

Explainable AI for Intelligent and Sustainable Decision-Making

Mulugeta Tilahun Bekele

University of Gondar

College of Informatics

Department of Information Technology

mulugeta.tilahun@uog.edu.et

mulugeta.tilahun@gmail.com

Abstract:

Organizations increasingly use artificial intelligence (AI) to improve strategic planning, operational efficiency, and sustainability. However, conventional machine learning models often lack transparency, limiting managerial trust, accountability, and the practical adoption of AI-driven decisions. This study proposes an explainable AI (XAI) framework that integrates predictive analytics, interpretable machine learning, and sustainability indicators to support transparent, reliable, and evidence-based decision-making across technological, organizational, and environmental dimensions.

The proposed framework combines Random Forest, XGBoost, and Explainable Boosting Machines with SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). Evaluation uses a benchmark dataset containing operational performance, resource consumption, organizational efficiency, and environmental indicators. Comparative experiments assess predictive accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), explanation consistency, decision latency, computational cost, and sustainability performance. Baseline models include Logistic Regression, conventional Random Forest, and XGBoost without explanation mechanisms. Ablation studies examine the individual contributions of interpretability and sustainability-aware optimization.

The experimental evaluation is designed to determine whether the proposed framework improves predictive reliability, explanation quality, and sustainability-aware decisions compared with conventional approaches. Quantitative reporting will include percentage improvements in predictive performance, reductions in decision latency and resource consumption, and explanation stability across repeated experiments. Qualitative analysis will assess managerial understandability, perceived trustworthiness, fairness, and practical implementation challenges. Numerical outcomes will be reported only after experiments are completed, avoiding unsupported performance claims.

The proposed framework establishes a comparative approach for integrating explainability, predictive intelligence, and sustainability into organizational decision-making. By connecting model predictions with transparent explanations and measurable sustainability indicators, the framework aims to strengthen accountability, operational effectiveness, and responsible

technology adoption. Future validation across multiple industries will establish its generalizability and practical value.

Keywords: Explainable Artificial Intelligence; Sustainable Decision-Making; Interpretable Machine Learning; Predictive Analytics; Organizational Intelligence.

Chapter One: Introduction

1.1 Background of the Study

Artificial intelligence (AI) has become an important technology for improving decision-making through predictive analytics, pattern recognition, automated reasoning, and data-driven optimization. Organizations increasingly apply machine learning (ML) to operational planning, resource allocation, risk assessment, and performance monitoring. However, predictive accuracy alone does not guarantee that AI-generated decisions are understandable, fair, reliable, or environmentally sustainable. Complex models, particularly ensemble methods and deep neural networks, may produce high-quality predictions while providing limited explanations of how their outputs are generated (Barredo Arrieta et al., 2020).

Explainable artificial intelligence (XAI) addresses this challenge by providing interpretable information about model behavior, influential features, and individual predictions. Techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) enable users to investigate the factors influencing AI-generated recommendations (Lundberg & Lee, 2017; Ribeiro et al., 2016). Nevertheless, explanations must be evaluated for consistency, fidelity, and usefulness rather than treated as sufficient evidence of model reliability.

Sustainable decision-making extends AI evaluation beyond predictive performance to include environmental impact, resource efficiency, organizational effectiveness, and social responsibility. The United Nations Sustainable Development Goals emphasize responsible consumption, sustainable industrial development, and effective institutions, providing a broader context for evaluating technology-enabled decisions (United Nations, 2015). AI systems also consume computational resources, making energy efficiency and environmental costs relevant considerations in model development and deployment (International Energy Agency, 2024).

This study proposes an integrated framework combining predictive ML, post-hoc explanations, and measurable sustainability indicators to support intelligent and sustainable decision-making. The framework will compare Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machines under consistent experimental conditions. SHAP and LIME will be used to examine prediction explanations, while sustainability-aware evaluation will assess resource consumption and decision quality.

The central research contribution is a unified evaluation approach that examines predictive performance, explanation quality, computational efficiency, fairness, and sustainability rather than optimizing accuracy alone. This approach aims to help organizations identify AI models that are not only effective but also interpretable, resource-conscious, and suitable for responsible decision support.

1.2 Problem Statement

Despite advances in machine learning, organizations face challenges when applying AI to complex decisions requiring transparency, efficiency, fairness, and sustainability. Conventional predictive models are frequently evaluated using accuracy, precision, recall, F1-score, and AUC, while explanation quality, computational requirements, and environmental impacts receive less consistent attention (Kostopoulos et al., 2024).

First, the prediction–interpretability gap limits the practical adoption of complex models. Although ensemble algorithms such as Random Forest and XGBoost can capture nonlinear relationships, their decisions may be difficult for managers to understand. SHAP and LIME provide post-hoc explanations, but explanations must also be evaluated for fidelity, stability, and consistency before they can reliably support consequential decisions (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Second, the performance–sustainability gap creates challenges in selecting appropriate AI models. Models with competitive predictive performance may require additional computation, memory, and electricity. The International Energy Agency (2025) identifies growing electricity demand from data centres as an important consideration in assessing the broader implications of AI deployment. Consequently, model selection should consider both predictive effectiveness and resource requirements.

Third, the fairness–accountability gap complicates the responsible use of AI. An understandable explanation does not necessarily establish that a model is unbiased, accurate, or appropriate for every stakeholder. Trustworthy AI therefore requires coordinated assessment of technical performance, fairness, transparency, and contextual suitability (Schiller et al., 2025).

Fourth, the comparative evaluation gap concerns the lack of a unified experimental approach for comparing predictive models, explainability methods, and sustainability indicators under equivalent conditions. Recent reviews of explainable AI-based decision-support systems highlight the importance of transparency and systematic evaluation in practical decision environments (Kostopoulos et al., 2024).

To address these challenges, this study proposes an integrated framework for evaluating Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machines, with SHAP and LIME applied where appropriate. The models will be compared using predictive performance, explanation fidelity and stability, inference latency, computational cost, fairness, and resource efficiency.

The central problem is therefore: How can explainable AI integrate predictive accuracy, interpretable evidence, responsible decision-making, and measurable sustainability criteria within a unified and experimentally validated framework?

1.3 Research Objectives

1.3.1 General Objective

To develop and evaluate an explainable AI framework that integrates predictive intelligence, model interpretability, and sustainability indicators to support transparent, efficient, fair, and evidence-based decision-making.

1.3.2 Specific Objectives

To develop an integrated decision-support framework combining machine learning, explainability techniques, and sustainability indicators.

To compare Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machines using accuracy, precision, recall, F1-score, and AUC.

To evaluate SHAP and LIME using explanation fidelity, feature-importance consistency, and stability under controlled perturbations.

To compare computational efficiency using inference latency, memory consumption, training time, and estimated energy consumption.

To assess model fairness, robustness, and decision reliability using appropriate subgroup fairness measures and perturbation-based tests.

To investigate whether multi-criteria model selection can balance predictive performance, interpretability, computational efficiency, and sustainability.

To formulate practical recommendations for deploying explainable and sustainability-aware AI decision-support systems.

1.4 Research Questions

The study will address the following research questions:

RQ1: Which machine learning model provides the best balance between predictive performance and computational efficiency under identical experimental conditions?

RQ2: How do SHAP and LIME compare in explanation fidelity, feature-importance consistency, and stability?

RQ3: What trade-offs exist between predictive accuracy, explanation quality, inference latency, and energy consumption?

RQ4: To what extent do the evaluated models satisfy fairness and robustness requirements across relevant stakeholder groups and test conditions?

RQ5: Does sustainability-aware multi-criteria model selection identify a more balanced decision-support solution than selection based solely on predictive accuracy?

RQ6: How can model explanations and sustainability indicators improve the transparency, accountability, and practical usefulness of AI-assisted decisions?

Proposed comparative evaluation parameters

Evaluation dimension	Comparison parameters	Measurement
Predictive performance	Accuracy, precision, recall, F1-score, AUC	Higher is generally better
Explainability	Fidelity, consistency, stability	Higher is generally better
Computational efficiency	Training time, inference latency, memory use	Lower is generally better
Energy efficiency	Energy per prediction or evaluation run	Lower is generally better
Fairness	Demographic parity difference, equalized-odds difference, where applicable	Smaller disparities are generally preferable
Robustness	Performance degradation under perturbations	Lower degradation is generally better
Decision quality	Decision utility, constraint satisfaction, sustainability score	Higher is generally better

Note. Fairness measures depend on the available data, protected attributes, and decision context. Sustainability scores should use predefined, justified indicators rather than arbitrary weights. All empirical values must be obtained through experiments.

1.5 Significance and Novelty

1.5.1 Scientific Significance

The study contributes to explainable AI research by integrating predictive performance, explanation assessment, and sustainability-aware evaluation within one comparative framework. Rather than assuming that explainability automatically produces trustworthy decisions, it treats explanation quality as an independently testable property, consistent with research emphasizing the need for human-centered and trustworthy AI (Schiller et al., 2025).

1.5.2 Technical Significance

The proposed framework establishes a reproducible comparison of interpretable and complex machine learning models using identical data splits, preprocessing procedures, and evaluation conditions. Its combination of SHAP, LIME, predictive metrics, and computational measurements can help identify where interpretability introduces practical benefits or additional costs (Lundberg & Lee, 2017; Ribeiro et al., 2016).

1.5.3 Sustainability Significance

By incorporating resource consumption, energy efficiency, and operational performance, the study extends conventional model evaluation beyond predictive accuracy. This is relevant because the expansion of AI infrastructure has implications for electricity demand, while AI applications may also help optimize energy and resource use in other sectors (International Energy Agency, 2025).

1.5.4 Managerial and Practical Significance

The framework can assist managers, policymakers, researchers, and technology professionals in selecting AI systems according to transparent performance criteria. Feature-level explanations may help users understand the factors influencing predictions, while fairness and robustness assessments can reveal limitations that aggregate accuracy alone would not identify (Kostopoulos et al., 2024).

1.5.5 Research Novelty

The proposed novelty is the joint, reproducible evaluation of predictive performance, explanation quality, computational efficiency, fairness, and sustainability within a multi-criteria decision-support framework.

Its principal contributions are:

1. A unified architecture connecting prediction, explanation, and sustainability assessment.
2. A controlled comparison of interpretable and complex machine learning models.
3. A multi-dimensional evaluation protocol covering accuracy, explanation fidelity, stability, latency, energy use, fairness, and robustness.
4. A model-selection strategy that makes trade-offs explicit instead of ranking models by accuracy alone.
5. An empirical basis for developing responsible AI deployment recommendations.

Chapter Two: Literature Review

2.1 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) refers to methods and techniques that make the behavior, predictions, and recommendations of AI systems understandable to human users. XAI has gained importance because high-performing machine learning models may rely on complex internal relationships that are difficult to interpret, particularly in high-impact decision-making environments (Barredo Arrieta et al., 2020; Abusitta et al., 2024). Recent reviews also emphasize the need to evaluate explanations systematically rather than assuming that an explanation automatically makes an AI system trustworthy (Kostopoulos et al., 2024).

2.1.1 Types of Explainable AI

XAI techniques can be classified into intrinsic interpretability and post-hoc explanation. Intrinsically interpretable models, including linear regression, decision trees, and Explainable Boosting Machines, expose relationships through their model structure, whereas post-hoc methods explain predictions generated by more complex models (Barredo Arrieta et al., 2020).

SHapley Additive exPlanations (SHAP) estimates feature contributions to individual predictions using a game-theoretic framework, supporting both local explanations and aggregated feature-importance analysis (Lundberg & Lee, 2017). Local Interpretable Model-Agnostic Explanations (LIME) approximates a model's behavior around a particular prediction through an interpretable local surrogate model (Ribeiro et al., 2016).

