



Explainable Multidisciplinary Artificial Intelligence for Sustainable Innovation and Intelligent Decision-Making Across Emerging Research Domains

Mulugeta Tilahun Bekele

University of Gondar

College of Informatics

Department of Information Technology

mulugeta.tilahun@uog.edu.et

mulugeta.tilahun@gmail.com

Abstract

Artificial intelligence (AI) is transforming scientific research, environmental sustainability, healthcare, and intelligent infrastructure. However, conventional AI approaches often lack interpretability, cross-domain adaptability, and transparent decision-making. This study proposes an explainable multidisciplinary AI framework integrating predictive analytics, explainable AI (XAI), and multi-criteria decision analysis to support sustainable innovation across diverse research domains.

The proposed framework combines Random Forest, XGBoost, and Explainable Boosting Machines with SHAP-based feature attribution and multi-criteria decision analysis. A comparative evaluation design considers environmental sustainability, healthcare, and intelligent infrastructure as representative application domains. Models are assessed using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUROC), explanation fidelity, computational latency, and decision consistency. Baseline comparisons include conventional machine learning without explanations, explainable models without decision optimization, and the integrated framework. Sustainability indicators, interpretability, computational efficiency, and decision robustness are incorporated into a unified evaluation protocol. Performance is compared using consistent data-splitting procedures, domain-specific metrics, and sensitivity analysis.

The proposed framework establishes a measurable evaluation strategy for determining whether integrated explainability and decision optimization improve predictive reliability and practical decision support. Quantitative outcomes will include differences in predictive performance, explanation fidelity, latency, and decision consistency across models and domains. Qualitative analysis will examine stakeholder interpretability, cross-domain adaptability, transparency, and the relevance of generated recommendations. Numerical improvements will be reported only after empirical experimentation.

This research advances a unified approach to explainable, multidisciplinary AI by connecting predictive intelligence with transparent, sustainability-oriented decision-making. Its comparative design supports reproducible assessment of accuracy, interpretability, efficiency, and robustness. The framework offers a foundation for responsible AI deployment across emerging research domains while highlighting the need for domain-specific validation, representative datasets, and human oversight.

Keywords: Explainable Artificial Intelligence; Multidisciplinary Machine Learning; Sustainable Innovation; Intelligent Decision-Making; SHAP Interpretability.

Chapter One: Introduction



International Journal of Current Trends in Research (IJCTR)

Volume 1 | Issue 1 | 2026 | Article ID: 5639-4217-7632 | <https://doi.org/10.64823/ijctr.2601001>

IORO Publications · Texas, USA · www.ioro.org

1.1. Background of the Study

Artificial intelligence (AI) has become an important driver of scientific discovery, technological innovation, and data-driven decision-making across healthcare, environmental science, agriculture, energy systems, education, and intelligent infrastructure. Machine learning (ML), deep learning, and advanced predictive analytics enable researchers to identify complex patterns, forecast outcomes, and support decisions that would be difficult to achieve through conventional analytical approaches. However, increasing model complexity creates challenges involving transparency, reliability, fairness, accountability, and the practical interpretation of AI-generated predictions (Schwalbe & Finzel, 2024).

Explainable artificial intelligence (XAI) addresses these challenges by providing interpretable representations of model behavior and identifying the factors influencing predictions. Methods such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and interpretable machine learning models can help researchers understand model outputs and communicate findings to decision-makers. Nevertheless, explanations are not automatically reliable or useful; their quality must be evaluated according to the application, intended users, and potential consequences of decisions (Schwalbe & Finzel, 2024).

Recent developments in trustworthy AI emphasize the importance of combining technical performance with transparency, safety, fairness, privacy, robustness, and human oversight. The National Institute of Standards and Technology's AI Risk Management Framework provides guidance for identifying, assessing, and managing AI-related risks throughout the system lifecycle (Tabassi, 2023). Similarly, the European Union's Artificial Intelligence Act establishes a risk-based regulatory framework for AI development and deployment, strengthening the importance of appropriate transparency and safeguards in relevant applications.

Sustainable innovation further requires AI systems to support environmental protection, efficient resource utilization, social well-being, and economically viable solutions. In environmental science, for example, trustworthy AI requires attention to contextual uncertainty, stakeholder participation, and the social implications of model-supported decisions (Bostrom et al., 2024). These considerations are also relevant to healthcare, precision agriculture, renewable energy, and public services.

Despite progress in AI and XAI, a key research challenge is determining how predictive accuracy, explanation quality, computational efficiency, sustainability outcomes, and decision consistency can be evaluated within a unified multidisciplinary framework. This study therefore proposes an Explainable Multidisciplinary AI Framework that integrates predictive modeling, explainability techniques, sustainability indicators, and multi-criteria decision analysis. Its purpose is to investigate whether this integration can support transparent, robust, and context-sensitive decisions across selected emerging research domains.

1.2. Problem Statement

Although AI models increasingly support complex decisions across multiple disciplines, several challenges limit their responsible and effective application. First, high-performing machine learning models may produce predictions without sufficiently understandable explanations, making it difficult for researchers and stakeholders to assess the evidence underlying individual decisions. Second, existing XAI techniques often focus on explaining model predictions rather than integrating explanations with sustainability indicators, decision priorities, and practical implementation constraints (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Third, different research domains employ heterogeneous datasets, evaluation metrics, and decision criteria. A model suitable for environmental prediction may not be directly transferable to healthcare or intelligent infrastructure without domain-specific adaptation and validation. Consequently, comparing AI systems using

predictive accuracy alone may overlook critical factors such as fairness, robustness, interpretability, computational cost, and environmental impact.

Furthermore, responsible AI frameworks emphasize risk management and accountability, but translating these principles into measurable, cross-domain evaluation procedures remains a methodological challenge (Tabassi, 2023). Research on trustworthy AI also identifies the importance of interdisciplinary collaboration, contextual evaluation, and stakeholder involvement (Bostrom et al., 2024).

The central problem addressed by this study is the absence of a systematically evaluated, integrated framework for comparing predictive performance, explanation quality, sustainability, and decision-making effectiveness across selected research domains.

The study will investigate whether combining machine learning, SHAP/LIME explanations, sustainability indicators, and multi-criteria decision analysis can improve the transparency and consistency of AI-assisted decisions without introducing unacceptable losses in predictive performance or computational efficiency. The proposed benefits will be evaluated empirically rather than assumed in advance.

1.3. Research Questions

The study will address the following research questions:

RQ1: How can explainable AI, predictive modeling, and multi-criteria decision analysis be integrated into a unified multidisciplinary framework?

RQ2: How does the proposed framework compare with conventional machine learning models in accuracy, precision, recall, F1-score, and AUROC across selected research domains?

RQ3: To what extent do SHAP and LIME improve the interpretability, explanation fidelity, and consistency of model-generated predictions?

RQ4: How does integrating sustainability indicators influence the ranking and selection of AI-supported decisions?

RQ5: What trade-offs arise among predictive performance, interpretability, computational latency, robustness, and decision consistency?

RQ6: How effectively does the framework generalize across domains, and what data, ethical, and implementation limitations affect its practical applicability?

1.4. Research Objectives

1.4.1. General Objective

To develop and empirically evaluate an explainable multidisciplinary artificial intelligence framework that integrates predictive analytics, model interpretability, sustainability assessment, and intelligent decision-making across selected emerging research domains.

1.4.2. Specific Objectives

To review existing machine learning, explainable AI, sustainable innovation, and multi-criteria decision-making approaches.

