feature importance, computational efficiency, and ease of interpretation. SHAP and LIME explanations will be generated for the selected models using the same dataset and comparable evaluation procedures. Global feature importance summaries will help identify influential variables across observations, while local explanations will clarify why particular predictions were produced. The study will also examine whether explanations remain consistent across repeated evaluations and whether they adequately represent the underlying model behaviour.

The research is expected to establish a systematic framework for comparing the transparency of interpretable and complex machine learning models. It will help identify the strengths and limitations of SHAP and LIME in explaining different algorithms. However, explainability alone does not establish that a model is fair, accurate, or trustworthy. Therefore, the results will be interpreted alongside predictive performance and fairness assessments to support more transparent and responsible AI-based decision-making.

8. THE FAIRNESS OF MACHINE LEARNING MODELS BY EXAMINING GROUP-BASED PERFORMANCE DIFFERENCES, POTENTIAL ALGORITHMIC BIAS, AND APPROPRIATE FAIRNESS METRICS ACROSS RELEVANT DEMOGRAPHIC GROUPS

Fairness is an essential component of responsible Artificial Intelligence (AI) because machine learning systems increasingly influence decisions in banking, recruitment, healthcare, education, and other socially significant areas. Although these systems can improve predictive efficiency, they may reproduce historical discrimination, sampling inequalities, and biases embedded in training datasets. Therefore, evaluating machine learning models requires examining not only overall accuracy but also whether predictions and error rates differ across relevant demographic groups. Mehrabi et al. (2021) highlighted the importance of understanding data-related, algorithmic, and societal sources of bias when designing fair machine learning systems.

This study proposes assessing the fairness of four machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. The evaluation will examine group-based performance differences across relevant demographic attributes, such as age and gender, where appropriate to the dataset and research application. It will investigate whether particular groups experience higher false-positive rates, lower true-positive rates, or unequal access to favourable predictions. Pessach and Shmueli (2022) explained that algorithmic unfairness may arise from biased historical data, unrepresentative samples, missing information, and optimization objectives that favour majority groups. Consequently, fairness assessment must consider both the quality of input data and the outcomes produced by trained models.

The proposed research will employ established fairness metrics, including demographic parity difference, disparate impact, equal opportunity difference, and equalized odds difference. Demographic parity difference compares positive prediction rates across groups, while disparate impact evaluates the ratio of those rates. Equal opportunity difference compares true-positive rates, and equalized odds difference examines disparities in both true-positive and false-positive rates. Mitchell et al. (2021) emphasized that fairness definitions involve different assumptions and value judgments; therefore, no single metric is sufficient for every decision-making context.

The study will calculate these metrics for each algorithm using consistent test data and clearly defined demographic groups. Results will be compared to identify potential disparities and determine whether predictive performance varies substantially between groups. Where disparities are observed, possible mitigation approaches, including data rebalancing, threshold adjustment, and fairness-aware training, may be considered. The findings are expected to support a systematic comparison of algorithmic fairness alongside accuracy and transparency, helping researchers identify machine learning models better suited to responsible decision-making.

Dataset preparation

Component	Proposed specification
Dataset	Adult Census Income
Target variable	Income above $50,000: Yes/No
Training data	80% of observations
Test data	20% of observations
Data splitting	Stratified train-test split
Random seed	42
Algorithms	Logistic Regression, Decision Tree, Random Forest, Gradient Boosting

Sensitive attributes	Sex, race, and age group
Performance metrics	Accuracy, precision, recall, F1-score, ROC-AUC
Fairness metrics	Demographic parity difference, disparate impact, equal opportunity difference, equalized odds difference

Table 2

The 80:20 split and random seed are proposed experimental settings, not completed experimental results. Preprocessing, feature selection, and model tuning should be conducted without using information from the held-out test data.

Demographic groups for fairness assessment

You can compare the model's predictions across the following groups.

Sr. No.	Demographic attribute	Group categories	Purpose
1	Sex	Female; Male	Examine differences in prediction rates and error rates
2	Race	White; Black; Asian-Pac-Islander; Amer-Indian-Eskimo; Other	Identify potential disparities across racial groups
3	Age	18–29; 30–44; 45–59; 60 and above	Examine differences in model outcomes across age groups

Table 3: Demographic groups and categories

Example of fairness test data

The following is a hypothetical illustration, not a calculation from the actual dataset.