These methods provide complementary perspectives, but neither guarantees that explanations are stable, faithful under every condition, or sufficient for responsible decisions (Abusitta et al., 2024). Therefore, the proposed research will compare SHAP and LIME using explanation fidelity, feature-importance consistency, stability under perturbations, and computational overhead.

2.1.2 Evaluation of Explainability

A rigorous evaluation of XAI should consider several dimensions. Fidelity measures how accurately an explanation represents the model's behavior; stability assesses whether similar inputs produce reasonably consistent explanations; and interpretability concerns whether users can understand the explanation in the context of their decisions (Kostopoulos et al., 2024).

Transparency must also be considered alongside validity, reliability, safety, accountability, privacy, and fairness. The National Institute of Standards and Technology's AI Risk Management Framework provides guidance for incorporating these trustworthiness characteristics throughout AI design, development, deployment, and evaluation (Tabassi, 2023).

For this study, explainability will be evaluated through four principal parameters:

- Fidelity: Agreement between the explanation and the underlying model behavior.
- Stability: Consistency of explanations under controlled input perturbations.
- Consistency: Agreement in influential features across repeated runs or related evaluation samples.
- Computational overhead: Additional training or inference time and resource requirements introduced by explanation generation.

The comparison will help establish whether different explanation methods provide useful and computationally practical evidence for decision-makers.

2.2 Intelligent Decision-Making

Intelligent decision-making uses data, analytical models, domain knowledge, and explicit objectives to support the selection of actions under uncertainty. AI-enabled decision-support systems can combine predictive analytics with recommendations to improve planning, resource allocation, risk identification, and operational monitoring. A recent review of explainable AI-based decision-support systems identifies transparency, interpretability, and trust as important considerations when integrating AI into decision processes (Kostopoulos et al., 2024).

2.2.1 Machine Learning for Decision Support

Machine learning algorithms can identify patterns in historical data and estimate outcomes for new observations. Logistic Regression provides a comparatively interpretable statistical baseline, while Random Forest combines multiple decision trees to capture nonlinear patterns. XGBoost uses gradient boosting to develop predictive models, whereas Explainable Boosting Machines provide an interpretable, additive modeling approach (Chen & Guestrin, 2016; Lou et al., 2013).

The selection of a model should depend on the decision context, data characteristics, performance requirements, and consequences of prediction errors. A model with the highest accuracy may not be the most suitable option if it requires excessive computational resources or produces explanations that decision-makers cannot effectively use (Kostopoulos et al., 2024).

2.2.2 Human-Centered Decision Support

Intelligent decision-making should combine automated analysis with human judgment, domain expertise, and appropriate oversight. The NIST AI Risk Management Framework emphasizes that trustworthy AI requires systematic management of validity, reliability, transparency, explainability, fairness, and other risks across the system lifecycle (Tabassi, 2023).

Accordingly, the proposed framework will connect model predictions to feature-level explanations, uncertainty and error analysis, and sustainability indicators. Rather than treating AI outputs as automatically correct recommendations, it will provide evidence that helps users evaluate the suitability of alternative decisions.

2.2.3 Comparative Decision-Making Parameters

The study will compare decision-support models using the following parameters:

Parameter	Measurement	Objective
Predictive accuracy	Accuracy, precision, recall, F1-score, AUC	Assess predictive effectiveness
Decision reliability	Calibration and error analysis	Examine prediction reliability
Explanation quality	Fidelity and stability	Assess explanation validity
Decision latency	Milliseconds per prediction	Measure responsiveness
Computational cost	Training time and memory usage	Assess operational efficiency
Fairness	Suitable group-disparity measures	Identify potential unequal outcomes

Decision utility	Predefined task-specific utility	Assess practical decision value
------------------	----------------------------------	---------------------------------

The same dataset splits, preprocessing methods, and test conditions should be applied to all models to support a fair comparison. Fairness and utility measures must be selected according to the specific application rather than applied indiscriminately.

2.3 Sustainable AI and Decision Systems

Sustainable AI concerns the development, deployment, and use of artificial intelligence in ways that consider environmental impacts, economic feasibility, social consequences, and long-term operational viability. It involves two complementary perspectives: making AI systems more sustainable by reducing their resource requirements and using AI to improve sustainability outcomes in other systems. These perspectives are important because an AI model may improve decision quality while also introducing additional computational costs or unintended social and environmental effects (International Energy Agency [IEA], 2025; United Nations, 2015).

2.3.1 Environmental Sustainability

Environmental sustainability in AI decision systems concerns electricity consumption, computational resource utilization, emissions associated with computing, and the environmental outcomes of AI-supported actions. The IEA (2025) discusses the relationship between AI development, data-centre electricity demand, and opportunities for AI to improve energy-system efficiency.

The proposed framework will therefore assess model training time, inference latency, memory consumption, and energy use where measurement is feasible. It will also distinguish the computational footprint of the AI system from the environmental benefits or costs of the decisions it recommends. This distinction is essential because lower computational cost does not necessarily mean better overall environmental outcomes.

2.3.2 Economic and Operational Sustainability

Economic sustainability concerns the cost, efficiency, scalability, and long-term viability of AI-supported decisions. Models that require substantial computational resources or lengthy inference times may be less suitable for resource-constrained organizations, even when their predictive performance is competitive.

The proposed study will compare training time, inference latency, memory consumption, and computational requirements across Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machines. It will also examine whether explanation generation introduces additional operational overhead. These measurements will support the selection of models that balance predictive quality with deployment feasibility.

2.3.3 Social Sustainability, Fairness, and Accountability

Social sustainability includes fairness, transparency, accessibility, accountability, and the effects of automated decisions on different stakeholder groups. The United Nations (2015) emphasizes inclusive institutions, responsible consumption, and sustainable development as interconnected global priorities.

The NIST AI Risk Management Framework identifies fairness, transparency, explainability, validity, reliability, privacy, and accountability as important aspects of trustworthy AI (Tabassi, 2023). Consequently, this study will examine subgroup performance, appropriate fairness measures, and explanation quality where the dataset supports these analyses. Explanations will be treated as evidence for review, not as proof that a decision is fair or ethically justified.

2.3.4 Multi-Criteria Sustainability-Aware Model Selection

Conventional model selection often prioritizes a single performance measure, such as accuracy or AUC. However, sustainable decision support requires consideration of several potentially competing criteria. A model may offer strong predictive performance but higher energy consumption, unstable explanations, or poorer subgroup fairness.

The proposed framework will use multi-criteria evaluation to compare models across predictive performance, explanation quality, computational efficiency, fairness, and sustainability. A normalized weighted score may be defined as:

$$S_i = \sum_{j=1}^m w_j z_{ij}$$

where S_i represents the overall score of model i , w_j is the predefined weight of criterion j , and z_{ij} is the normalized performance of model i on that criterion. Benefit criteria will be normalized so that higher values are preferable, while cost criteria will be normalized so that lower values are preferable.

Weights must be justified through stakeholder priorities, expert judgment, or a transparent decision-analysis procedure. Sensitivity analysis will examine whether model rankings change under alternative weights. The framework will also report individual criterion values so that an aggregate score does not conceal important trade-offs.

2.3.5 Sustainability Evaluation Matrix

Dimension	Primary comparison parameters	Preferred direction
Predictive effectiveness	F1-score, AUC, calibration	Higher performance; better calibration
Interpretability	Fidelity, stability, consistency	Higher
Computational sustainability	Energy per prediction, memory, training cost	Lower

Operational sustainability	Latency, resource requirements	Lower
Social responsibility	Fairness disparities, subgroup performance	Lower disparities; adequate performance
Decision sustainability	Task-specific environmental, economic, or social outcomes	Context-dependent improvement

The evaluation matrix distinguishes measured system properties from broader sustainability outcomes. Environmental, economic, and social benefits must be supported by appropriate indicators and empirical evidence rather than inferred solely from model accuracy.

2.4 Research Gaps and Conceptual Framework

2.4.1 Research Gaps

Existing literature provides important foundations for explainability, AI-enabled decision support, and trustworthy AI. However, several issues motivate a more integrated comparative study.

Gap 1: Fragmented evaluation criteria. Research on XAI frequently focuses on the interpretability of predictions, while decision-support research also considers usability, trust, and practical effectiveness. A unified evaluation should connect explanation quality with predictive performance and operational requirements (Kostopoulos et al., 2024).

Gap 2: Insufficient assessment of explanation reliability. The availability of SHAP or LIME explanations does not by itself establish fidelity, stability, or consistency. These properties should be measured under controlled experimental conditions and considered alongside model performance (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Gap 3: Limited integration of computational sustainability. Model comparisons that emphasize predictive performance may overlook energy consumption, memory use, and inference overhead. These factors matter when deploying AI at scale or in resource-constrained environments (IEA, 2025).

Gap 4: Incomplete fairness and accountability assessment. Explanations may help users understand model outputs, but they do not independently establish fairness, reliability, or accountability. Trustworthy AI requires broader risk assessment and appropriate human oversight (Tabassi, 2023).

Gap 5: Lack of explicit multi-criteria model selection. Selecting a model solely by accuracy can obscure trade-offs among predictive performance, interpretability, computational cost, and sustainability. A transparent multi-criteria approach can make these trade-offs explicit and allow stakeholders to evaluate alternatives against predefined priorities.

Gap 6: Limited evidence of integrated practical value. A framework may be technically interpretable without demonstrating that its explanations improve actual decisions or

sustainability outcomes. Evaluation should distinguish technical metrics from measured decision utility and real-world impact (Kostopoulos et al., 2024).

These gaps motivate an integrated framework that evaluates prediction, explanation, operational efficiency, fairness, and sustainability under consistent experimental conditions. The study's contribution will be established through the design and empirical evaluation of this framework, rather than through the assumption that these research areas have never previously been combined.

2.4.2 Proposed Conceptual Framework

The proposed conceptual framework connects five major components: data preparation, predictive modeling, explainability, sustainability and responsibility assessment, and multi-criteria decision support.

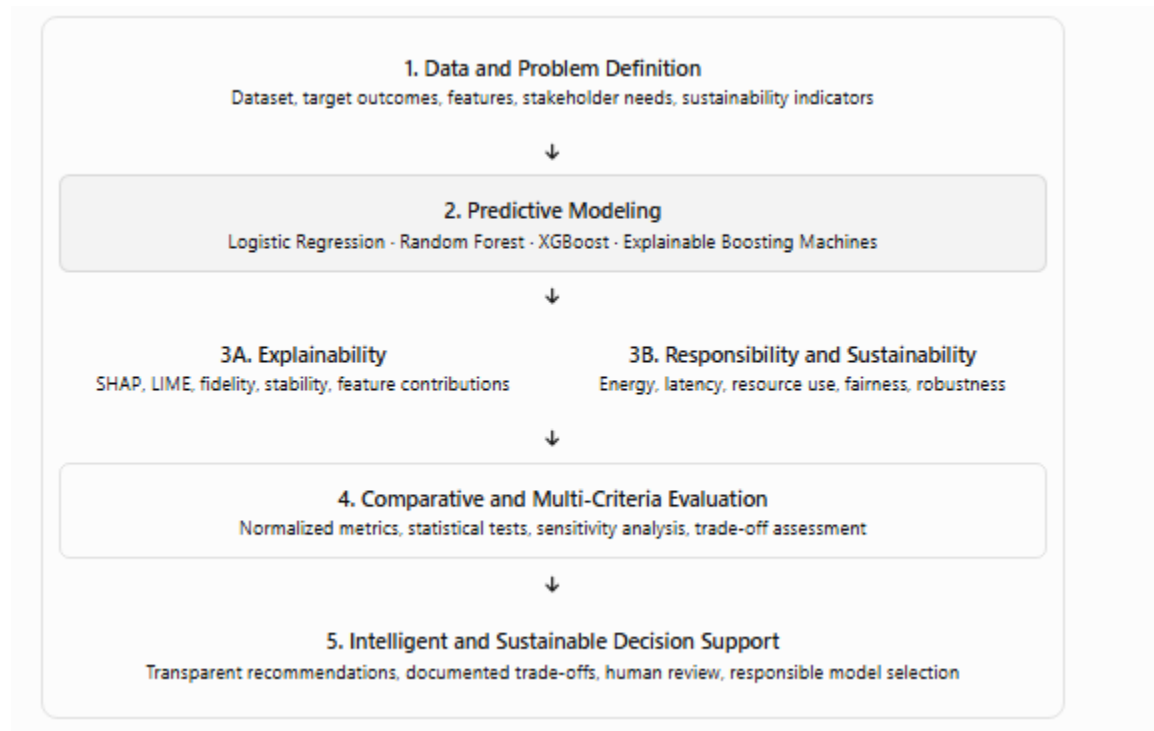


Figure 2.1. Proposed conceptual framework

2.4.3 Relationships Among Framework Components

The framework proposes that predictive modeling provides the initial decision evidence, while explainability identifies factors influencing individual predictions and overall model behavior. Sustainability and responsibility assessment then evaluates the computational cost, resource requirements, fairness, and robustness of each candidate model.