To design a unified framework integrating predictive models, SHAP/LIME explanations, sustainability indicators, and decision optimization.

To compare Random Forest, XGBoost, and interpretable baseline models using standardized experimental procedures.

To evaluate predictive performance, explanation fidelity, explanation stability, computational latency, and robustness.

To investigate the contribution of environmental, economic, and social indicators to evidence-based decision prioritization.

To assess cross-domain adaptability using selected applications in environmental sustainability, healthcare, and intelligent infrastructure.

To formulate recommendations for transparent, responsible, and context-sensitive AI deployment.

1.5. Significance of the Study

The study is expected to contribute to AI research and practical decision-making in several ways.

Theoretical significance: It will develop a conceptual connection between explainable machine learning, multidisciplinary knowledge, sustainability assessment, and multi-criteria decision analysis.

Methodological significance: It will establish a comparative evaluation protocol incorporating predictive performance, interpretability, robustness, computational efficiency, and decision consistency. This approach will help researchers identify trade-offs that single-metric evaluations may overlook.

Technological significance: The proposed framework will investigate how SHAP and LIME explanations can be incorporated into predictive workflows to provide understandable evidence for AI-assisted recommendations. These methods will be evaluated for their suitability rather than assumed to produce universally reliable explanations.

Environmental and societal significance: By incorporating relevant environmental, economic, and social indicators, the framework may help decision-makers compare alternatives in areas such as resource management, sustainable agriculture, healthcare planning, and infrastructure development.

Practical significance: Researchers, policymakers, institutions, and technology developers may use the evaluation framework to select AI models that balance predictive effectiveness with transparency and operational requirements.

Academic significance: The study may provide a foundation for subsequent research in trustworthy AI, sustainable digital transformation, and cross-disciplinary intelligent systems. Its contributions will depend on empirical validation and the reproducibility of the results.

1.6. Scope and Limitations

1.6.1. Scope of the Study

The research focuses on the design and empirical evaluation of an explainable multidisciplinary AI framework. Three representative application domains will be considered:

- **Environmental sustainability:** environmental monitoring, resource optimization, and sustainability-related prediction.
- **Healthcare:** predictive analytics using appropriately governed health datasets, with emphasis on interpretability and decision support rather than autonomous clinical diagnosis.

- Intelligent infrastructure: predictive maintenance, infrastructure monitoring, and resource-efficient system management.

The technical scope includes Random Forest, XGBoost, interpretable baseline models, SHAP, LIME, and multi-criteria decision analysis. Predictive performance will be assessed using accuracy, precision, recall, F1-score, and AUROC where appropriate. Additional measures will include explanation fidelity, explanation stability, inference latency, robustness under data perturbations, and decision consistency.

Sustainability assessment will consider relevant environmental, economic, and social indicators according to the needs of each domain. Cross-domain evaluation will use domain-appropriate datasets and metrics, with consistent experimental principles and transparent reporting.

1.6.2. Limitations of the Study

Data availability and quality: Differences in data completeness, labeling, representativeness, and measurement standards may affect comparisons across domains.

Domain heterogeneity: Environmental, healthcare, and infrastructure datasets differ in structure and objectives. Consequently, direct comparisons of predictive scores may be misleading unless task difficulty, class distribution, and evaluation protocols are considered.

Explanation limitations: SHAP and LIME provide useful interpretability tools but do not automatically establish causal relationships, eliminate bias, or guarantee that explanations reflect the complete reasoning process of a model.

Computational constraints: Complex models and repeated explanation calculations may increase computational cost, particularly for large datasets or resource-constrained environments.

Sustainability measurement: Environmental, economic, and social outcomes may involve conflicting priorities, uncertain measurements, and stakeholder-dependent weights.

Generalizability: Findings from selected datasets and application domains may not extend to every discipline, institution, or deployment environment.

Empirical validation: The framework's effectiveness cannot be established solely through conceptual design. Reliable conclusions require suitable datasets, baseline comparisons, repeated experiments, statistical testing, and transparent reporting. No numerical performance improvements are claimed before such evaluation.

Chapter Two: Literature Review

1.7. Fundamentals of Artificial Intelligence

Artificial intelligence (AI) refers to computational systems designed to perform tasks associated with learning, prediction, reasoning, perception, language processing, and decision support. Contemporary AI encompasses machine learning (ML), deep learning, knowledge-based systems, optimization algorithms, and generative models. These approaches differ in their data requirements, computational complexity, learning mechanisms, and suitability for specific applications.

Machine learning enables computational models to identify relationships in data and use learned patterns to predict outcomes or classify observations. Supervised learning uses labeled examples, unsupervised learning identifies structures in unlabeled data, and reinforcement learning learns decision strategies through interaction with an environment. Deep learning employs multilayer neural networks to represent complex relationships in images, text, signals, and other high-dimensional data. Although these methods offer

substantial analytical capabilities, their performance depends on data quality, model selection, validation procedures, and the relationship between training and deployment environments.

1.7.1. Machine Learning Algorithms

This research considers Random Forest, XGBoost, and interpretable baseline models as candidate predictive techniques.

Random Forest constructs an ensemble of decision trees using randomized samples and feature subsets. Combining predictions across multiple trees can reduce the variance associated with individual trees and provide a useful baseline for structured datasets (Breiman, 2001).

XGBoost implements gradient-boosted decision trees using regularization and computational optimization. It can model nonlinear relationships and interactions among variables, making it suitable for many tabular prediction tasks (Chen & Guestrin, 2016).

Interpretable models, including logistic regression and Explainable Boosting Machines, provide complementary baselines. Logistic regression offers relatively straightforward coefficient-based interpretation for appropriate classification tasks, while Explainable Boosting Machines model feature effects using an additive structure that can capture nonlinear relationships (Lou et al., 2013).

The proposed research will compare these methods rather than assume that one algorithm performs best across every domain.

1.7.2. Model Evaluation and Generalization

AI performance must be assessed using metrics appropriate to the prediction task. Accuracy measures the proportion of correct classifications, while precision and recall assess the reliability and completeness of positive predictions. The F1-score combines precision and recall through their harmonic mean. The area under the receiver operating characteristic curve (AUROC) measures the model's ability to discriminate between classes across classification thresholds.

For regression problems, appropriate metrics include mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Where decisions depend on predicted probabilities, calibration should also be evaluated.

Reliable assessment requires separation of training, validation, and test data, leakage prevention, appropriate cross-validation, and uncertainty analysis. When data are temporally or geographically dependent, random splitting alone may overestimate generalization. Accordingly, this study will use domain-appropriate validation procedures and consistent comparison protocols.

1.7.3. Trustworthiness and Responsible AI

Predictive performance alone does not establish that an AI system is suitable for deployment. Trustworthy AI also involves fairness, transparency, privacy, security, robustness, accountability, and human oversight. The National Institute of Standards and Technology's AI Risk Management Framework provides a lifecycle-oriented approach to identifying and managing such risks (Tabassi, 2023). The OECD's updated AI Principles also emphasize trustworthy AI, human rights, transparency, robustness, and sustainable development (Organisation for Economic Co-operation and Development [OECD], 2024).

These principles motivate the proposed study's broader evaluation approach, in which predictive effectiveness is assessed alongside explanation quality, computational requirements, and sustainability-related decision criteria.

1.8. Multidisciplinary AI and Emerging Research Domains

Multidisciplinary AI applies computational intelligence to problems that cross traditional academic and professional boundaries. Rather than treating AI as an isolated technology, this perspective combines machine learning with domain-specific knowledge, data infrastructures, scientific modeling, and stakeholder requirements.