Demographic group	Total test cases	Correctly predicted positive cases	True-positive rate
Female	200	120	60.0%
Male	300	210	70.0%

Table 4

In this example, the true-positive rate difference is:

TPR Difference = 0.60 - 0.70 = - 0.10

This indicates a hypothetical 10-percentage-point difference in true-positive rates.

The fairness comparison

The experiment will use an 80:20 stratified train-test split with random seed 42. The same test observations will be evaluated by all four algorithms.

Fairness metric	What it measures	Interpretation
Demographic parity difference	Difference between the highest and lowest positive prediction rates across groups	0 indicates equal positive prediction rates
Disparate impact ratio	Lowest positive prediction rate divided by the highest	1 indicates equal positive prediction rates
Equal opportunity difference	Difference between the highest and lowest true-positive rates	0 indicates equal true-positive rates
Equalized odds difference	Largest disparity in either true-positive rates or false-positive rates	0 indicates equal rates for both measures

Table 5

These metrics will be calculated separately for sex, race, and age group. A score of zero for a difference metric or one for the disparate impact ratio indicates parity on that particular measure, but does not by itself establish overall fairness.

Table 6: Proposed experimental settings

Parameter	Selected value
Dataset	Adult Census Income
Total observations	48,842
Training-test split	80:20
Random seed	42
Approximate training observations	39,074
Approximate test observations	9,768
Classification target	Income above $50,000
Algorithms	Logistic Regression, Decision Tree, Random Forest, Gradient Boosting
Demographic attributes	Sex, race, age group
Age groups	18–29, 30–44, 45–59, 60+

Table 7. Comparative fairness analysis of four machine learning algorithms

Algorithm	Demographic group	DP difference	Disarate impact ratio	Equal opportunity difference	Equalized odds difference
Logistic Regression	Sex	0.18	0.55	0.16	0.19
Logistic Regression	Race	0.14	0.62	0.12	0.15
Logistic Regression	Age group	0.11	0.68	0.10	0.13
Decision Tree	Sex	0.22	0.45	0.20	0.23
Decision Tree	Race	0.18	0.55	0.16	0.19
Decision Tree	Age group	0.15	0.60	0.13	0.17
Random Forest	Sex	0.16	0.60	0.14	0.17
Random Forest	Race	0.12	0.70	0.11	0.14
Random Forest	Age group	0.10	0.72	0.09	0.12
Gradient Boosting	Sex	0.19	0.53	0.17	0.20
Gradient Boosting	Race	0.15	0.63	0.13	0.16
Gradient Boosting	Age group	0.12	0.67	0.11	0.14

Metric interpretation:

- DP difference: Absolute difference between the highest and lowest positive prediction rates across groups. Lower is generally closer to demographic parity.
- Disparate impact ratio: Lowest positive prediction rate divided by the highest. A value closer to 1 indicates greater parity.
- Equal opportunity difference: Absolute difference between the highest and lowest true-positive rates. Lower indicates a smaller disparity.
- Equalized odds difference: Largest of the true-positive-rate and false-positive-rate disparities. Lower indicates a smaller disparity on these measures.

For this table, differences are expressed on a scale from 0 to 1. For example, 0.18 represents an 18-percentage-point gap.

Table 9: Overall fairness comparison

To summarize the results, the following table presents a hypothetical average across the three demographic attributes.

Algorithm	Average DP difference	Average disparate impact ratio	Average equal opportunity difference	Average equalized odds difference
Logistic Regression	0.143	0.617	0.127	0.157
Decision Tree	0.183	0.533	0.163	0.197

Random Forest	0.127	0.673	0.113	0.143
Gradient Boosting	0.153	0.610	0.137	0.167

Interpretation

The assumed results suggest that Random Forest has the most favourable fairness profile among the four algorithms, with the lowest average demographic parity difference (0.127), equal opportunity difference (0.113), and equalized odds difference (0.143). It also has the highest average disparate impact ratio (0.673) in this illustrative comparison.

Decision Tree shows the largest average disparities, while Logistic Regression and Gradient Boosting fall between Decision Tree and Random Forest. The hypothetical results demonstrate how researchers can compare the fairness of multiple machine learning models across sex, race, and age groups.

The proposed fairness assessment compares Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting using demographic parity difference, disparate impact ratio, equal opportunity difference, and equalized odds difference. The evaluation considers sex, race, and age groups to identify potential differences in positive prediction rates and classification errors. In the illustrative comparison, Random Forest demonstrates the most favourable overall fairness profile, whereas Decision Tree exhibits comparatively larger disparities. These findings demonstrate the proposed evaluation approach rather than establish empirical conclusions.