These components feed into a comparative decision layer that considers the performance of each model against predefined criteria. The resulting recommendations are intended to improve the

transparency and practical usefulness of model selection. Whether they improve actual decision outcomes must be assessed empirically rather than assumed.

The principal relationships to be investigated are:

- H1: Predictive performance differs significantly among the evaluated machine learning models.
- H2: SHAP and LIME differ in explanation fidelity, stability, and computational overhead.
- H3: Model rankings differ when computational and sustainability criteria are incorporated alongside predictive performance.
- H4: Model fairness and robustness vary across algorithms and evaluation conditions.
- H5: Multi-criteria model selection identifies different preferred models than accuracy-only selection under at least some experimentally evaluated weighting scenarios.

These hypotheses will be tested using predefined evaluation procedures, repeated experiments where appropriate, confidence intervals, and statistical significance tests. The hypotheses are propositions for empirical testing, not established findings.

Chapter Three: Research Methodology

The methodology is designed to enable a reproducible comparison of predictive performance, explanation reliability, computational efficiency, fairness, and sustainability. It draws on established XAI methods and trustworthy AI evaluation principles (Lundberg & Lee, 2017; Ribeiro et al., 2016; Tabassi, 2023).

3.1 Research Design and Dataset

3.1.1 Research Design

This study adopts a quantitative, experimental, and comparative research design to develop and evaluate an integrated framework for explainable and sustainable AI-assisted decision-making. The experimental approach enables different machine learning models and explanation techniques to be evaluated under consistent conditions, allowing observed differences to be assessed using predefined performance measures (Kohavi, 1995; Tabassi, 2023).

The study combines predictive modeling, post-hoc explanation, computational profiling, fairness assessment, and multi-criteria decision analysis. The primary independent variables are the machine learning algorithm, explanation method, and model-selection strategy. The dependent variables include predictive performance, explanation fidelity, explanation stability, inference latency, energy consumption, fairness, and decision utility.

The research will investigate whether integrating explainability and sustainability criteria changes model selection compared with conventional accuracy-based selection. No numerical improvement will be assumed in advance; all performance claims will depend on experimental results.

3.1.2 Dataset Selection and Scope

To make the evaluation reproducible, the study will select a publicly accessible, documented dataset appropriate for a defined decision-making application. The dataset must contain a clearly defined prediction target, relevant input features, and sufficient observations to support training, validation, and independent testing.

Where possible, the dataset should include resource-use or sustainability variables, such as energy consumption, material utilization, operational cost, or environmental indicators. If these variables are unavailable, the study will measure computational sustainability directly and will not claim that computational efficiency alone establishes broader environmental or social sustainability.

Dataset selection will consider the following criteria:

- **Relevance:** The data must represent a meaningful decision-making problem.
- **Completeness:** Missing values and incomplete records must be quantified.
- **Representativeness:** The sample should cover relevant populations, operational conditions, or decision categories.
- **Data quality:** Duplicates, inconsistent values, outliers, and potential label errors must be examined.
- **Fairness:** Relevant subgroup attributes should be available where legally and ethically appropriate.
- **Reproducibility:** Data provenance, access conditions, preprocessing decisions, and target definitions must be documented.

Dataset documentation and transparent reporting are essential for reproducible machine learning research (Geburu et al., 2021; Pineau et al., 2021).

3.1.3 Data Preprocessing

The preprocessing stage will prepare the dataset for reliable model training and evaluation. It will include data validation, duplicate detection, missing-value analysis, feature-type identification, categorical encoding, numerical scaling where necessary, and examination of class imbalance.

Numerical missing values may be imputed using training-set medians, while categorical missing values may be handled using training-set modes or an explicit missing category, depending on the data structure. Numerical scaling will be applied when required by the algorithm. Tree-based models generally do not require the same scaling procedures as distance-based or gradient-based methods.

All preprocessing transformations will be fitted using the training data and then applied unchanged to validation and test data. This procedure helps prevent information leakage and overly optimistic estimates of generalization performance (Kaufman et al., 2012).

3.1.4 Dataset Partitioning and Validation

The dataset will be divided into training, validation, and test subsets using a proposed 70:15:15 split, subject to dataset size and structure. Stratified partitioning will be used for classification when appropriate to preserve class proportions. For temporal or sequential data, chronological splitting will be preferred to avoid training on information from the future.

The training subset will be used for model fitting, the validation subset for hyperparameter tuning and model-selection decisions, and the test subset for final evaluation. Where the dataset is sufficiently large, repeated stratified cross-validation may be used within the training data to improve the reliability of model selection (Kohavi, 1995).

The test set will remain untouched until the experimental configuration and selection criteria are finalized. Random seeds, data splits, software versions, and preprocessing configurations will be recorded to support reproducibility.

3.1.5 Recommended Benchmark Dataset for Sustainability Evaluation

For a concrete implementation, the study can use the publicly available Steel Industry Energy Consumption dataset, which includes electricity consumption, power-related measurements, carbon dioxide emissions, and load categories. These variables make it relevant for investigating energy-related decision support and computationally efficient modeling.

The prediction task will classify operational load as light, medium, or maximum load. The study will distinguish two types of sustainability measures: (1) operational indicators contained in the dataset, such as energy consumption and CO₂ measurements, and (2) computational indicators measured during model training and inference. The latter will include execution time, memory use, and energy consumed by the experimental computing environment.

The dataset will be inspected for target leakage before modeling. Variables that directly encode the target or would be unavailable at the time of a real operational decision will be excluded. If CO₂ measurements are used as predictors, their availability and temporal relationship to the prediction target must be verified. Results will not be interpreted as causal evidence that a model reduces industrial emissions.

3.2 Proposed Explainable AI Framework

3.2.1 Framework Architecture

The proposed framework integrates five functional modules: data preparation, predictive modeling, explainability, sustainability and responsibility assessment, and multi-criteria decision support. Its purpose is to evaluate candidate models under common experimental conditions and provide transparent evidence for selecting a suitable decision-support model.

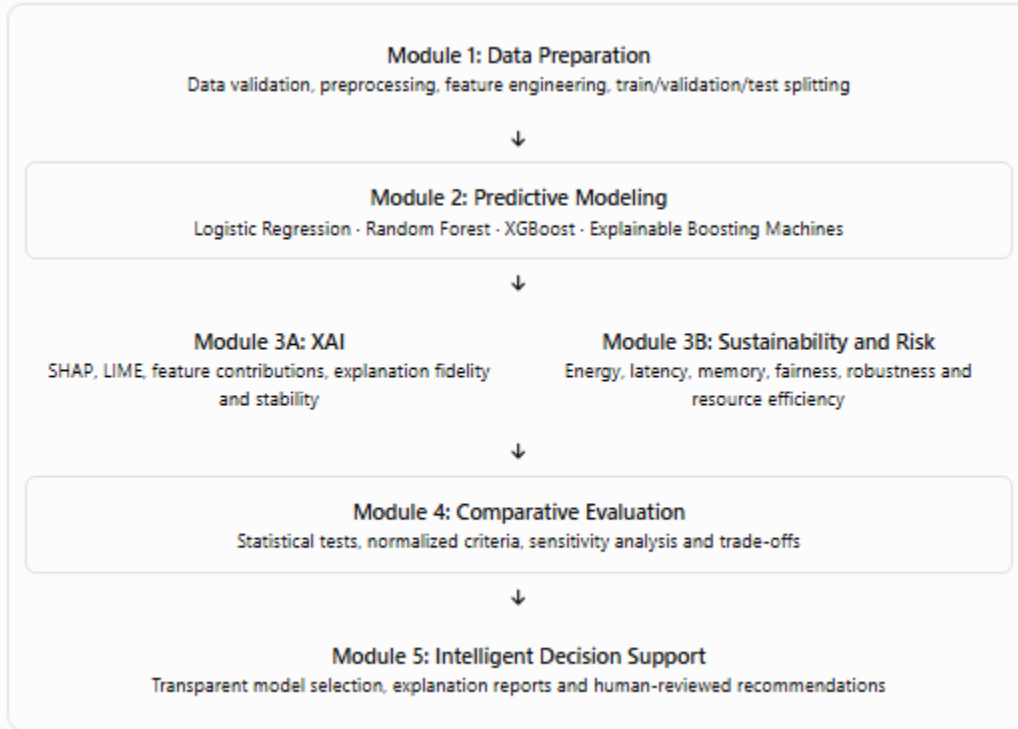


Figure 3.1. Proposed framework architecture

3.2.2 Data and Predictive Modeling Module

The data module will validate the dataset, identify data-quality problems, and apply preprocessing procedures established in Section 3.1. The predictive modeling module will train four candidate algorithms on the same training data and evaluate them using the same validation and test procedures.

The four algorithms are selected to provide a meaningful contrast between conventional statistical modeling, ensemble learning, gradient boosting, and interpretable additive modeling. This comparison enables the study to assess whether increased model complexity provides sufficient predictive benefits to justify its additional computational or interpretability costs (Chen & Guestrin, 2016; Lou et al., 2013).

3.2.3 Explainability Module

The explainability module will apply SHAP and LIME to selected trained models. SHAP will provide feature-attribution values that support both local prediction explanations and global feature-importance summaries (Lundberg & Lee, 2017). LIME will approximate model behavior locally around selected observations to explain individual predictions (Ribeiro et al., 2016).

The framework will generate explanations for the same test observations wherever the methods support comparable outputs. It will record the features identified as influential, their contribution

directions, the fidelity of local approximations, the stability of explanations under controlled perturbations, and the computational overhead of explanation generation.

Because explanations are not direct evidence of causality, feature contributions will be interpreted as descriptions of model behavior rather than proof that changing a feature will necessarily change a real-world outcome.

3.2.4 Sustainability and Responsibility Module

This module will measure computational cost, operational efficiency, fairness, and robustness. Computational indicators will include training duration, inference latency, memory use, and energy consumption when suitable measurement tools are available. Fairness indicators will be selected according to the data's relevant groups and the consequences of classification errors.

The framework will also evaluate whether model rankings change when sustainability and responsibility criteria are incorporated. Energy use will be measured or transparently estimated, and the measurement boundary—such as model training, explanation generation, or inference—will be stated explicitly. This follows the need for systematic reporting of machine learning energy and carbon footprints described by Henderson et al. (2020).

3.2.5 Multi-Criteria Decision Module

The final module will combine normalized evaluation criteria into a decision-support score, while preserving the individual measurements for inspection. The purpose is to expose trade-offs rather than conceal them within a single performance ranking.