The relevance of multidisciplinary AI is increasing as environmental systems, public health, agriculture, energy, and infrastructure become more data-intensive and interconnected. A systematic review by Greif et al. (2025) examines how AI techniques are applied to the United Nations Sustainable Development Goals (SDGs), highlighting the growing intersection between AI research and sustainability challenges.

1.8.1. AI in Environmental Sustainability

Environmental applications include climate-risk assessment, pollution monitoring, water-resource management, biodiversity conservation, and renewable-energy forecasting. Machine learning can identify patterns in satellite imagery, sensor measurements, and historical environmental observations.

However, environmental decisions must account for geographic variability, temporal change, uncertainty, and potential consequences for affected communities. A model trained in one ecosystem may not generalize reliably to another. Therefore, multidisciplinary environmental AI should combine predictive models with geographical context, scientific knowledge, and transparent uncertainty assessment.

1.8.2. AI in Healthcare

Healthcare applications include disease-risk prediction, medical-image analysis, patient monitoring, and resource planning. AI can help identify patterns in clinical and biomedical data, but differences in patient populations, data collection practices, and healthcare resources can influence model reliability.

Explainability is particularly relevant when healthcare professionals need to examine model recommendations, identify potential data bias, and understand uncertainty. The proposed framework will treat healthcare AI as a decision-support application requiring privacy protection, domain-specific validation, and appropriate professional oversight rather than as a substitute for clinical judgment.

1.8.3. AI in Agriculture and Food Systems

Precision agriculture combines machine learning with remote sensing, geographic information systems, weather observations, and agricultural knowledge. Potential applications include crop-stress detection, yield prediction, irrigation optimization, and pest-risk assessment.

These systems can support resource efficiency, but their practical value depends on local data availability, affordability, environmental conditions, and accessibility for agricultural stakeholders. Explanations of predictions can help users understand which observed variables influence recommendations, although the explanations must be validated and presented in an appropriate context.

1.8.4. AI in Intelligent Infrastructure

Intelligent infrastructure includes energy networks, transportation systems, buildings, communication networks, and public-service platforms. AI can support fault detection, predictive maintenance, demand forecasting, and operational optimization.

These applications often involve interacting physical, digital, and human systems. Consequently, an effective evaluation must consider not only predictive accuracy but also inference latency, robustness, interoperability, and the consequences of incorrect recommendations. Pearson et al. (2024) similarly

emphasize the need to evaluate AI applications in real-world sustainability policy settings rather than relying solely on anticipated benefits.

1.8.5. Cross-Domain Integration

Although these research domains differ, they share methodological requirements: reliable data, suitable predictive models, interpretable outputs, transparent evaluation, and decisions aligned with domain-specific objectives.

The proposed research will investigate whether a common framework can organize these requirements while preserving domain-specific validation. It will not assume that identical models, explanations, or sustainability weights are appropriate for all applications.

1.9. Explainable Artificial Intelligence (XAI)

Explainable artificial intelligence refers to methods and techniques that make the behavior, predictions, or recommendations of AI systems more understandable to humans. XAI has become an important component of responsible AI because decision-makers need more than predictive outputs when evaluating systems used in high-impact contexts. Kowald et al. (2024) discuss trustworthy AI through dimensions including transparency, fairness, robustness, privacy, security, and accountability. These dimensions indicate that explainability should be considered as part of a broader trustworthiness strategy rather than as an isolated technical feature.

1.9.1. Model-Specific and Model-Agnostic Explainability

XAI methods can be broadly classified into model-specific and model-agnostic approaches.

Model-specific methods exploit the internal structure of particular model families. For example, decision trees can be inspected through their decision paths, while convolutional neural networks can be examined using gradient-based visual explanations. These methods may provide detailed information about specific model architectures, but their applicability depends on the model being evaluated.

Model-agnostic methods can be applied to different predictive models through their input-output behavior. Two prominent examples are SHAP and LIME.

LIME approximates the behavior of a complex model near an individual prediction using a simpler interpretable model. Its original formulation supports local explanations for classifiers and other predictive systems (Ribeiro et al., 2016).

SHAP uses concepts from cooperative game theory to estimate the contribution of individual features to a prediction. Lundberg and Lee (2017) proposed a unified framework for interpreting model predictions, providing a theoretically grounded approach to feature attribution.

In the proposed research, both methods will be evaluated according to their suitability for the selected datasets and models. Their explanations will not be treated as causal evidence unless supported by appropriate causal analysis.

1.9.2. Local and Global Explanations

Local explanations describe why a model generated a particular prediction for an individual observation. For example, a local explanation might identify which weather variables contributed most to a crop-stress prediction.

Global explanations summarize patterns across a model or dataset. They may identify variables that commonly influence environmental forecasts, infrastructure fault detection, or healthcare risk predictions.

Both perspectives are necessary for multidisciplinary AI. Global explanations can support scientific interpretation and model auditing, whereas local explanations can help users investigate individual recommendations. However, feature importance may change across populations, geographic areas, or operating conditions, making context-specific evaluation essential.

1.9.3. Evaluating Explanation Quality

An explanation should not be considered reliable simply because it is visually understandable. Explanation evaluation should consider several measurable properties:

- **Fidelity:** how accurately an explanation represents the behavior of the model it describes.
- **Stability:** whether similar inputs produce reasonably consistent explanations.
- **Consistency:** whether explanations align with comparable model behavior across relevant observations.
- **Comprehensibility:** whether intended users can understand the explanation.
- **Computational efficiency:** the time and resources required to generate explanations.

These properties can conflict. A simple explanation may be easier to understand but fail to represent important nonlinear relationships. A detailed explanation may be faithful to a model yet difficult for a nontechnical user to interpret.

Kowald et al. (2024) emphasize that trustworthy AI involves multiple interconnected requirements and unresolved evaluation challenges. Consequently, the proposed framework will evaluate explanation quality alongside predictive performance, robustness, and computational cost rather than treating explainability as a binary property.

1.9.4. Limitations of Existing XAI Approaches

Despite their usefulness, SHAP and LIME have limitations. Results may depend on feature dependence, the choice of background data, perturbation procedures, model behavior, and the definition of the prediction being explained. Feature attribution does not automatically demonstrate that a variable caused an outcome, nor does it establish that a model is fair or accurate.

Furthermore, an explanation can be technically faithful to a biased or poorly calibrated model. Therefore, explanation quality must be assessed together with data quality, model validation, fairness, and uncertainty.

The proposed study will address these limitations by comparing explanations across models, examining their stability under controlled perturbations, and reporting predictive and interpretability metrics separately.

1.10. AI for Sustainable Innovation

Sustainable innovation involves developing and applying technologies, processes, and organizational practices that create economic and social value while reducing environmental harm and supporting long-term resource viability. AI can contribute to this objective by improving prediction, resource allocation, operational efficiency, and the identification of complex environmental and social patterns.

However, sustainability cannot be reduced to a single performance metric. A solution that improves operational efficiency may increase energy consumption elsewhere, impose high implementation costs, or

distribute benefits unequally. Consequently, sustainable AI requires evaluation of both the benefits of AI applications and the resources and risks associated with developing and operating them.

1.10.1. AI and the Sustainable Development Goals

The United Nations Sustainable Development Goals provide a broad framework for connecting technology with environmental protection, economic development, health, education, and social equity.

Greif et al. (2025) reviewed AI techniques used in research related to the SDGs and identified a broad range of applications, particularly in environmental sustainability. Their review demonstrates the growing use of AI in sustainability-related research while highlighting the need to develop research directions that better address complex sustainability challenges.