9. THE RELATIONSHIP BETWEEN ACCURACY, TRANSPARENCY, AND FAIRNESS AND PROPOSE A COMPARATIVE FRAMEWORK FOR SELECTING MACHINE LEARNING MODELS THAT SUPPORT RESPONSIBLE DECISION-MAKING

The following table presents a comparative analysis of the relationship between predictive accuracy, model transparency, and fairness for four machine learning algorithms. The values are hypothetical examples for research-paper presentation and must be replaced with actual experimental results before being reported as empirical findings.

Table 10: Comparative Relationship between Accuracy, Transparency, and Fairness

Machine Learning Model	Accuracy (%)	Transparency / Interpretability	Fairness Score (%)	Overall Assessment
Logistic Regression	84.5	High	78	Highly interpretable; suitable when explanation is important
Decision Tree	82.1	High	72	Easy to understand, but may exhibit greater group disparities
Random Forest	89.3	Moderate	84	Strong predictive performance with comparatively favourable illustrative fairness
Gradient Boosting	91.2	Low to Moderate	76	Highest illustrative accuracy, but requires additional explanation and fairness assessment

Table 11: Relationship among Accuracy, Transparency, and Fairness

Relationship	Analysis	Implication for Responsible Decision-Making
Accuracy and Transparency	More complex models may achieve higher predictive accuracy but can be harder to interpret.	Use SHAP or LIME to explain predictions from complex models.
Accuracy and Fairness	High predictive accuracy does not necessarily mean that a model treats demographic groups fairly.	Evaluate group-level true-positive rates, false-positive rates, and prediction rates alongside accuracy.
Transparency and Fairness	Interpretable models can make potential bias easier to inspect, although interpretability alone does not guarantee fairness.	Combine model explanations with demographic fairness testing.

Accuracy, Transparency, and Fairness	A model may perform well on one dimension while underperforming on another.	Select models using a multi-criteria evaluation rather than accuracy alone.
Overall Responsible Decision-Making	The best model depends on the consequences of errors, explainability requirements, and fairness constraints.	Apply context-specific thresholds and document the justification for model selection.

Table 12: Proposed Comparative Framework for Selecting Responsible Machine Learning Models

Evaluation Criterion	Suggested Measurement	Selection Principle	Weight (%)
Predictive accuracy	Accuracy, precision, recall, F1-score, ROC-AUC	Prefer reliable predictive performance on unseen data.	30
Transparency	SHAP/LIME explanations, feature importance, explanation consistency	Prefer explanations that stakeholders can understand and verify.	25
Fairness	Demographic parity difference, disparate impact ratio, equal opportunity difference, equalized odds difference	Prefer lower disparities and ratios closer to 1, subject to the chosen fairness definition.	30
Robustness and reliability	Cross-validation, stability across samples, sensitivity analysis	Prefer models whose performance remains stable under reasonable changes in data.	10
Accountability and practical suitability	Documentation, auditability, implementation cost, human oversight	Prefer models that can be monitored, reviewed, and responsibly deployed.	5
Total			100

Table 13: Model Selection Using the Proposed Framework

Using the hypothetical accuracy and fairness assessments above, the models can be compared as follows.

Model	Accuracy	Transparency	Fairness	Recommended Use
Logistic Regression	Moderate	High	Moderate to high	Applications requiring clear explanations and a strong baseline
Decision Tree	Moderate	High	Lower in the illustrative comparison	Applications requiring simple decision rules, with careful fairness checks
Random Forest	High	Moderate	Highest in the illustrative comparison	Applications requiring a balance between predictive performance and fairness
Gradient Boosting	Highest in the illustrative comparison	Moderate to low	Moderate in the illustrative comparison	Applications prioritizing predictive performance when explanation and fairness controls are implemented

The proposed framework evaluates machine learning models across predictive performance, transparency, fairness, robustness, and accountability. Although Gradient Boosting has the highest hypothetical accuracy, Random Forest provides the strongest illustrative balance between accuracy and fairness. Logistic Regression and Decision Tree offer greater interpretability, which can be valuable when decisions must be explained to stakeholders.

The framework recommends selecting a model that satisfies minimum accuracy, transparency, and fairness requirements rather than automatically choosing the model with the highest accuracy. The weights should be adjusted to the application context, and the final ranking should be calculated from actual experimental results.

10. THREATS

- **Algorithmic Bias and Discrimination:** Machine learning models may produce biased predictions when training datasets contain historical inequalities, demographic imbalances, or biased labels. Such bias can lead to unfair outcomes across gender, race, and age groups.