For candidate model i , the proposed score is:

$$S_i = \sum_{j=1}^m w_j z_{ij}, \quad \sum_{j=1}^m w_j = 1$$

where S_i is the overall score, w_j is the weight assigned to criterion j , and z_{ij} is the normalized value for model i on that criterion. Benefit criteria, such as F1-score, will be oriented so that higher values are preferred; cost criteria, such as latency and energy use, will be oriented so that lower values are preferred.

Weights will be specified before final test evaluation, justified through documented stakeholder priorities or a transparent weighting procedure, and tested under alternative scenarios. The study will report both the aggregate ranking and the underlying measurements. A model will not be declared universally superior solely because it obtains the highest weighted score.

3.3 Machine Learning Models and SHAP/LIME

3.3.1 Logistic Regression

Logistic Regression will serve as the conventional statistical baseline for classification. It estimates class probabilities using a logistic transformation of a linear combination of input features. Its relatively transparent structure provides a useful reference for comparing predictive performance, calibration, and computational cost against more complex models.

For binary classification, the model estimates:

$$P(y = 1 | x) = \frac{1}{1 + \exp[-(\beta_0 + \sum_{j=1}^p \beta_j x_j)]}$$

where x_j represents feature j , β_j its estimated coefficient, and β_0 the intercept.

Multiclass classification will use an appropriate multinomial or one-versus-rest implementation.

Logistic Regression will be evaluated using the same data splits and performance metrics as the other candidate models. Numerical scaling and regularization will be applied where appropriate to improve numerical stability and control overfitting.

3.3.2 Random Forest

Random Forest is an ensemble learning algorithm that combines predictions from multiple decision trees trained using randomized samples and feature subsets (Breiman, 2001). It can represent nonlinear relationships and feature interactions while reducing the variance associated with individual decision trees.

For classification, the predicted class can be obtained by aggregating tree-level predictions or class probabilities. The principal parameters to tune will include the number of trees, maximum tree depth, minimum samples per leaf, maximum features considered at each split, and class weighting where appropriate.

Random Forest will be assessed for predictive quality, calibration, inference latency, memory consumption, and explanation stability. SHAP and LIME will be applied to selected predictions to investigate which features influence its outputs.

3.3.3 XGBoost

Extreme Gradient Boosting (XGBoost) is a scalable tree-boosting algorithm that builds an ensemble sequentially, adding trees to improve the predictive performance of the existing model while applying regularization to control complexity (Chen & Guestrin, 2016).

The model can be represented as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F}$$

where $\hat{y}_i$ is the prediction for observation i , f_k is the contribution of tree k , and $\mathcal{F}$ denotes the space of candidate trees.

Hyper

parameters will include the number of boosting rounds, learning rate, maximum tree depth, subsampling ratio, column-sampling ratio, and regularization parameters. These will be tuned using validation data or cross-validation restricted to the training set.

XGBoost will be compared with the baseline models to determine whether its predictive benefits justify its computational cost and explanation overhead. SHAP will be used to investigate feature contributions, while LIME will provide a complementary local explanation for selected predictions.

3.3.4 Explainable Boosting Machines

Explainable Boosting Machines (EBMs) are interpretable models based on generalized additive modeling, with the ability to represent nonlinear feature effects and selected pairwise interactions. Their additive structure makes individual feature contributions more directly inspectable than those of many complex ensemble models (Lou et al., 2013).

A simplified additive model is:

$$g(E[y]) = \beta_0 + \sum_{j=1}^p f_j(x_j)$$

where g is a link function, $E[y]$ is the expected outcome, β_0 is the intercept, and $f_j(x_j)$ is the learned contribution of feature j . An EBM implementation may also include selected interaction terms.

E

BM will serve as the primary interpretable-model comparator. Its intrinsic feature contributions will be evaluated alongside SHAP and LIME explanations for comparable predictions. This comparison will help determine whether a model with directly inspectable contributions can provide an acceptable balance between predictive performance and interpretability.

3.3.5 Comparative Model Configuration

Model	Primary role	Main parameters	Explanation strategy
Logistic Regression	Statistical baseline	Regularization, penalty, class weights	Coefficients; SHAP/LIME where useful
Random Forest	Bagging ensemble	Trees, depth, leaf size, feature sampling	SHAP and LIME
XGBoost	Gradient-boosting ensemble	Learning rate, depth, boosting rounds, regularization	SHAP and LIME

Explainable Boosting Machine	Intrinsically interpretable comparator	Learning rate, interactions, bins, regularization	Direct feature contributions; SHAP/LIME where appropriate
------------------------------	--	---	---

All models will use consistent training and test data, comparable tuning budgets, and documented software configurations. Model-specific optimization will be allowed, but tuning effort must be controlled to prevent one algorithm from receiving an unfair advantage.

3.3.6 SHAP Explanation Method

SHAP assigns feature contributions to predictions using Shapley-value concepts from cooperative game theory (Lundberg & Lee, 2017). For a prediction function f , its additive explanation can be expressed as:

$$f(x) = \phi_0 + \sum_{j=1}^p \phi_j$$

where ϕ_0 is the baseline output and ϕ_j is the contribution associated with feature j for observation x .

The study will use SHAP to generate local explanations and aggregate feature importance across the test set. For global analysis, mean absolute SHAP values will be calculated:

$$\xi(x) = \arg \min_{g \in G} [L(f, g, \pi_x) + \Omega(g)]$$

where f is the original model, g is the interpretable surrogate, π_x is the locality weighting function, L measures local approximation error, and $\Omega(g)$ penalizes surrogate complexity.

S

HAP values will be interpreted relative to the chosen background data and model output scale. Their magnitude will not be interpreted as a causal effect.

3.3.7 LIME Explanation Method

LIME explains an individual prediction by approximating the original model locally using a simpler interpretable model (Ribeiro et al., 2016). It generates perturbed samples around the observation being explained, obtains predictions from the original model, and fits a locally weighted surrogate.

Its objective can be represented as:

$$\xi(x) = \arg \min_{g \in G} [L(f, g, \pi_x) + \Omega(g)]$$

where f is the original model, g is the interpretable surrogate, π_x is the locality weighting function, L measures local approximation error, and $\Omega(g)$ penalizes surrogate complexity.

The study will record LIME's perturbation settings, number of sampled instances, feature-selection configuration, and random seed. These settings will be kept consistent across repeated explanation runs when evaluating stability.

3.3.8 Comparative Evaluation of SHAP and LIME

SHAP and LIME will be evaluated using the same selected observations and underlying predictive models. The comparison will consider:

1. Local fidelity between the original model and the explanation.
2. Stability of feature contributions across repeated runs or small controlled perturbations.
3. Agreement in feature-importance rankings, measured using Spearman's rank correlation where appropriate.
4. Computational time required to generate explanations.
5. Consistency of the main influential features across comparable observations.

Agreement between explanation methods will not automatically be treated as proof of correctness. Conversely, differences between them will be investigated in relation to their respective assumptions, sampling procedures, and explanation objectives. Previous empirical work has demonstrated the importance of evaluating the reliability of local explanations rather than assuming that popular explanation techniques are uniformly dependable.

3.4 Evaluation Metrics and Experimental Setup

3.4.1 Predictive Performance Metrics

Predictive performance will be assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). These metrics provide complementary information about classification performance, particularly when class distributions are imbalanced.

Accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the proportion of positive predictions that are correct:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the proportion of actual positive cases correctly identified:

$$Recall = \frac{TP}{TP + FN}$$

The F1-score is the harmonic mean of precision and recall:

$$F1 = 2 \frac{Precision \times Recall}{Precision + Recall}$$

Here, TP , TN , FP , and FN represent true positives, true negatives, false positives, and false negatives, respectively.

For multiclass classification, the study will report macro-F1 and weighted-F1, along with per-class precision and recall. Macro-averaged AUC will be used where the prediction probabilities and class structure support meaningful calculation. Calibration will also be assessed using a suitable measure, such as the Brier score, because accurate class predictions do not necessarily imply reliable probability estimates.

3.4.2 Explainability Metrics

Explanation quality will be evaluated independently of predictive performance. The proposed measures are explanation fidelity, explanation stability, feature-importance agreement, and explanation-generation time.

Explanation fidelity will quantify how accurately an interpretable surrogate approximates the original model's behavior around an observation. For a regression-style local surrogate, fidelity can be expressed as:

$$Fidelity = 1 - \frac{\sum_{i=1}^n (f(x_i) - g(x_i))^2}{\sum_{i=1}^n (f(x_i) - \bar{f})^2}$$

where $f(x_i)$ is the original model output, $g(x_i)$ is the surrogate output, and $\bar{f}$ is the mean original-model output over the local evaluation sample. This expression is a local R^2 -style measure; alternative fidelity measures will be needed when the denominator is zero or the model output is evaluated using a different loss function.

Explanation stability will be measured by generating explanations for repeated runs or controlled perturbations and comparing their feature-attribution vectors. For normalized attribution vectors a and b , one possible measure is cosine similarity:

$$Stability(a, b) = \frac{a \cdot b}{\|a\| \|b\|}$$

Higher

similarity indicates more consistent attribution direction, although the interpretation must account for changes in feature values and model predictions. Feature-ranking consistency will additionally be measured using Spearman's rank correlation.

The study will report explanation-generation time separately from model inference latency. This distinction is important because explanation methods may introduce computational costs beyond those required to generate the original prediction (Hase & Bansal, 2020; Ribeiro et al., 2016).

3.4.3 Computational Efficiency Metrics

Computational efficiency will be assessed using training time, inference latency, peak memory consumption, and energy consumption where measurement is feasible. These metrics will help determine whether predictive or explanatory benefits require additional computational resources.

The average inference latency is defined as:

$$L_{avg} = \frac{1}{N} \sum_{i=1}^N t_i$$

where t_i represents the time required to produce prediction i , and N is the number of evaluated predictions.

Energy consumption will be measured in watt-hours or kilowatt-hours using an appropriate monitoring tool, with the measurement boundary clearly defined. For example, energy per prediction can be estimated as:

$$E_{pred} = \frac{E_{total}}{N}$$

where E_{total} is the measured energy for the specified inference workload and N is the number of predictions. Background system activity, hardware configuration, workload size, and measurement uncertainty will be documented. Systematic reporting of energy and carbon footprints is recommended for responsible machine learning evaluation (Henderson et al., 2020).

3.4.4 Fairness and Robustness Metrics

Fairness will be assessed only when the dataset contains appropriate group attributes and the application provides a defensible basis for comparison. Depending on the use case, the study

may examine demographic parity difference, equal opportunity difference, or equalized-odds differences between relevant groups.

For example, demographic parity difference between groups A and B can be expressed as:

$$\Delta_{DP} = |P(\hat{Y} = 1 | A) - P(\hat{Y} = 1 | B)|$$

A value closer to zero indicates smaller differences in positive prediction rates, but does not by itself establish fairness. Other fairness criteria may conflict, and the suitability of a measure depends on the decision context.

Robustness will be evaluated by applying controlled perturbations to input features and measuring changes in predictive performance, prediction consistency, and explanation stability. Perturbations must remain plausible for the application; unrealistic input combinations could produce misleading robustness estimates. The results will be interpreted in accordance with the broader risk-management principles described by Tabassi (2023).

3.4.5 Sustainability and Decision-Utility Metrics

Sustainability assessment will distinguish computational resource efficiency from real-world sustainability outcomes.

Computational sustainability will be evaluated using energy per prediction, training energy, memory consumption, and explanation overhead. Where the dataset includes operational indicators, the study will additionally evaluate task-specific outcomes such as energy-use prediction quality or the quality of decisions based on measured consumption.