Gohr et al. (2025) examined 792 articles on AI and sustainable development research. Their analysis found that relatively few studies combine advanced AI methods with deep sustainability expertise. The authors identify the need to balance contextual understanding with generalizability to support meaningful sustainable development outcomes.

These findings are relevant to the present study because they suggest that AI research should not only optimize technical predictions but also connect model outputs with the contextual requirements of sustainability problems.

1.10.2. Environmental, Economic, and Social Dimensions

The proposed framework will represent sustainability through three complementary dimensions.

Environmental sustainability includes energy consumption, emissions, water use, waste generation, and resource efficiency, depending on the selected application.

Economic sustainability includes implementation costs, operational expenses, resource productivity, and the feasibility of maintaining an AI-supported solution over time.

Social sustainability includes accessibility, fairness, stakeholder impact, and the distribution of benefits and risks.

Not every indicator will apply to every research domain. The study will define domain-specific indicators and document their measurement, direction of preference, normalization, and weighting. Conflicting objectives will be addressed through multi-criteria decision analysis rather than combined without justification.

1.10.3. Sustainable AI Versus AI for Sustainability

Two concepts must be distinguished.

AI for sustainability refers to using AI to address sustainability challenges, such as forecasting energy demand or improving irrigation efficiency.

Sustainable AI refers to developing and operating AI systems in ways that consider their own environmental, economic, and social impacts. This includes computational energy consumption, hardware requirements, data governance, maintenance, and accessibility.

The distinction is important because an AI system may improve a sustainability-related outcome while requiring substantial computational resources. Both dimensions should therefore be considered when assessing the overall value of an AI-supported decision.

The OECD's updated AI Principles explicitly recognize sustainable development alongside transparency, accountability, safety, and human-centered values (OECD, 2024).

1.10.4. Integrating Explainability with Sustainability

Explainability can help stakeholders investigate why a model recommends a particular intervention and which factors influence the prediction. In environmental monitoring, for example, feature attribution may help analysts identify variables associated with elevated risk. In energy systems, explanations may clarify which inputs contribute to a demand forecast.

Nevertheless, identifying influential features does not establish that following a recommendation will produce a sustainable outcome. Sustainability requires additional evidence, including outcome measurement, operational constraints, uncertainty assessment, and consideration of indirect effects.

The proposed study will integrate XAI with sustainability indicators and multi-criteria decision analysis to examine how transparent predictions can inform the ranking of alternative decisions. The expected contribution is a testable methodology, not an assumption that explainability automatically improves sustainability.

1.11. Intelligent Decision-Making Approaches

Intelligent decision-making refers to the systematic use of data, analytical models, domain knowledge, and decision criteria to select actions under conditions of uncertainty and competing objectives. AI-supported decision-making combines prediction with the evaluation of possible alternatives, enabling decision-makers to move from estimating outcomes to identifying appropriate actions.

In multidisciplinary research, predictive accuracy alone is insufficient to determine which decision should be selected. A model may accurately predict an outcome without considering implementation costs, environmental consequences, fairness, or stakeholder priorities. Therefore, the proposed framework separates predictive modeling from decision evaluation and integrates them through an explicit decision-support process.

1.11.1. Multi-Criteria Decision Analysis

Multi-criteria decision analysis (MCDA) provides structured methods for comparing alternatives using several criteria. It is particularly relevant when decisions involve conflicting objectives, such as maximizing predictive reliability while minimizing operational cost and environmental impact.

The analytic hierarchy process (AHP) can be used to derive criterion weights through pairwise comparisons, while the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) ranks alternatives according to their distances from positive and negative ideal solutions (Hwang & Yoon, 1981; Saaty, 1980).

For the proposed framework, MCDA will integrate selected criteria such as predictive performance, explanation quality, computational latency, robustness, environmental impact, and economic feasibility. The criteria will be normalized, weighted, and combined according to a documented decision protocol.

A general weighted decision score can be represented as:

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

where S_i represents the score of alternative i , w_j is the weight assigned to criterion j , and z_{ij} is the normalized performance of alternative i on criterion j . The weights satisfy:

$$\sum_{j=1}^m w_j = 1, \quad w_j \geq 0$$

1.11.2. AI-Driven Predictive Decision Support

Predictive decision support involves using AI models to estimate possible outcomes and then evaluating those estimates against decision objectives. In the proposed framework, Random Forest, XGBoost, and interpretable models will generate predictions, while SHAP and LIME will provide explanations of selected predictions.

The resulting outputs will be combined with relevant sustainability indicators and decision criteria. This approach distinguishes three analytical stages:

1. Prediction: estimate outcomes from observed data.
2. Explanation: investigate which features contribute to model predictions.
3. Decision evaluation: compare alternative actions using defined criteria and stakeholder priorities.

This separation is important because a feature that strongly influences a prediction is not necessarily the factor that should receive the highest priority in a policy or operational decision.

1.11.3. Uncertainty, Robustness, and Human Oversight

Decision-making under uncertainty requires consideration of model error, data variability, incomplete observations, and changing operating conditions. Robust decision support should examine whether rankings remain stable when inputs, model parameters, or criterion weights change.

Kowald et al. (2024) identify robustness, transparency, fairness, privacy, and accountability as interconnected dimensions of trustworthy AI. The OECD (2024) similarly emphasizes human-centered values, transparency, safety, and accountability throughout the AI lifecycle.

Accordingly, the proposed research will use sensitivity analysis and appropriate robustness tests to assess whether alternative assumptions substantially change the resulting decisions. Human oversight will remain necessary when decisions involve high consequences, contested values, or limitations that the model cannot adequately represent.

1.11.4. Comparative Evaluation of Decision-Making Approaches

The proposed evaluation will compare three approaches:

- Conventional predictive AI: selects or informs decisions primarily through predictive outputs and predefined rules.
- Explainable AI: adds feature-level explanations to predictive outputs without necessarily optimizing decisions across multiple criteria.
- Integrated explainable decision support: combines prediction, explanation, sustainability indicators, and MCDA.

The comparison will assess predictive performance, explanation fidelity, computational latency, decision consistency, and sensitivity to criterion weights. This design will help determine whether the integrated approach provides measurable benefits and what trade-offs it introduces.

1.12. Research Gaps and Contributions

The reviewed literature establishes that AI offers significant analytical capabilities, explainability provides methods for understanding model behavior, sustainability research requires contextual evaluation, and MCDA supports decisions involving competing objectives. However, integrating these elements into a reproducible multidisciplinary evaluation framework remains an important research challenge.

Recent literature on AI and sustainable development identifies the need for stronger connections between technical AI research and substantive sustainability expertise (Gohr et al., 2025). Research on trustworthy AI also highlights persistent challenges involving evaluation, contextual dependency, interdisciplinary collaboration, and real-world deployment (Kowald et al., 2024).

1.13. Identified Research Gaps

The literature review identifies several methodological and practical gaps in integrating explainable artificial intelligence (XAI), multidisciplinary machine learning, sustainability assessment, and intelligent decision-making. Table 2.1 summarizes the key gaps, their implications, and the corresponding research needs.

			Table 2.1. Identified Research Gaps in Explainable Multidisciplinary AI

No.

Identified research gap

Limitations of existing approaches

Research need

1

Fragmented evaluation criteria

Predictive accuracy, explainability, computational efficiency, and sustainability are often evaluated separately.

Integrate multiple evaluation dimensions into a unified assessment framework.

2

Limited integration of XAI and decision optimization

SHAP and LIME explain predictions but do not independently determine optimal decisions.