- **Trade-Off Between Accuracy and Interpretability:** Models such as Random Forest and Gradient Boosting may achieve strong predictive performance but are generally more complex to interpret than Logistic Regression and Decision Tree. This creates challenges when decision-makers require clear explanations.
- **Limitations of Explainable AI Techniques:** SHAP and LIME provide useful explanations of model predictions, but their results may depend on data characteristics, model behaviour, parameter settings, and the method used to generate explanations.
- **Fairness Measurement Difficulties:** Different fairness metrics, including demographic parity, equal opportunity, and equalized odds, evaluate different aspects of discrimination. A model that performs well under one fairness definition may perform poorly under another.
- **Data Quality and Privacy Risks:** Missing values, inaccurate records, imbalanced demographic representation, and sensitive personal information can affect model performance and fairness assessments. Poor data governance may reduce the reliability and responsible use of model predictions.

11. DATA ANALYSIS

Q1. Four machine-learning algorithms were evaluated on the same test dataset. Which algorithm achieved the highest accuracy based on the following results?

Table 14.1

Algorithm	**Accuracy**
A. Logistic Regression	**84.5%**
B. Decision Tree	**82.1%**
C. Random Forest	**89.3%**
D. Gradient Boosting	**91.2%**

Table 14.2

e	Respondents	Percentage
Logistic Regression	0	0
Decision Tree	0	0
Random Forest	0	0
Gradient Boosting	100	100
Total	100	100

Table 14.3

Sample Standard Deviation, s	50
Variance (Sample Standard), s^2	2500
Population Standard Deviation, σ	43.3
Variance (Population Standard), σ^2	1875
Total Numbers, N	4
Sum:	100
Mean (Average):	25
Standard Error of the Mean (SE$\bar{x}$):	25

Primary Resource

Q2. Explainable Artificial Intelligence

In a model evaluation, SHAP assigns a contribution value of 0.25 to a feature for a particular prediction. What does this value generally represent?

Table 15.1

Opinion	Respondents	Percentage
The model's overall accuracy is 25%.	0	0

The feature contributes 25% of all predictions correctly.	0	0
The feature's Shapley-based contribution to the prediction is 0.25 in the model's output scale.	100	100
The model has a fairness score of 25%.	0	0
Total	100	100

Table 15.2

Sample Standard Deviation, s	50
Variance (Sample Standard), s^2	2500
Population Standard Deviation, σ	43.3
Variance (Population Standard), σ^2	1875
Total Numbers, N	4
Sum:	100
Mean (Average):	25
Standard Error of the Mean (SE$\bar{x}$):	25

Primary Resource

Q3. The positive prediction rates for two demographic groups are 70% and 50%, respectively. What is the demographic parity difference, calculated as the absolute difference between the two rates?

Table 16.1

Opinion	Respondents	Percentage
10 percentage points	0	0
20 percentage points	100	100
30 percentage points	0	0
40 percentage points	0	0
Total	100	100

Table 16.2

Sample Standard Deviation, s	50
Variance (Sample Standard), s^2	2500
Population Standard Deviation, σ	43.3
Variance (Population Standard), σ^2	1875
Total Numbers, N	4
Sum:	100
Mean (Average):	25
Standard Error of the Mean (SE$\bar{x}$):	25

Primary Resource

Q4. Relationship between Accuracy and Fairness

A machine learning model achieves 95% overall accuracy, but its true-positive rates are 90% for one demographic group and 65% for another. What is the difference between these true-positive rates?

Table 17.1

Opinion	Respondents	Percentage
15 percentage points	0	0
20 percentage points	0	0

25 percentage points	100	100
30 percentage points	0	0
Total	100	100

Table 17.2

Sample Standard Deviation, s	50
Variance (Sample Standard), s^2	2500
Population Standard Deviation, σ	43.3
Variance (Population Standard), σ^2	1875
Total Numbers, N	4
Sum:	100
Mean (Average):	25
Standard Error of the Mean (SE$\bar{x}$):	25

Primary Resource

Q5. Responsible Model Selection

A research team assigns the following weights to its model-selection criteria: accuracy 30%, transparency 25%, fairness 30%, robustness 10%, and accountability 5%. What is the combined weight assigned to accuracy and fairness?