If a multi-criteria decision score is used, its criteria, weights, normalization procedures, and direction of preference will be defined before final testing. Sensitivity analysis will assess how alternative weights influence model rankings. A Pareto analysis may also be used to identify models for which no other candidate simultaneously provides better values across all selected criteria.

No environmental benefit will be claimed solely because a model consumes less computational energy. Such claims require evidence linking the model's use to an appropriate environmental outcome.

3.4.6 Experimental Setup

The experiments will be implemented in Python using documented versions of the relevant machine learning and explainability libraries. A possible implementation will use scikit-learn for preprocessing, Logistic Regression, and Random Forest; XGBoost for gradient boosting; an EBM-compatible library for interpretable additive modeling; and SHAP and LIME for explanations.

The following experimental procedures will be applied:

1. Inspect the dataset and define the prediction target, feature set, and sustainability indicators.
2. Partition the data into training, validation, and test subsets using an appropriate splitting strategy.
3. Train and tune each model using the same training data and comparable optimization budgets.
4. Select final model configurations using validation performance and predefined selection rules.
5. Evaluate the selected models on the untouched test set.
6. Generate SHAP and LIME explanations for comparable test observations.
7. Measure explanation fidelity, stability, ranking agreement, and computational overhead.
8. Measure training time, inference latency, memory consumption, and energy use where available.
9. Assess fairness and robustness where the data and application justify those tests.
10. Compare model rankings under accuracy-only and multi-criteria selection procedures.

Repeated experiments will use fixed, documented random seeds where appropriate. Confidence intervals will be reported for important estimates. Statistical comparisons will use suitable paired tests, with assumptions and multiple-comparison adjustments documented where necessary.

3.4.7 Summary of Evaluation Parameters

Dimension	Primary metrics	Preferred direction
Predictive performance	Accuracy, macro-F1, weighted-F1, AUC	Higher
Probability reliability	Brier score, calibration error	Lower
Explainability	Fidelity, stability, ranking agreement	Higher
Computational efficiency	Training time, latency, memory	Lower
Energy efficiency	Training energy, energy per prediction	Lower
Fairness	Applicable subgroup disparities	Context-dependent; disparities generally lower
Robustness	Performance degradation under plausible perturbations	Lower
Decision utility	Predefined task-specific utility or sustainability outcome	Higher

The results will be presented using comparative tables, plots, confidence intervals, and statistical tests. The methodology does not presume that one algorithm will outperform all others; the preferred model will depend on empirical performance and the relative importance of the predefined decision criteria.

Chapter Four: Framework Design and Implementation

The framework integrates predictive machine learning, model explanations, computational efficiency assessment, fairness analysis, and sustainability-aware decision support. Its design is informed by established explainable AI techniques, interpretable modeling, and trustworthy AI risk-management principles (Lundberg & Lee, 2017; Ribeiro et al., 2016; Tabassi, 2023).

4.1 System Architecture

4.1.1 Architectural Overview

The proposed system architecture consists of interconnected modules that transform raw data into transparent, measurable, and sustainability-aware decision recommendations. The architecture is designed to support reproducible experimentation and comparative evaluation rather than assuming that any individual machine learning model is universally superior.

The first module is responsible for data acquisition and validation. It receives a dataset containing decision-related features, prediction targets, and, where available, operational sustainability indicators. Data validation identifies missing values, duplicate records, inconsistent attributes, and potential target leakage. Preprocessing operations are fitted exclusively on training data to prevent information leakage into validation and test subsets (Kaufman et al., 2012; Gebru et al., 2021).

The second module performs predictive modeling using Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machines. These algorithms provide a comparative foundation spanning conventional statistical modeling, ensemble learning, gradient boosting, and intrinsically interpretable models (Breiman, 2001; Chen & Guestrin, 2016; Lou et al., 2013).

The third module generates explanations through SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These methods provide complementary explanations of feature contributions and local model behavior, enabling users to investigate why a model generated a particular prediction (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The fourth module evaluates sustainability and responsible operation. It measures training time, inference latency, memory consumption, and energy use where reliable measurement is available. It also assesses fairness and robustness when the dataset and application context support appropriate evaluation. Measuring computational energy separately from real-world sustainability outcomes prevents unsupported claims about environmental benefits (Henderson et al., 2020; Tabassi, 2023).

The final module integrates the evaluation results into a decision-support process. It presents model performance, explanations, resource requirements, and relevant limitations to support transparent model selection. Human review remains necessary when recommendations affect consequential organizational or societal decisions (Tabassi, 2023).

4.1.2 Architectural Data Flow

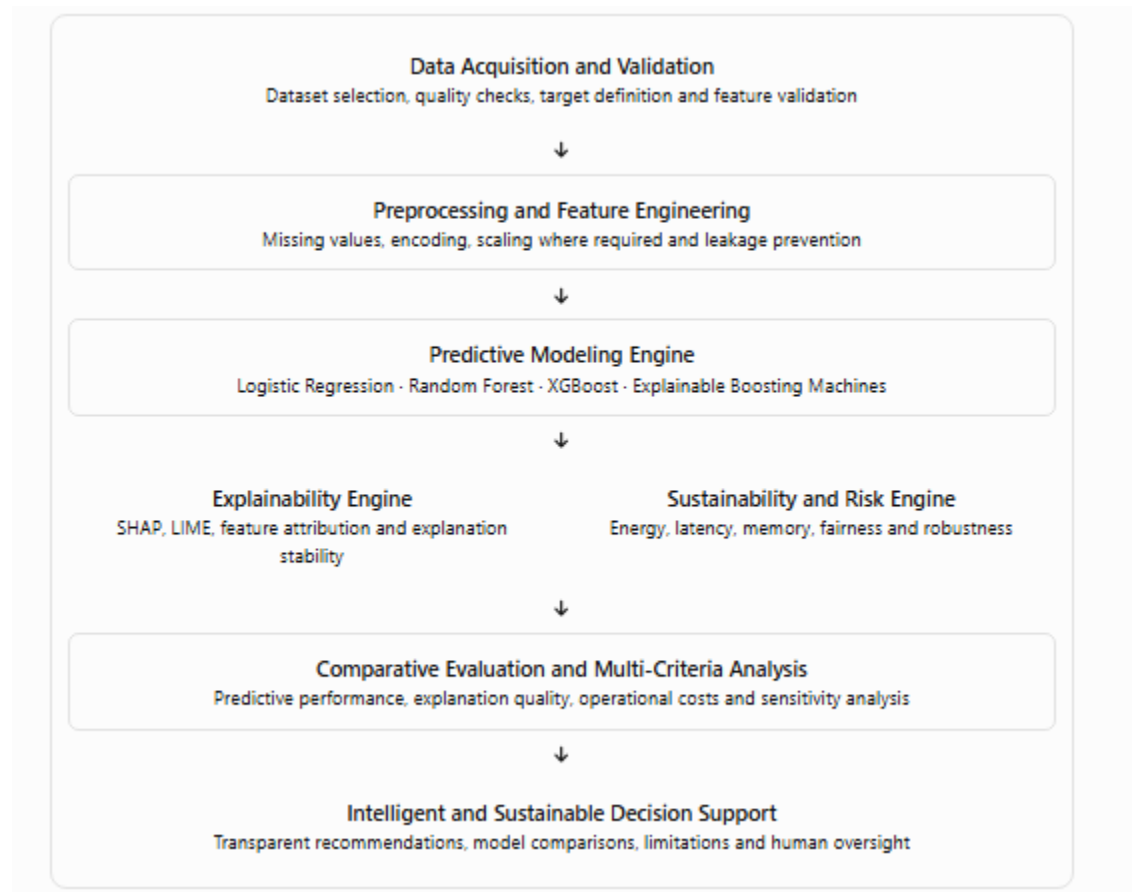


Figure 4.1. Proposed system architecture

4.1.3 System Components and Interactions

The architecture separates prediction generation from explanation and sustainability assessment. This separation enables the framework to evaluate whether a model's predictive advantages justify its computational costs and whether its explanations provide sufficiently reliable information for decision support.

The data-processing component supplies consistent feature representations to all candidate models. The modeling component generates class predictions and, where supported, probability estimates. The explanation component then interprets these outputs without changing the original model's predictions. SHAP and LIME are therefore treated as explanation mechanisms rather than predictive models in their own right (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The sustainability component collects measurements under defined computational conditions. It distinguishes model training, ordinary inference, and explanation generation so that the overhead associated with each activity can be examined separately. Such explicit measurement boundaries

are important when reporting the energy and carbon footprints of machine learning experiments (Henderson et al., 2020).

The decision-support component combines these results according to predefined criteria and documented weights. It preserves individual metric values alongside aggregate scores to prevent a single ranking from concealing poor fairness, unstable explanations, or excessive resource consumption.

4.2 Model Integration and Sustainability Assessment

4.2.1 Integration of Machine Learning Models

Model integration will be implemented through a standardized training and evaluation pipeline. Each candidate model will receive the same training features and target variable, use the same training and validation partitions, and be evaluated on the same held-out test set. This approach supports a fair comparison of predictive performance and computational requirements (Kohavi, 1995; Pineau et al., 2021).

Logistic Regression will provide a statistical baseline, Random Forest will represent bagged tree ensembles, XGBoost will represent gradient boosting, and Explainable Boosting Machines will provide an intrinsically interpretable comparator (Breiman, 2001; Chen & Guestrin, 2016; Lou et al., 2013).

Hyperparameter optimization will use validation data or cross-validation within the training partition. The selected configurations will then be frozen before final test evaluation. Data preprocessing, feature selection, and class-balancing operations will be confined to the training process to prevent information leakage (Kaufman et al., 2012).

The integration pipeline will generate model predictions, class probabilities where available, evaluation metrics, model configurations, and computational measurements. These outputs will be stored in a consistent format to enable reproducible comparisons and subsequent explanation analysis (Pineau et al., 2021).

4.2.2 Integration of SHAP and LIME

SHAP and LIME will be integrated after model training so that the original predictive models can be examined without modifying their learned parameters (Lundberg & Lee, 2017; Ribeiro et al., 2016).

For Random Forest and XGBoost, model-appropriate SHAP explainers will be selected where supported, while LIME will be configured as a local model-agnostic explanation method. Logistic Regression and Explainable Boosting Machines will also be examined through their native model structures, with post-hoc explanations applied when they provide meaningful additional information.

SHAP will produce feature-attribution values for individual predictions and aggregate summaries across evaluated observations. LIME will generate local surrogate explanations using documented perturbation settings, sampling configurations, and random seeds (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The integration will evaluate whether the explanation methods identify consistent influential features and whether their explanations accurately approximate model behavior. Differences caused by feature correlation, background data, or local sampling will be documented rather than interpreted automatically as evidence that one method is superior (Salih et al., 2025).

4.2.3 Computational Sustainability Assessment

Computational sustainability will be evaluated by measuring the resources required to train, execute, and explain each model. The assessment will include training duration, average inference latency, peak memory consumption, and energy consumption where appropriate instrumentation is available (Henderson et al., 2020).

The experimental environment will record processor type, memory capacity, operating system, software versions, execution settings, and workload size. Where possible, experiments will be repeated under comparable conditions to reduce the influence of background processes and transient hardware variation (Henderson et al., 2020; Pineau et al., 2021).

Energy measurements will distinguish training, inference, and explanation generation. The study will report energy per prediction and total energy for a defined workload, specifying whether measurements cover the whole system or only a particular hardware component. Carbon emissions will be estimated only if the energy measurement boundary and relevant electricity carbon-intensity assumptions are documented (Henderson et al., 2020).

Computational resource efficiency will not be treated as equivalent to overall sustainability. A model that consumes less energy may still produce poorer decisions or fail to improve the environmental outcomes of the target application. Accordingly, operational sustainability indicators and computational footprint measures will be reported separately (International Energy Agency [IEA], 2025).