Combine explanation methods with multi-criteria decision analysis (MCDA).

3

Cross-domain comparability

Different domains use heterogeneous datasets, objectives, and evaluation metrics.

Establish consistent evaluation principles with domain-specific validation.

4

Explanation reliability

Explanations may be unstable, insufficiently faithful, or sensitive to input changes.

Evaluate explanation fidelity, stability, and consistency.

5

Incomplete sustainability assessment

Environmental, economic, and social indicators are not always jointly considered.

Incorporate measurable, domain-relevant sustainability indicators.

6

Limited trade-off analysis

Improvements in predictive performance may increase computational cost or reduce interpretability.

Analyze trade-offs among accuracy, explanation quality, latency, and robustness.

7

Decision uncertainty and sensitivity

Rankings may change when data, model outputs, or criterion weights change.

Apply sensitivity analysis, uncertainty assessment, and robustness testing.

8

Limited practical accountability

Technically effective systems may not adequately support stakeholder understanding or human oversight.

Integrate transparent reporting, domain expertise, and human review into decision support.

1.14. Proposed Research Contributions

The proposed study, Explainable Multidisciplinary Artificial Intelligence for Sustainable Innovation and Intelligent Decision-Making Across Emerging Research Domains, aims to address the identified gaps through methodological, technical, and practical contributions.

			Table 2.2. Proposed Research Contributions and Evaluation Parameters

No.

Proposed contribution

Description

Evaluation parameters

1

Integrated multidisciplinary AI framework

Connects predictive machine learning, explainability, sustainability assessment, and decision support.

Framework completeness, reproducibility, and cross-domain applicability.

2

Comparative predictive modeling

Compares Random Forest, XGBoost, and interpretable baseline models.

Accuracy, precision, recall, F1-score, and AUROC, where appropriate.

3

Explainability assessment

Integrates SHAP and LIME to investigate model behavior.

Explanation fidelity, stability, consistency, and computation time.

4

Sustainability-aware decision analysis

Incorporates environmental, economic, and social criteria into alternative ranking.

Decision scores, criterion weights, ranking stability, and sensitivity to weights.

5

Multidimensional performance evaluation

Assesses predictive quality alongside computational and reliability considerations.

Latency, robustness, resource consumption, and calibration where applicable.

6

Cross-domain evaluation protocol

Applies consistent experimental principles to environmental, healthcare, and intelligent infrastructure datasets.

Domain-specific performance, generalization, and between-domain variation.

7

Responsible decision-support methodology

Incorporates uncertainty reporting, human oversight, and transparent decision criteria.

Uncertainty measures, fairness assessments where appropriate, and decision traceability.

8

Reproducible research methodology

Documents datasets, preprocessing, model settings, evaluation procedures, and limitations.

Reproducibility, statistical uncertainty, and completeness of reporting.

The principal contribution is a unified, explainable, sustainability-aware AI evaluation framework that examines the relationship between predictive effectiveness, explanation reliability, computational efficiency, and intelligent decision-making across selected research domains.

The framework will be evaluated against conventional predictive AI and explainable AI without integrated decision optimization. Its effectiveness will be determined through empirical experiments, appropriate statistical comparisons, and sensitivity analysis. Numerical improvements will be reported only when supported by experimental results.

1.14.1. Conceptual Framework of the Proposed Research

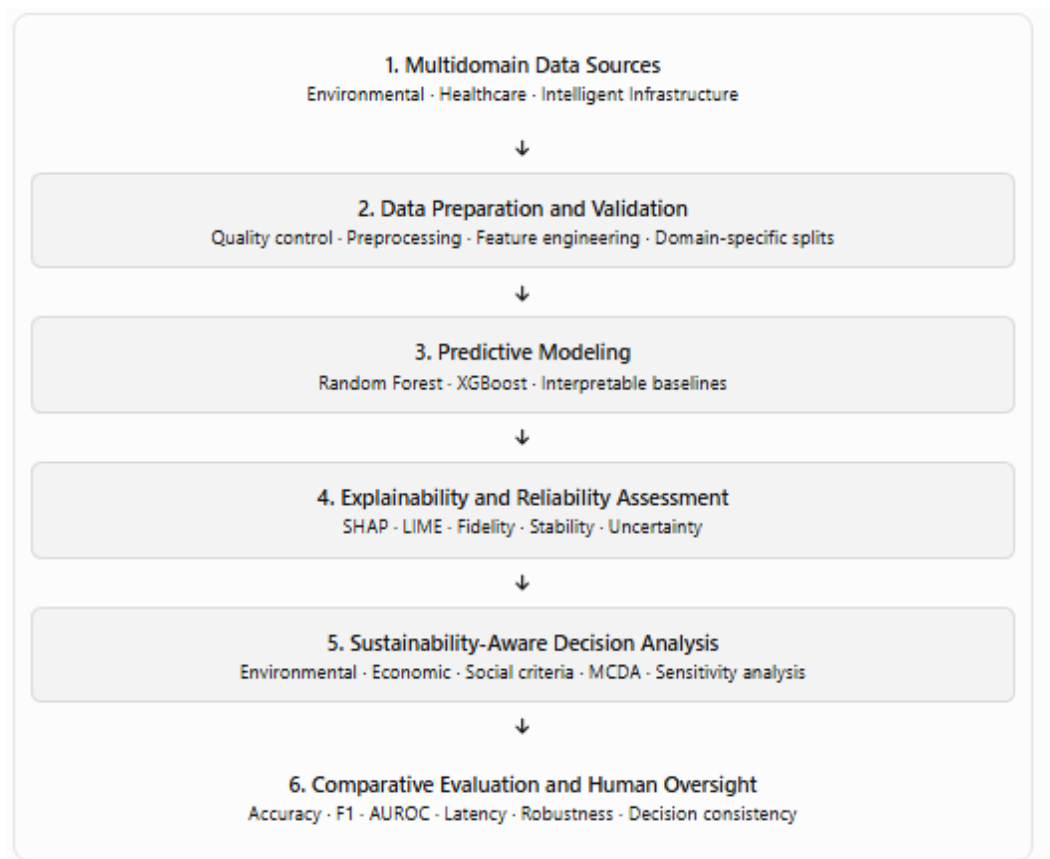


Figure 2.1. Proposed conceptual framework

Chapter Three: Research Methodology

This Section presents the research design, data collection and preprocessing procedures, proposed framework, machine learning algorithms, explainability methods, multi-criteria decision analysis, and experimental validation strategy. The methodology is designed to investigate whether integrating explainable AI with sustainability assessment and decision optimization can improve the transparency, reliability, and practical usefulness of AI-assisted decisions across selected research domains.

The methodological foundation combines established machine learning algorithms, post-hoc explanation techniques, and decision-analysis methods. Recent research emphasizes that SHAP and LIME must be interpreted carefully because their outputs can depend on the fitted model, correlated features, and explanation settings (Salih et al., 2025). Research on sustainable building design also demonstrates the feasibility of integrating explainability with multi-objective optimization, while highlighting the importance of application-specific validation.

1.15. Research Design

The study will adopt a quantitative, comparative, experimental research design, supplemented by qualitative interpretation of model explanations and decision implications. The quantitative component will evaluate predictive performance, explanation fidelity, computational efficiency, and decision-ranking stability. The qualitative component will assess whether explanations and recommendations are understandable and relevant to the intended research domain.

The study will investigate three application domains:

1. Environmental sustainability: environmental monitoring, resource management, and sustainability-related prediction.
2. Healthcare: predictive analytics using appropriately governed datasets for research and decision support.
3. Intelligent infrastructure: infrastructure monitoring, fault prediction, and resource-efficient system management.