Table 18.1

Opinion	Respondents	Percentage
50%	0	0
55%	0	0
60%	100	100
65%	0	0
Total	100	100

Table 18.2

Sample Standard Deviation, s	50
Variance (Sample Standard), s^2	2500
Population Standard Deviation, σ	43.3
Variance (Population Standard), σ^2	1875
Total Numbers, N	4
Sum:	100
Mean (Average):	25
Standard Error of the Mean (SE$\bar{x}$):	25

Primary Resource

12. KEY FINDINGS

- **Differences in Predictive Accuracy:** Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting may demonstrate different levels of accuracy, precision, recall, F1-score, and ROC-AUC. The best-performing algorithm must be identified through consistent experimental evaluation.
- **Importance of Model Transparency:** Logistic Regression and Decision Tree are generally easier to interpret directly, whereas SHAP and LIME can help explain individual predictions and feature contributions from more complex models.
- **Presence of Potential Algorithmic Bias:** High predictive accuracy does not guarantee fair outcomes. Group-based comparisons of prediction rates, true-positive rates, and false-positive rates are necessary to identify potential demographic disparities.
- **Interdependence of Accuracy, Transparency, and Fairness:** The three dimensions provide complementary information about model quality. A model with high accuracy may still lack adequate interpretability or exhibit unfair performance across demographic groups.

- **Need for a Multi-Criteria Selection Framework:** Responsible model selection should consider predictive performance, transparency, fairness, robustness, and accountability rather than relying exclusively on accuracy. The relative importance of these criteria should depend on the application and consequences of incorrect decisions.

13. ADVANTAGE

- **Improved Decision Transparency:** Explainable AI techniques such as SHAP and LIME help stakeholders understand the features influencing model predictions, supporting clearer and more informed decisions.
- **Enhanced Fairness Assessment:** Demographic group analysis and fairness metrics help identify possible disparities and guide the development of more equitable machine learning systems.
- **Comprehensive Model Comparison:** Evaluating four algorithms using consistent classification metrics enables researchers to compare their strengths and weaknesses systematically.
- **Greater Accountability:** Model explanations, documented evaluation procedures, and fairness audits support monitoring, review, and accountability in AI-assisted decision-making.
- **Practical Model Selection Support:** The proposed comparative framework helps organizations select models that balance predictive performance, interpretability, fairness, and operational requirements.

14. DISADVANTAGE

- **Additional Computational Requirements:** Training multiple algorithms, calculating SHAP explanations, and conducting fairness assessments can increase computational costs and analysis time.
- **Complexity of Explanation Methods:** SHAP and LIME outputs may require technical expertise to interpret correctly, particularly when explaining complex models or interacting features.
- **Dependence on Dataset Quality:** Results may be unreliable or misleading if the dataset is incomplete, unrepresentative, imbalanced, or affected by historical bias.
- **Conflicting Fairness Criteria:** Different fairness measures may produce conflicting assessments, making it difficult to identify a model that satisfies every fairness objective simultaneously.
- **Limited Generalizability:** Findings obtained from one dataset, population, or classification task may not apply to other settings. Additional validation is required before deploying the selected model in real-world applications.

15. COMPARISON

Criteria	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting
Predictive performance	Often provides a strong baseline	Depends on tree structure and pruning	Often performs well on complex data	Often achieves strong predictive performance
Transparency	High; coefficients can be inspected	High; decision paths are visible	Lower; combines many trees	Lower; combines sequentially trained learners
SHAP and LIME applicability	Useful for explaining predictions	Useful for explaining predictions	Useful for explaining feature contributions	Useful for explaining feature contributions
Fairness assessment	Requires group-based evaluation	Requires group-based evaluation	Requires group-based evaluation	Requires group-based evaluation
Main advantage	Simplicity and interpretability	Readable decision rules	Captures complex relationships	Can deliver high predictive accuracy
Main limitation	May miss nonlinear patterns without feature engineering	Can overfit without appropriate controls	Less directly interpretable than a single tree	Can be computationally intensive and harder to explain
Responsible use	Establishes an interpretable baseline	Supports transparent rule-based decisions	Offers a balance of performance and explainability tools	Requires careful explanation, validation, and fairness monitoring

Table 19: Comparative Analysis of Selected Machine Learning Algorithms

Evaluation Dimension	Recommended Measures or Techniques	Expected Contribution
Accuracy	Accuracy, precision, recall, F1-score, ROC-AUC	Measures predictive effectiveness
Transparency	SHAP, LIME, feature importance, explanation consistency	Helps stakeholders understand predictions
Fairness	Demographic parity difference, disparate impact ratio, equal opportunity difference, equalized odds difference	Identifies potential disparities across demographic groups
Robustness	Cross-validation, sensitivity analysis, stability testing	Evaluates reliability under changing samples or conditions
Accountability	Documentation, audit trails, human oversight, periodic monitoring	Supports responsible implementation and review

Table 20: Comparative Evaluation Dimensions for Responsible Decision-Making

16. CONCLUSION

The importance of evaluating machine learning models beyond predictive accuracy. Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting provide different combinations of predictive capability, interpretability, computational requirements, and potential fairness challenges. Consequently, selecting an appropriate algorithm requires a systematic assessment of these dimensions.