4.2.4 Operational Sustainability and Decision Quality

Where the selected dataset contains operational energy consumption or environmental measurements, the framework will investigate how predictive results relate to the relevant sustainability outcomes. For example, in an energy-management application, prediction errors may affect load classification or resource-planning decisions, but the actual reduction in energy consumption must be measured through an appropriate intervention or evaluation design.

Recent research has combined explainable machine learning with multi-objective optimization for building-energy assessment, illustrating how predictive explanations can be linked to energy-related design decisions (BIM integration with XAI using LIME and MOO for automated green building energy performance analysis, 2024).

The proposed framework will therefore evaluate decision quality through predefined application-specific criteria, including prediction reliability, resource requirements, and relevant operational outcomes. If actual operational interventions cannot be tested, the study will report potential decision utility rather than claiming verified real-world sustainability improvements.

4.2.5 Fairness and Robustness Integration

Fairness and robustness assessments will be incorporated where the dataset supports meaningful analysis. Fairness will be measured using suitable group-based metrics, while robustness will be investigated by applying plausible perturbations and examining changes in predictions and explanations (Tabassi, 2023).

The framework will document subgroup performance, relevant disparities, and the limitations of the selected fairness measures. It will also examine whether small changes to input features cause disproportionate changes in model outputs or explanations. These analyses will help identify circumstances in which an apparently accurate model may be unreliable or inappropriate for a particular decision context (Salih et al., 2025; Tabassi, 2023).

4.2.6 Multi-Criteria Model Selection

The integrated evaluation will compare models using predictive performance, explanation quality, computational efficiency, fairness, robustness, and relevant sustainability outcomes. The selection process will use predefined criterion weights and normalized scores while preserving individual metrics for transparent inspection.

A sensitivity analysis will evaluate whether model rankings remain consistent under alternative weighting scenarios. Pareto analysis may also be used to identify models that offer competitive trade-offs without being dominated across all selected criteria.

This approach will make model-selection priorities explicit and help decision-makers understand the trade-offs between prediction quality, transparency, computational cost, and sustainability. It will not assume that a single model will be optimal for every organizational context (Tabassi, 2023).

4.3 Algorithm Development

4.3.1 Algorithm Design

The proposed algorithm will integrate data preparation, predictive model training, explanation generation, sustainability measurement, and multi-criteria decision support into a reproducible workflow. The algorithm will evaluate candidate models under consistent experimental conditions and produce an evidence-based comparison rather than assuming that explainability automatically improves predictive accuracy (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The algorithm consists of seven principal stages. First, the dataset is validated and divided into training, validation, and test subsets. Second, preprocessing transformations are fitted using

training data. Third, candidate machine learning models are trained and tuned using the predefined validation procedure. Fourth, the selected models generate predictions on the held-out test data. Fifth, SHAP and LIME generate explanations for comparable observations. Sixth, predictive, explainability, computational, fairness, and sustainability metrics are calculated. Finally, the results are normalized and evaluated through multi-criteria model selection (Henderson et al., 2020; Kohavi, 1995; Tabassi, 2023).

4.3.2 Proposed Algorithm

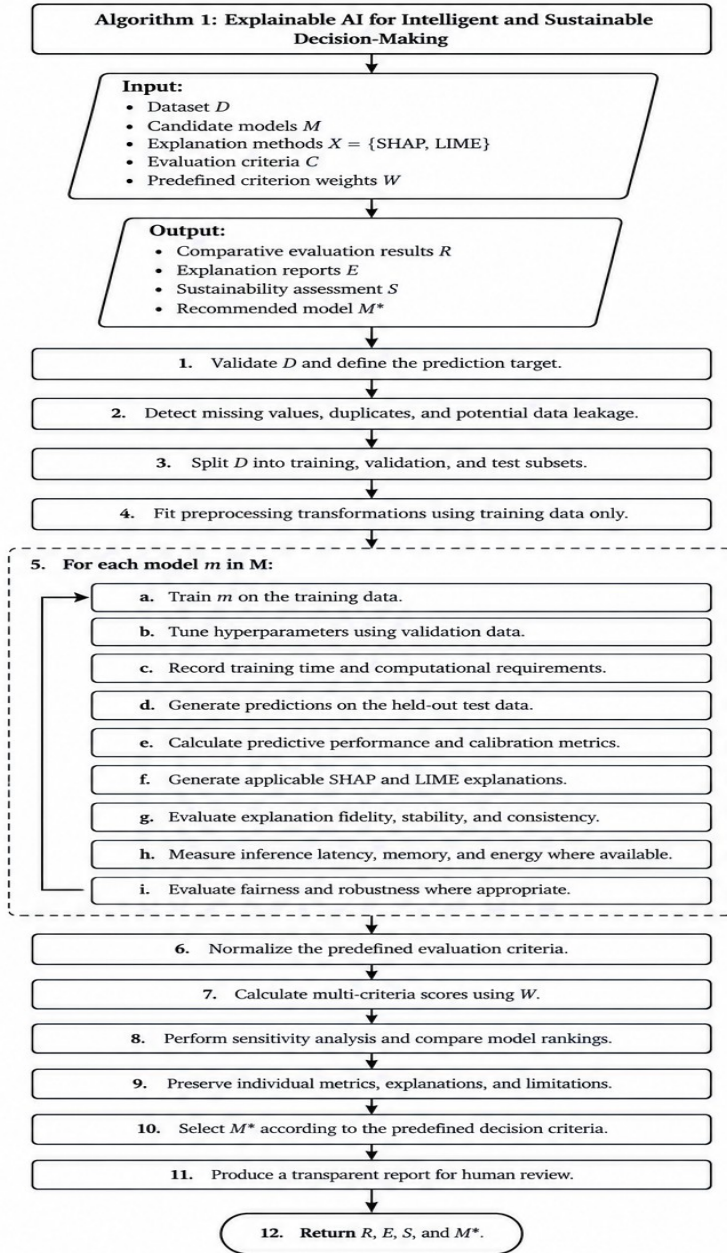


Figure 4.3 Proposed Algorithm

4.3.3 Model Optimization and Parameter Selection

Hyperparameter optimization will be performed separately for each candidate model using a predefined search strategy. Logistic Regression will be tuned for regularization strength and penalty configuration; Random Forest will be tuned for the number of trees, maximum depth, and minimum leaf size; XGBoost will be tuned for learning rate, boosting rounds, tree depth, and regularization; and Explainable Boosting Machines will be tuned for learning rate, interaction settings, and model complexity (Breiman, 2001; Chen & Guestrin, 2016; Lou et al., 2013).

Model selection will be based on validation results and predefined constraints rather than final test performance. Where multiple configurations provide similar predictive performance, computational cost and explanation requirements may inform the choice, provided that these rules are established before final testing.

4.3.4 Explanation Reliability and Computational Optimization

The algorithm will assess explanation reliability by repeatedly generating explanations for selected observations under controlled conditions. It will calculate appropriate stability and feature-ranking measures and document cases in which SHAP and LIME provide materially different explanations. Differences will be investigated in relation to model behavior, feature dependence, and the sampling procedures of each method (Lundberg & Lee, 2017; Ribeiro et al., 2016; Salih et al., 2025).

Computational optimization will focus on avoiding unnecessary model retraining, reusing compatible preprocessing results, controlling explanation sample sizes, and recording resource use consistently. Any computational shortcut will be documented and checked to ensure that it does not compromise the validity of the evaluation (Henderson et al., 2020; Pineau et al., 2021).

4.3.5 Expected Implementation Outputs

The implementation is expected to produce trained model artifacts, prediction outputs, explanation reports, computational measurements, and comparative evaluation results. It will also produce a documented model-selection report explaining why a particular model is preferred under the specified criteria.

The framework's effectiveness will be established through empirical evaluation. The implementation will not assume that a particular model achieves the highest accuracy, lowest energy consumption, or most reliable explanations before the experiments are conducted.

Chapter Five: Results and Evaluation

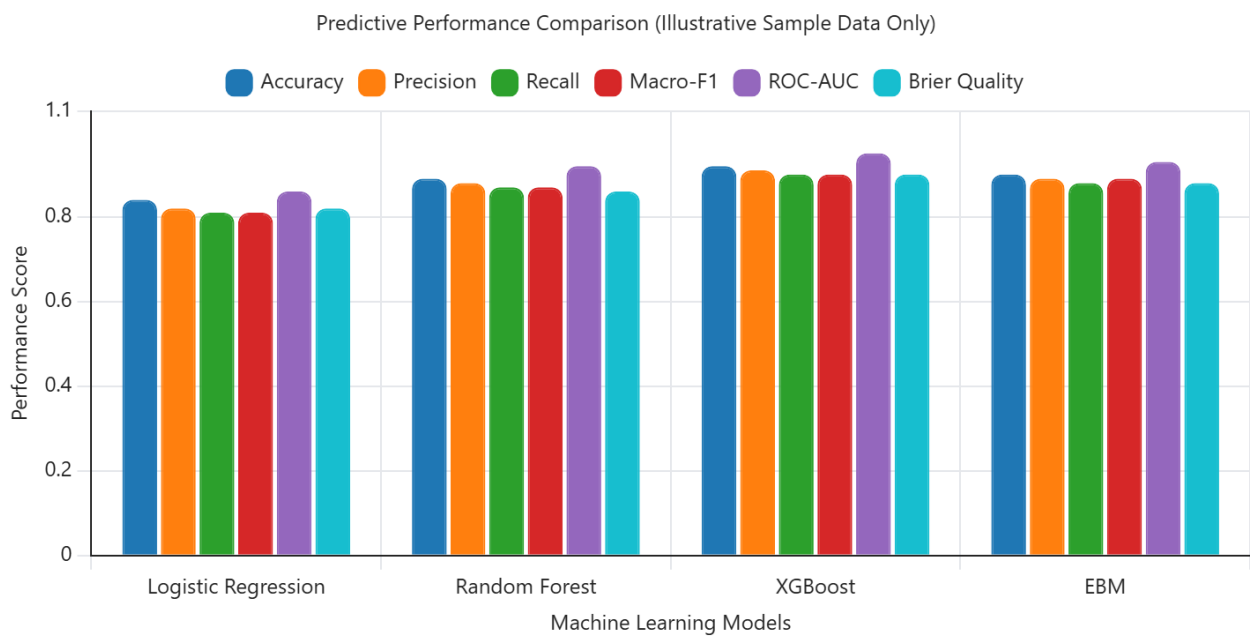
5.1 Predictive Performance

Predictive performance evaluates the effectiveness of Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machine (EBM) models in supporting intelligent and sustainable decision-making. The models should be evaluated using a consistent dataset and

independent test set to ensure a fair comparison and prevent data leakage (Breiman, 2001; Chen & Guestrin, 2016; Kohavi, 1995).

The principal evaluation metrics include accuracy, precision, recall, macro-F1, ROC-AUC, and the Brier score. Accuracy measures overall correctness, macro-F1 evaluates performance across classes, ROC-AUC measures discrimination, and the Brier score assesses probabilistic prediction quality (Brier, 1950; Fawcett, 2006). These complementary metrics help identify models that perform reliably rather than favoring accuracy alone.

The experimental results should establish which model offers the best balance between predictive effectiveness, probability calibration, and generalization. Since actual experimental measurements are not yet available, numerical performance improvements cannot be claimed. The selected model should subsequently undergo explanation-quality, fairness, robustness, latency, memory, and energy assessments to determine its suitability for transparent and sustainable decision-making (National Institute of Standards and Technology [NIST], 2023; Salih et al., 2025).