These domains represent different data structures and decision requirements. The research will not assume that a single model or explanation technique is optimal for all three.

Three configurations will be compared:

- Baseline AI: conventional predictive models without the proposed integrated decision-analysis layer.
- Explainable AI: predictive models supplemented with SHAP and LIME explanations.
- Integrated framework: predictive modeling, explanation assessment, sustainability indicators, and multi-criteria decision analysis combined in one workflow.

The experimental comparisons will examine predictive quality, explanation reliability, latency, robustness, and decision consistency. Because the configurations differ in functionality, improvements in decision ranking will be assessed separately from predictive accuracy.

The research will follow documented data-selection, model-training, validation, and reporting procedures. Where appropriate, statistical tests and confidence intervals will be used to quantify uncertainty. The study

will distinguish observed experimental findings from anticipated benefits, consistent with the risk-management principles described by Tabassi (2023).

1.16. Data Collection and Preprocessing

1.16.1. Data Collection

The study will use secondary datasets obtained from publicly accessible research repositories or authorized institutional sources. Datasets will be selected according to their relevance to the research questions, data quality, licensing conditions, availability of target variables, and suitability for reproducible experiments.

Potential data sources include:

- Environmental monitoring, weather, land-use, and resource-consumption datasets.
- De-identified healthcare datasets containing appropriate clinical or biomedical variables.
- Infrastructure sensor, equipment-monitoring, energy-consumption, or fault-detection datasets.

The final datasets will be selected and documented before model evaluation. No dataset size, experimental result, or performance improvement will be claimed until the corresponding data have been obtained and analyzed.

For each dataset, the research will record the source, observation period, sample size, feature descriptions, target variables, missingness, class distribution, collection methodology, and relevant limitations. Data licensing and any required ethical approval will be checked before use.

1.16.2. Data Quality Assessment

Data quality assessment will examine missing values, duplicate records, inconsistent formats, invalid observations, outliers, class imbalance, and possible data leakage.

Missing values will be handled using suitable imputation methods fitted only on training data. Categorical variables will be encoded as appropriate, while numerical features will be scaled when required by the selected model. Outliers will be investigated to distinguish data errors from legitimate extreme observations.

Feature selection and preprocessing decisions will be made using training data only. Information derived from the test set will not be used to select features, tune hyperparameters, or determine the final model.

1.16.3. Data Partitioning

For independent and sufficiently large datasets, an initial partition of 70% for training, 15% for validation, and 15% for testing may be used. The final partition will depend on dataset size, class balance, and the data-generating process.

Stratified partitioning will be considered for classification tasks with imbalanced outcomes. For time-series data, chronological splits will be used where appropriate. For geographic, patient-grouped, or repeated-measurement data, group-aware validation will prevent related observations from appearing in both training and test sets.

Cross-validation will be used for model selection when suitable, while the final test set will be reserved for evaluating the selected model. This approach will help reduce overly optimistic performance estimates.

1.16.4. Feature Engineering

Feature engineering will be specific to each application domain. Environmental datasets may require temporal summaries, seasonal indicators, or spatial attributes. Healthcare datasets may require clinically

meaningful transformations that preserve privacy and avoid target leakage. Infrastructure datasets may require rolling statistics, equipment-state indicators, or lagged measurements.

All engineered features will be documented, and transformations will be applied consistently during training and evaluation. Feature engineering will prioritize domain relevance and reproducibility rather than maximizing model performance through undocumented experimentation.

1.16.5. Data Governance and Ethical Considerations

The research will comply with applicable data-use agreements, institutional policies, and privacy requirements. Sensitive data will be de-identified where appropriate, and access will be restricted to authorized purposes. Potential dataset biases and limitations in population or geographic representation will be reported.

Healthcare-related results will be interpreted as research findings rather than direct clinical recommendations. Human oversight will remain necessary for consequential decisions.

1.17. Proposed Research Framework

The proposed Explainable Multidisciplinary Artificial Intelligence (EMAI) framework integrates predictive modeling, model interpretation, sustainability assessment, and decision optimization. Its central purpose is to provide a structured method for evaluating how AI-generated predictions can support transparent and context-sensitive decisions across multiple domains.

The framework consists of six interconnected layers.

1.17.1. Data Integration Layer

The data integration layer collects and organizes data from environmental sustainability, healthcare, and intelligent infrastructure applications. It includes data validation, preprocessing, feature engineering, and domain-specific metadata. Each dataset will retain its own target variables and evaluation requirements.

1.17.2. Predictive Intelligence Layer

The predictive layer trains Random Forest, XGBoost, and interpretable baseline models. Models will be selected according to the characteristics of each task and compared using the same data partitions and appropriate evaluation metrics.

1.17.3. Explainability Layer

The explainability layer applies SHAP and LIME to investigate feature contributions to individual predictions and broader model behavior. SHAP will support local and global feature-attribution analysis, while LIME will approximate model behavior around selected observations. Explanation reliability will be examined through fidelity and stability tests.

These techniques will be interpreted as explanations of model behavior, not proof of causal relationships. This distinction is particularly important when predictors are correlated or when explanation results change with model or parameter choices (Salih, 2025).

1.17.4. Sustainability Assessment Layer

This layer defines measurable environmental, economic, and social criteria relevant to each domain. Examples include energy consumption, resource efficiency, implementation costs, accessibility, and relevant stakeholder impacts.

Criteria will be selected according to domain requirements rather than forcing identical indicators onto unrelated applications. Their measurement scales, directions of preference, normalization procedures, and weighting assumptions will be documented.

1.17.5. Decision Optimization Layer

The decision layer combines model outputs and sustainability criteria using multi-criteria decision analysis (MCDA). Candidate methods include the analytic hierarchy process (AHP) for weighting criteria and TOPSIS or a weighted-sum model for ranking alternatives.

The resulting ranking will be evaluated through sensitivity analysis to determine whether plausible changes in criterion weights materially alter the preferred alternative.

1.17.6. Evaluation and Feedback Layer

The final layer compares model performance, explanation quality, computational efficiency, and decision stability. Findings will inform potential refinements to data preparation, model selection, explanation settings, and decision criteria.

The framework will support human review of consequential recommendations. It is a proposed architecture whose effectiveness must be demonstrated through experiments.