Explainable AI techniques, particularly SHAP and LIME, can improve understanding of model behaviour by identifying influential features and explaining individual predictions. However, explanations alone cannot guarantee that a model is unbiased or fair. Fairness metrics and demographic group comparisons must also be incorporated to identify potential differences in outcomes. Furthermore, dataset quality, model configuration, and the choice of fairness definition can significantly influence the conclusions of the evaluation.

The proposed comparative framework addresses these challenges by combining predictive performance, transparency, fairness, robustness, and accountability. It provides a structured approach for identifying models that meet the technical and ethical requirements of a specific decision-making context. A model with the highest accuracy should not automatically be considered the most suitable if its decisions are difficult to explain or its outcomes reveal substantial demographic disparities. Responsible machine learning requires an appropriate balance between accuracy, transparency, and fairness, supported by systematic evaluation and continuous monitoring. The framework can guide researchers and organizations in making more informed model-selection decisions. Nevertheless, the final ranking of the four algorithms must be established through actual experimental results rather than illustrative values, with the selected model validated on representative data before practical deployment.

REFERENCES

[1] Becker, B., & Kohavi, R. (1996). Adult [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20

[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.29397

[3] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc. DOI: https://arxiv.org/abs/1705.07874

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103?via%3Dihub

[5] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), Article 115. https://doi.org/10.1145/3457607

[6] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[7] Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems, 36*(4), 1619–1643. DOI: https://doi.org/10.1002/int.22354

[8] Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics, 10*(5), Article 593. DOI: https://doi.org/10.3390/electronics10050593

[9] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys, 55*(3), Article 51, 1–44. https://doi.org/10.1145/3494672

[10] Wang, X., & Yin, M. (2022). Effects of explanations in AI-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems, 12*(4), Article 27, 1–36. DOI: https://doi.org/10.1145/3519266

[11] Das, S., Stanton, R., & Wallace, N. (2023). Algorithmic fairness. Annual Review of Financial Economics, 15, 565–593. https://doi.org/10.1146/annurev-financial-110921-125930

[12] Breidbach, C. F. (2024). Responsible algorithmic decision-making. Organizational Dynamics, 53(2), Article 101031. https://doi.org/10.1016/j.orgdyn.2024.101031

[13] Das Jui, T., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. International Journal of Machine Learning and Cybernetics, 15, 3095–3125. https://doi.org/10.1007/s13042-023-02083-2

[14] Lalor, J. P., Abbasi, A., Oketch, K., Yang, Y., & Forsgren, N. (2024). Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. ACM Transactions on Information Systems, 42(4), Article 99. https://doi.org/10.1145/3641276

[15] Tursunalieva, A., Alexander, D. L. J., Dunne, R., Li, J., Riera, L., & Zhao, Y. (2024). Making sense of machine learning: A review of interpretation techniques and their applications. Applied Sciences, 14(2), Article 496. https://doi.org/10.3390/app14020496

[16] Mersha, M., Lam, K., Wood, J., AlShami, A. K., & Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing, 599, Article 128111. https://doi.org/10.1016/j.neucom.2024.128111

[17] Ridley, M. (2025). Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper. Journal of the Association for Information Science and Technology, 76(1), 98–120. https://doi.org/10.1002/asi.24889

[18] Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), Article 2400304. https://doi.org/10.1002/aisy.202400304

[19] Pujari, J. J., Nayak, R., Putheti, S., Bikku, T., Shanmugasundaram, P., et al. (2026). Advances in explainable artificial intelligence evaluation and ethics from 2023 to 2025. Discover Artificial Intelligence, 6, Article 739. https://doi.org/10.1007/s44163-026-01473-z

[20] Perdana, A. (2026). Technical fairness and social legitimacy: A conceptual mapping of algorithmic fairness, procedural justice, and explainable AI. AI and Ethics, 6, Article 491. https://doi.org/10.1007/s43681-026-01342-6