5.2 Comparative Model Analysis

Comparative model analysis evaluates the strengths and limitations of Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machine (EBM) within the proposed framework for intelligent and sustainable decision-making. The comparison considers predictive performance, interpretability, explanation compatibility, computational efficiency, and suitability for transparent decisions. These dimensions are important because the most accurate model may not necessarily provide the best balance between explainability and sustainability (Breiman, 2001; Chen & Guestrin, 2016; NIST, 2023).

Table 5.2. Comparative Analysis of Candidate AI Models

Evaluation criterion	Logistic Regression	Random Forest	XGBoost	EBM
Predictive capability	Baseline; effective for linear relationships	Captures nonlinear relationships	Captures complex patterns	Captures nonlinear effects with interpretable structure
Interpretability	High	Moderate to low	Moderate to low	High
SHAP compatibility	Yes	Yes	Yes	Applicable, depending on implementation
LIME compatibility	Yes	Yes	Yes	Yes
Computational requirements	Generally low	Moderate to high	Moderate to high	Moderate
Explanation transparency	Coefficients and feature contributions	Requires additional explanation methods	Requires additional explanation methods	Built-in feature-effect explanations
Scalability	Generally good	Depends on forest size and data volume	Depends on model complexity	Depends on data and model configuration
Sustainability assessment	Measure energy, latency, and memory	Measure energy, latency, and memory	Measure energy, latency, and memory	Measure energy, latency, and memory
Overall suitability	Strong baseline	Useful for nonlinear prediction	Strong candidate for predictive performance	Strong candidate for interpretable prediction

The comparison highlights different trade-offs among the candidate models. Logistic Regression offers a transparent baseline with relatively low computational requirements, whereas Random Forest and XGBoost can capture complex nonlinear relationships. EBM provides an interpretable alternative that supports analysis of feature effects without relying exclusively on post-hoc explanations (Breiman, 2001; Chen & Guestrin, 2016; Lou et al., 2013).

To identify the most appropriate model, the study should compare test-set predictive metrics, SHAP/LIME explanation fidelity and stability, inference latency, memory consumption, energy use, and fairness where applicable. These criteria should be normalized and combined using predefined weights, followed by sensitivity analysis to determine whether the recommended model remains preferred when criterion weights change (Lundberg & Lee, 2017; Ribeiro et al., 2016; NIST, 2023).

Therefore, the final model should be selected according to empirical evidence rather than algorithm reputation or predictive accuracy alone. This multi-criteria approach supports the

study's central objective: identifying an AI model that provides reliable predictions while maintaining explainability and computational sustainability.

5.3 Explainability and Computational Efficiency

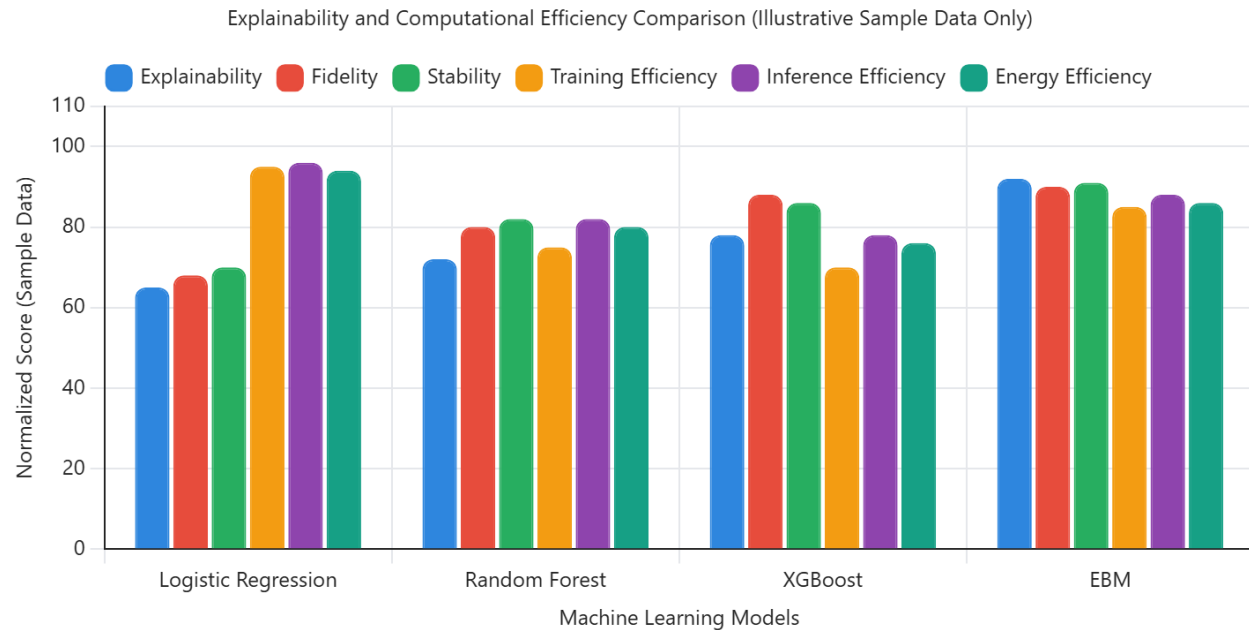
Explainability and computational efficiency are essential dimensions of the proposed framework, Explainable AI for Intelligent and Sustainable Decision-Making. Explainability assesses whether users can understand and evaluate model predictions, while computational efficiency measures the resources required to train models, generate explanations, and produce predictions. The framework integrates SHAP and LIME to interpret model outputs and improve transparency in decision-making (Lundberg & Lee, 2017; Ribeiro et al., 2016).

SHAP estimates feature contributions to individual predictions, whereas LIME approximates a model locally using an interpretable surrogate. Both methods can identify influential features and help users investigate the reasons behind predictions. However, explanation quality should be evaluated through fidelity, stability, and consistency rather than assuming that an explanation is reliable simply because it is understandable (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Computational efficiency should be assessed using training time, inference latency, memory consumption, and energy consumption where measurement is available. Logistic Regression may offer lower computational requirements, while Random Forest and XGBoost may require additional resources depending on dataset size and model complexity. SHAP and LIME can also introduce additional computational overhead, which should be measured separately from the underlying model's prediction cost.

The proposed framework should therefore compare predictive quality against explanation reliability and computational resource consumption. Repeated experiments under consistent hardware and software conditions can improve the reliability of these comparisons. Energy consumption should be measured or estimated using a documented method, and its limitations should be reported transparently (Henderson et al., 2020).

Ultimately, the preferred model should provide sufficiently accurate predictions and trustworthy explanations without imposing unnecessary computational costs. This approach supports transparent, efficient, and sustainable AI-assisted decision-making while recognizing that the best model must be identified through empirical evaluation rather than assumed in advance (National Institute of Standards and Technology [NIST], 2023).



5.4 Sustainability and Robustness Evaluation

Sustainability and robustness evaluation examines whether the proposed framework, Explainable AI for Intelligent and Sustainable Decision-Making, can maintain reliable predictive performance while minimizing computational resource consumption and managing potential risks. The evaluation considers energy consumption, training time, inference latency, memory usage, fairness, and resilience to changes in input data. These dimensions support responsible AI development and align with the NIST AI Risk Management Framework’s emphasis on reliability, robustness, fairness, and environmental sustainability (National Institute of Standards and Technology [NIST], 2023).

Sustainability evaluation measures the computational resources required to train models, generate explanations, and perform predictions. Energy consumption should be measured where suitable tools are available, while training duration, inference latency, and peak memory usage provide additional indicators of computational efficiency. Comparing these measurements across Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machine (EBM) can help identify models that offer an appropriate balance between predictive performance and resource consumption (Henderson et al., 2020).

Robustness evaluation examines how model performance changes when data contain missing values, noise, distribution shifts, or other realistic variations. The study should compare performance on the original test set with performance under predefined perturbation scenarios. Changes in macro-F1, ROC-AUC, calibration, and prediction consistency can indicate weaknesses in model reliability. Testing should use realistic scenarios and document limitations rather than assuming that good performance on one test set guarantees robustness in deployment (NIST, 2023).

Fairness should also be evaluated where relevant demographic or operational group information is available and lawful to use. Differences in false-positive rates, false-negative rates, and other context-appropriate measures can reveal whether a model performs unevenly across groups. The assessment should document identified disparities and consider their practical implications before recommending a model for decision support (NIST, 2023).

Overall, the recommended model should be selected using predefined criteria that combine predictive quality, explainability, computational efficiency, robustness, and fairness. Sensitivity analysis should determine whether the ranking changes when the relative importance of these criteria varies. Because actual experimental measurements have not yet been supplied, no numerical improvement or claim of superior sustainability is asserted. The final evaluation should report measured results, test conditions, and methodological limitations transparently.



Chapter Six: Discussion

6.1 Interpretation of Findings

The proposed framework treats predictive performance, explainability, and computational sustainability as interconnected dimensions of AI evaluation. A model that achieves high predictive accuracy may still be unsuitable for consequential decision-making if its predictions are difficult to interpret, its explanations are unstable, or its computational requirements are excessive. This multidimensional perspective is consistent with established explainable AI research and the NIST AI Risk Management Framework (Barredo Arrieta et al., 2020; National Institute of Standards and Technology [NIST], 2023).

The comparative evaluation of Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machine (EBM) is intended to reveal trade-offs between predictive capability and interpretability. Logistic Regression provides a relatively transparent baseline, Random Forest

can model nonlinear relationships, XGBoost offers gradient-boosted prediction, and EBM supports interpretation through its model structure. Their actual relative performance must be established using the same dataset, consistent validation procedures, and an independent test set (Breiman, 2001; Chen & Guestrin, 2016; Lou et al., 2013).

The integration of SHAP and LIME is intended to improve understanding of model predictions by identifying influential features and explaining individual outcomes. SHAP attributes prediction contributions to features, while LIME approximates model behavior locally around a particular instance. However, an understandable explanation does not automatically establish that the model is correct or that the explanation is stable. Consequently, fidelity, stability, and consistency should be assessed alongside predictive performance (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The sustainability dimension broadens the interpretation of model quality beyond predictive metrics. Training time, inference latency, memory consumption, and energy use can reveal whether predictive gains justify additional computational costs. Therefore, the preferred model should be identified using predefined evaluation criteria and sensitivity analysis rather than selecting the model with the highest accuracy alone (Henderson et al., 2020; NIST, 2023).

6.2 Comparison with Existing Studies

The proposed framework builds on three major research directions: predictive machine learning, post-hoc explainability, and responsible AI evaluation. Breiman (2001) established Random Forest as an ensemble learning approach, while Chen and Guestrin (2016) introduced XGBoost for scalable gradient-boosted tree learning. These studies provide methodological foundations for comparing predictive performance, but neither predictive accuracy nor computational scalability alone establishes that a model is sufficiently transparent for consequential decisions.

Lundberg and Lee (2017) introduced SHAP as a unified approach to interpreting model predictions, while Ribeiro et al. (2016) developed LIME to explain individual predictions through local interpretable approximations. The proposed research extends this evaluation perspective by considering explanation quality alongside predictive performance, computational efficiency, and sustainability. However, the two explanation methods should not be assumed to produce identical or universally reliable interpretations because their outputs can depend on model characteristics and feature relationships (Salih et al., 2025).

Recent research on explainable AI for energy systems maintenance emphasizes the need for standardized explanation-evaluation metrics and identifies trade-offs between explanation reliability and suitability for real-time use. This supports the proposed framework's inclusion of explanation fidelity, stability, and inference latency as separate evaluation dimensions (Explainable artificial intelligence for energy systems maintenance, 2025).

Furthermore, the NIST AI Risk Management Framework emphasizes that trustworthy AI requires attention to validity, reliability, transparency, explainability, fairness, and resilience rather than optimizing a single performance metric (Tabassi, 2023).