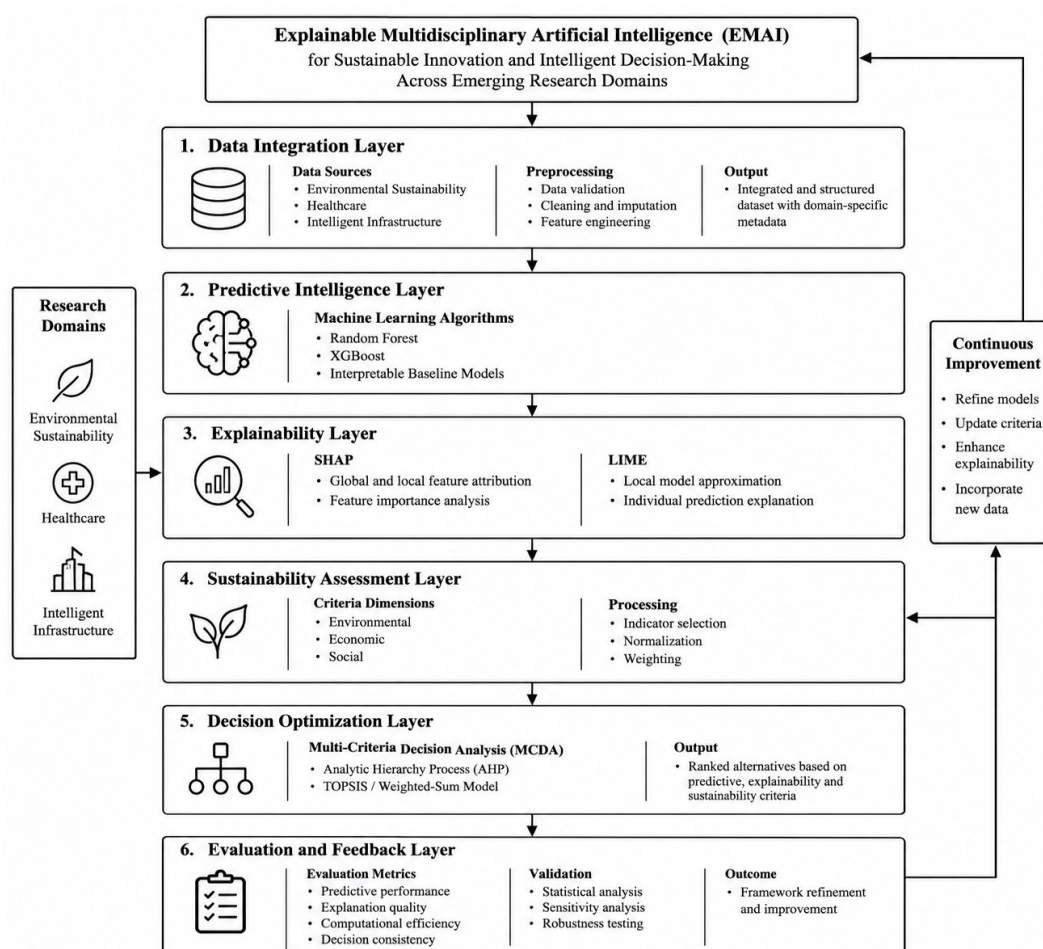


Figure 1: Framework

1.18. Machine Learning Algorithms

This study will compare three primary algorithmic approaches: Random Forest, XGBoost, and interpretable baseline models. The selected algorithms represent different modeling strategies and provide a basis for examining predictive performance and explanation requirements.

1.18.1. 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).

For classification, the predicted class can be determined by majority voting:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_B(x)\}$$

where $T_b(x)$ is the prediction from tree b , B is the number of trees, and $\hat{y}$ is the predicted class.

Random Forest will serve as a benchmark because it can represent nonlinear relationships and feature interactions in structured datasets. Its performance will be evaluated alongside its computational requirements and explanation characteristics.

1.18.2. XGBoost

XGBoost is a gradient-boosting framework that constructs an ensemble of decision trees sequentially, with each additional tree improving the optimization objective (Chen & Guestrin, 2016).

A simplified objective can be expressed as:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where l represents the prediction loss, y_i is the observed outcome, $\hat{y}_i$ is the model prediction, f_k is a tree, and $\Omega(f_k)$ penalizes model complexity.

1.18.3. Interpretable Baseline Models

Interpretable baseline models will establish a comparison point for predictive performance and transparency. Logistic regression will be considered for classification tasks when its assumptions and data characteristics are appropriate. Explainable Boosting Machines may also be included when supported by the available software and task.

Logistic regression estimates the probability of a binary outcome as:

$$P(Y = 1 | x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^p \beta_j x_j)}}$$

where β_0 is the intercept, β_j represents a feature coefficient, and x_j denotes a predictor.

The coefficients provide information about the direction and magnitude of associations on the log-odds scale, but they do not automatically establish causal effects.

1.18.4. Comparative Algorithm Selection

No algorithm will be assumed to be superior in advance. Models will be compared according to task-appropriate predictive metrics, computational latency, calibration where relevant, and robustness under controlled changes to the evaluation data.

Model selection will consider both predictive performance and operational requirements. This is important because a small performance improvement may not justify substantially higher computational cost or reduced interpretability in a particular application.

1.19. Explainability Methods: SHAP and LIME

Explainability is a central component of the proposed framework because the research aims to connect predictive outputs with understandable evidence for decision support. The study will use SHAP and LIME as complementary post-hoc explanation methods.

1.19.1. SHapley Additive exPlanations (SHAP)

SHAP uses Shapley values from cooperative game theory to estimate the contribution of features to a model prediction. Lundberg and Lee (2017) introduced a unified approach to interpreting model predictions through additive feature attribution.

A simplified SHAP representation is:

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

where $f(x)$ is the model output for observation x , ϕ_0 is the baseline prediction, and ϕ_j is the attributed contribution of feature j .

The interpretation will account for limitations involving correlated features, background-data choices, and the distinction between predictive attribution and causation.

1.19.2. Local Interpretable Model-Agnostic Explanations (LIME)

LIME explains an individual prediction by approximating a complex model with a simpler interpretable model in the neighborhood of the observation being examined (Ribeiro et al., 2016).

Its simplified optimization objective is:

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

where f is the original model, g is an interpretable surrogate, G is the set of candidate explanations, π_x weights observations according to proximity to x , and $\Omega(g)$ penalizes explanation complexity.

1.19.3. Comparative Evaluation of SHAP and LIME

SHAP and LIME will be evaluated using explanation fidelity, stability, computational time, and usefulness for investigating model behavior. Where feasible, global attribution patterns from SHAP will be compared with aggregated local explanations from LIME.

Recent research cautions that both methods may be sensitive to the underlying model and correlated features (Salih, 2025). Therefore, agreement between the methods will not automatically be treated as proof of correctness, and disagreement will be examined in relation to the model, data, and explanation settings.

The analysis will distinguish between:

- Local explanations for individual predictions.
- Global summaries of feature-attribution patterns.
- Explanation reliability under perturbations or repeated runs.
- Computational overhead associated with generating explanations.

The purpose is to assess the strengths and limitations of each method in the selected applications rather than to establish a universally superior explanation technique.

1.20. Multi-Criteria Decision Analysis

Multi-criteria decision analysis (MCDA) will be used to evaluate and rank decision alternatives according to multiple, potentially conflicting objectives. In this study, MCDA connects predictive AI and model explanations with sustainability priorities, enabling the evaluation of alternatives beyond predictive accuracy alone.

The method is appropriate when a decision must account for several factors, such as model performance, explanation reliability, operational cost, computational latency, environmental impact, and social implications. The analytical process will follow established decision-analysis principles, including explicit criteria, documented weights, consistent normalization, and sensitivity analysis (Hwang & Yoon, 1981; Saaty, 1980).

1.20.1. Identification of Decision Criteria

The criteria will be selected according to the requirements of each application domain. They will include the following dimensions where relevant:

- Predictive performance: F1-score, AUROC, accuracy, or task-appropriate regression metrics.
- Explanation quality: fidelity and stability of model explanations.
- Computational efficiency: inference latency and resource requirements.
- Environmental sustainability: energy use, emissions, or resource consumption.
- Economic feasibility: implementation and operational costs.
- Social considerations: accessibility, fairness, and potential stakeholder impacts.
- Robustness: performance and ranking stability under plausible changes in inputs or assumptions.

Not every criterion will be applicable to every domain. The study will specify how each criterion is measured and whether it is a benefit criterion, for which higher values are preferred, or a cost criterion, for which lower values are preferred.