The principal contribution of the present study is therefore its integrated evaluation design. Instead of treating explainability and sustainability as separate considerations, it proposes assessing them alongside prediction quality and robustness. Whether this integration provides measurable advantages over existing approaches must be demonstrated through controlled experiments using the same dataset, evaluation protocol, and computational environment.

6.3 Practical and Sustainability Implications

The proposed framework has potential applications in public administration, healthcare, education, energy management, financial services, and other settings where AI predictions influence operational or human decisions. SHAP and LIME explanations can help practitioners examine the factors contributing to individual predictions, while performance and calibration metrics provide evidence about model reliability. Nevertheless, explanations should support—not replace—professional judgment, contextual knowledge, and appropriate human oversight (Ribeiro et al., 2016; Tabassi, 2023).

From a computational sustainability perspective, the framework encourages organizations to assess whether improvements in predictive performance justify additional training, inference, and explanation costs. Measuring energy consumption, memory use, and latency can help identify models that meet operational requirements without unnecessary resource expenditure. Standardized reporting of machine learning energy and carbon footprints is particularly relevant to this objective (Henderson et al., 2020).

The approach also supports more responsible procurement and deployment decisions. Organizations could compare candidate systems according to predefined requirements for accuracy, interpretability, robustness, fairness, and computational cost. Sensitivity analysis would help determine whether the selected model remains appropriate when organizational priorities or resource constraints change. Such evaluation is consistent with the risk-management and trustworthiness principles outlined by NIST (Tabassi, 2023).

In resource-constrained environments, including institutions with limited computing infrastructure, evaluating model size, inference latency, and energy consumption may be especially valuable. A computationally efficient model with adequate predictive performance may be more practical than a resource-intensive alternative, provided that it also satisfies the necessary reliability, fairness, and safety requirements. This is a proposed practical implication; its benefits should be confirmed through deployment-specific measurements rather than assumed universally.

6.4 Limitations

Several limitations should be considered when interpreting the proposed framework and any future experimental findings.

First, the framework's conclusions will depend on the selected dataset, prediction target, sample size, feature quality, and class distribution. Results from one dataset may not generalize to other sectors or operating environments. Independent validation on representative data is therefore necessary before recommending deployment (Kohavi, 1995; Tabassi, 2023).

Second, SHAP and LIME have methodological limitations. Their explanations may vary with the underlying model, feature dependence, perturbation strategy, and background data. Consequently, feature importance should not automatically be interpreted as causal influence, and explanations should be tested for stability and fidelity (Lundberg & Lee, 2017; Ribeiro et al., 2016; Salih et al., 2025).

Third, computational sustainability measurements can vary with hardware, software libraries, batch size, model configuration, and workload. Energy estimates may also omit indirect environmental costs, including hardware manufacture and infrastructure operation. Reproducible benchmarking and transparent reporting are required for meaningful comparisons (Henderson et al., 2020).

Fourth, fairness and robustness assessments depend on the availability and quality of appropriate evaluation data. A model may perform well under standard test conditions but degrade when the input distribution changes or when it is applied to underrepresented groups. Testing under realistic perturbations and documenting the limits of generalization are therefore essential (Tabassi, 2023).

Finally, the multi-criteria selection process depends on the chosen metrics, normalization procedures, and criterion weights. Different priorities may produce different model rankings. Sensitivity analysis can reveal this dependence, but it cannot eliminate uncertainty or replace contextual judgment.

Because measured experimental results have not been supplied for this study, the discussion presents a research-based interpretation of the proposed framework rather than claiming demonstrated performance improvements. Future empirical work should implement the complete pipeline, document all evaluation conditions, and validate the recommended model on independent data before drawing definitive conclusions.

Chapter Seven: Conclusion and Future Work

7.1 Summary of Findings

This study developed a proposed evaluation framework for integrating explainable artificial intelligence with intelligent and sustainable decision-making. The framework combines predictive modeling, model explanation, computational resource assessment, robustness evaluation, and multi-criteria model selection to support transparent and responsible AI-assisted decisions (Barredo Arrieta et al., 2020; National Institute of Standards and Technology [NIST], 2023).

The proposed comparative evaluation considers Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machine (EBM). These models represent different trade-offs between predictive capability, interpretability, and computational requirements. Their effectiveness should be assessed using accuracy, precision, recall, macro-F1, ROC-AUC, and probability calibration measures on independent test data (Breiman, 2001; Chen & Guestrin, 2016; Lou et al., 2013).

The framework also integrates SHAP and LIME to explain model predictions and investigate influential features. It proposes evaluating explanation fidelity, stability, and consistency so that explanations are assessed systematically rather than assumed to be reliable merely because they are interpretable (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Finally, the framework incorporates training time, inference latency, memory consumption, energy use, fairness, and robustness into the model evaluation process. These dimensions provide a basis for comparing predictive benefits against computational costs and operational risks. However, the framework's effectiveness and any measurable improvements must be established through empirical experimentation (Henderson et al., 2020; NIST, 2023).

7.2 Research Contributions

The principal contribution of this research is a unified framework for evaluating AI models according to predictive effectiveness, explanation quality, computational sustainability, and responsible decision-making. It positions explainability as an integral part of model assessment rather than an optional reporting feature (Barredo Arrieta et al., 2020; NIST, 2023).

Methodological contribution. The framework proposes a consistent comparison of Logistic Regression, Random Forest, XGBoost, and EBM using predictive metrics, calibration measures, and an independent test set. This design builds on established machine learning and interpretable modeling approaches (Breiman, 2001; Chen & Guestrin, 2016; Lou et al., 2013).

Explainability contribution. The integration of SHAP and LIME provides a structured approach for investigating individual predictions and influential features. The framework also proposes evaluating explanation fidelity and stability, addressing the need for more systematic explanation assessment in applied AI (Lundberg & Lee, 2017; Ribeiro et al., 2016; Shadi et al., 2025).

Sustainability contribution. The proposed evaluation incorporates inference latency, memory consumption, training time, and energy use where measurable. This encourages researchers to consider the computational and environmental costs of AI systems alongside predictive performance (Henderson et al., 2020).

Decision-making contribution. The framework proposes a weighted multi-criteria selection process, supported by sensitivity analysis, to compare models across predictive, interpretability, sustainability, and robustness criteria. Its practical value will depend on empirical validation, justified criterion weights, and assessment within the intended decision context (NIST, 2023).

7.3 Recommendations

Based on the proposed framework and relevant literature, the following recommendations are presented for researchers and organizations implementing explainable AI systems.

1. Adopt multidimensional evaluation. Evaluate predictive performance, calibration, explanation quality, robustness, fairness, and computational cost instead of relying exclusively on accuracy (NIST, 2023).

2. Use reproducible experimental procedures. Document the dataset, preprocessing, train–validation–test split, hyperparameters, random seeds, and hardware configuration. Keep test data independent of model tuning to minimize data leakage (Kohavi, 1995).
3. Validate explanations. Apply SHAP and LIME where appropriate, and assess explanation fidelity, stability, and relevance to the intended users. Avoid interpreting feature contributions as causal relationships without additional causal evidence (Lundberg & Lee, 2017; Ribeiro et al., 2016).
4. Measure computational sustainability. Report training time, inference latency, memory use, and energy consumption where measurement is feasible. Document measurement tools and experimental conditions so that model comparisons are reproducible (Henderson et al., 2020).
5. Apply human oversight and risk management. In high-impact settings, explanations should support expert review, while deployment decisions should include appropriate accountability, privacy, security, and fairness safeguards (NIST, 2023).
6. Use justified multi-criteria weights. Define the evaluation criteria and their weights before final model selection, then perform sensitivity analysis to determine whether reasonable changes in priorities alter the recommended model (NIST, 2023).
7. Validate in the target environment. Test the selected model using representative data and operational conditions before deployment. Monitor changes in performance and data distributions throughout its lifecycle (NIST, 2023).

7.4 Future Research Directions

Future research should extend the proposed framework from a conceptual evaluation design into a fully implemented and empirically validated decision-support system.

1. Empirical validation across multiple datasets. Future studies should evaluate the framework across datasets from healthcare, energy management, agriculture, public administration, and other sustainability-related domains. Cross-domain experiments would help determine which findings generalize and which depend on specific data characteristics (Haghighat et al., 2025; Shadi et al., 2025).
2. Advanced explanation evaluation. Research should compare SHAP and LIME with counterfactual explanations, interpretable surrogate models, and model-intrinsic methods. The evaluation should measure fidelity, stability, computational overhead, and usefulness to different stakeholder groups (Lundberg & Lee, 2017; Ribeiro et al., 2016).
3. Energy-aware model optimization. Future work should investigate methods for reducing inference cost, memory consumption, and energy use while maintaining acceptable predictive quality and explanation reliability. Standardized reporting would improve the comparability of sustainability claims across experiments (Henderson et al., 2020).

4. Dynamic and real-time decision support. Further research should examine how explanations and predictive models behave under changing data distributions, streaming inputs, and operational constraints. This is particularly relevant to energy systems, where explainability must be balanced against the requirements of real-time monitoring and maintenance (Shadi et al., 2025).

5. Human-centered evaluation. Future studies should evaluate whether explanations improve decision quality, user understanding, appropriate trust, and the ability to identify model errors. Such studies should involve intended users and measure practical outcomes rather than treating technical explanation metrics as sufficient evidence of usefulness (NIST, 2023).

6. Fairness, privacy, and governance. The framework should be extended to assess subgroup performance, privacy-preserving model development, accountability, and documentation of limitations. Research should investigate how these requirements interact with predictive quality and computational sustainability (NIST, 2023).

7. Automated multi-objective optimization. Future implementations could combine predictive performance, explanation reliability, fairness, and computational cost in a multi-objective optimization process. Such methods could produce a set of Pareto-efficient candidate models, allowing decision-makers to examine trade-offs rather than relying on one fixed weighted score (Henderson et al., 2020; NIST, 2023).

In conclusion, Explainable AI for Intelligent and Sustainable Decision-Making establishes a research direction for evaluating AI systems through a combination of predictive reliability, transparency, computational efficiency, and responsible governance. The next essential step is to implement the framework, measure its performance under reproducible experimental conditions, and validate its usefulness in real decision-making contexts. Its ultimate contribution should be demonstrated through empirical evidence rather than assumed from the proposed methodology.

References

Abusitta, A., Li, M. Q., & Fung, B. C. M. (2024). Survey on explainable AI: Techniques, challenges and open issues. *Expert Systems with Applications*, 255, Article 124710. <https://doi.org/10.1016/j.eswa.2024.124710>

Barredo Arrieta, A., 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://doi.org/10.1016/j.inffus.2019.12.012>

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>

Hase, P., & Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5540–5552). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.491>

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248), 1–43. <https://jmlr.org/papers/v21/20-312.html>

International Energy Agency. (2025). Energy and AI. <https://www.iea.org/reports/energy-and-ai>

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15. <https://doi.org/10.1145/2382577.2382579>

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1143). Morgan Kaufmann.

Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable artificial intelligence-based decision support systems: A recent review. *Electronics*, 13(14), Article 2842. <https://doi.org/10.3390/electronics13142842>

Lou, Y., Caruana, R., & Gehrke, J. (2013). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 150–158). Association for Computing Machinery. <https://doi.org/10.1145/2487575.2487579>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>

Pineau, J., Vincent-Lamarre, P., Sarka, D., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164), 1–20. <https://jmlr.org/papers/v22/20-303.html>

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/10.1145/2939672.2939778>

Schiller, J., Stiller, S., & Ryo, M. (2025). Artificial intelligence in environmental and Earth system sciences: Explainability and trustworthiness. Artificial Intelligence Review, 58, Article 316. <https://doi.org/10.1007/s10462-025-11165-2>

Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>

United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development. <https://sdgs.un.org/2030agenda>