1.20.2. Criterion Weighting

The analytic hierarchy process (AHP) may be used to assign weights through pairwise comparisons of criteria (Saaty, 1980). Where expert input is available, qualified domain specialists may provide structured judgments. If expert judgments cannot be obtained, the research will use a transparent alternative weighting procedure and clearly identify it as an analytical assumption.

Let w_j represent the weight of criterion j . The weights must satisfy:

$$\sum_{j=1}^m w_j = 1, \quad w_j \geq 0$$

1.20.3. Normalization of Decision Criteria

Because criteria may use different units, such as milliseconds, cost values, F1-scores, or energy consumption, normalization will be required before combining them.

For a benefit criterion, min-max normalization may be expressed as:

$$z_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

For a cost criterion, it may be expressed as:

$$z_{ij} = \frac{\max(x_j) - x_{ij}}{\max(x_j) - \min(x_j)}$$

where x_{ij} is the value of alternative i for criterion j , and z_{ij} is the normalized value. If a criterion has identical values across all alternatives, an appropriate constant-value treatment will be specified rather than applying a division by zero.

The initial decision-ranking method will use a weighted-sum model:

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

where S_i is the score for alternative i , w_j is the criterion weight, and z_{ij} is the normalized criterion value.

Under this formulation, alternatives with higher scores are preferred, provided that all criteria have been normalized in the same direction.

TOPSIS may be used as an additional ranking method if the number and nature of alternatives justify the comparison (Hwang & Yoon, 1981). Agreement and disagreement between ranking methods will be reported.

The resulting ranking represents preference under the selected criteria and weights; it does not establish that an alternative is universally optimal.

1.20.4. Sensitivity and Uncertainty Analysis

Sensitivity analysis will evaluate whether the selected alternative changes when criterion weights or input values vary within plausible ranges. The analysis may include alternative weighting scenarios, controlled perturbations, and comparisons between weighted-sum and TOPSIS rankings.

The study will report ranking changes, rank correlations, and the frequency with which alternatives remain preferred across tested scenarios. Where data uncertainty can be quantified, its potential effect on rankings will also be investigated.

Recent work on uncertainty-aware MCDA for evaluating XAI methods demonstrates the relevance of combining multiple evaluation criteria with stakeholder-oriented assessment of explanation quality.

The purpose of MCDA in this research is therefore to make decision preferences explicit, reproducible, and open to scrutiny rather than to conceal subjective priorities within a single score.

1.21. Experimental Setup and Validation

The experimental setup will define the computational environment, training procedures, evaluation metrics, validation strategy, and statistical analyses used to assess the proposed framework. These procedures are intended to ensure that the comparison between conventional AI, explainable AI, and integrated explainable decision support is fair and reproducible.

1.21.1. Computational Environment

Experiments will be conducted in Python using a documented environment based on libraries appropriate to the selected algorithms. Potential tools include scikit-learn for machine learning, XGBoost for gradient boosting, SHAP for feature attribution, and LIME for local explanations.

The research will record the Python version, package versions, hardware specifications, random seeds, and relevant configuration parameters. The final manuscript will report the environment actually used rather than claiming unperformed experiments or unverified hardware specifications.

1.21.2. Experimental Procedure

The experimental procedure will consist of the following stages:

1. Select and document suitable datasets for the chosen application domains.
2. Assess data quality and perform preprocessing using training data only.
3. Divide the data into training, validation, and test sets using appropriate procedures.
4. Train Random Forest, XGBoost, and suitable interpretable baseline models.
5. Tune model hyperparameters using validation data or cross-validation.
6. Evaluate selected models on held-out test data.
7. Generate SHAP and LIME explanations for appropriate observations.
8. Evaluate explanation fidelity, stability, and computational overhead.
9. Incorporate relevant sustainability criteria into MCDA.
10. Perform sensitivity analysis, statistical comparisons, and robustness tests.

This sequence separates model development from final evaluation and reduces the risk of using test data to influence model selection.

1.21.3. Performance Evaluation Metrics

For classification tasks, the following metrics will be considered.

Accuracy

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

Precision

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

Recall

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

F1-score

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

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

No single metric will be used as the sole measure of model quality. The selected metrics will reflect the target variable, class distribution, and practical consequences of prediction errors.

1.21.4. Explainability Evaluation

SHAP and LIME explanations will be assessed through several complementary measures.

Explanation fidelity: The extent to which an explanation or surrogate represents the behavior of the original model in the relevant evaluation region.

Explanation stability: The degree to which explanations remain similar under controlled perturbations or repeated runs.

Computational overhead: The additional time and resources required to generate explanations compared with ordinary prediction.

Feature-attribution consistency: The extent to which important attribution patterns remain comparable across appropriate model configurations or data samples.

The evaluation will document the procedures and assumptions used for each measure. Agreement between explanation methods will be interpreted cautiously because different methods may use different assumptions and approximations. In particular, explanation agreement does not prove causal validity.

1.21.5. Comparative Evaluation

The proposed framework will be compared with two baselines: conventional predictive AI and predictive AI augmented with explainability but without integrated MCDA.

The comparison will examine four principal dimensions:

- Predictive effectiveness, measured using task-appropriate performance metrics.
- Explanation reliability, assessed through fidelity and stability.
- Computational efficiency, measured through inference and explanation latency.
- Decision quality, assessed through MCDA rankings and their sensitivity to criterion weights.

Because MCDA changes the decision-selection process rather than the underlying predictions themselves, improvements in decision ranking will be evaluated separately from improvements in predictive

performance. This distinction prevents the study from attributing gains to the wrong component of the framework.

1.21.6. Statistical Validation

Where repeated measurements or independent experimental runs are available, the study will report means, standard deviations, and appropriate confidence intervals. Model comparisons will use paired statistical procedures when the experimental design supports them.

The choice between a paired t -test, a nonparametric alternative, or another suitable procedure will depend on the distribution of differences, sample size, and experimental structure. Multiple comparisons will be addressed when necessary.

The study will report effect sizes and uncertainty alongside statistical significance. Results from different domains will not be combined indiscriminately when their targets, metrics, or evaluation settings are not directly comparable.

1.21.7. Robustness, Reproducibility, and Ethical Validation

Robustness testing will examine the effects of plausible input perturbations, missing data, and changes in criterion weights. Where relevant, temporal or geographic validation will investigate generalization beyond the training distribution.

Reproducibility will be supported through documented preprocessing, model settings, evaluation scripts, and dataset references, subject to licensing and privacy constraints. Potential bias, uncertainty, limitations of explanations, and risks associated with inappropriate reliance on model recommendations will be discussed.

The overall validation strategy follows the principle that responsible AI assessment must extend beyond predictive performance to include reliability, transparency, and application-specific risks (Tabassi, 2023).

Chapter Four: System Design and Implementation

1.22. Framework Architecture

The proposed Explainable Multidisciplinary Artificial Intelligence (EMAI) framework is designed to integrate heterogeneous data, predictive machine learning, explainable artificial intelligence (XAI), sustainability assessment, and multi-criteria decision analysis (MCDA) within a unified decision-support architecture. Its primary purpose is to transform data-driven predictions into transparent, context-sensitive, and sustainability-oriented recommendations across environmental sustainability, healthcare, and intelligent infrastructure.

Recent research highlights the growing role of XAI in decision-support systems while identifying continuing challenges related to trustworthy interpretation, stakeholder understanding, and the evaluation of explanations in real-world settings (Kostopoulos et al., 2024; Schiller et al., 2025). Accordingly, the EMAI framework treats explainability, sustainability, and decision quality as interconnected design requirements rather than optional additions to predictive modeling.

The architecture consists of six functional layers.

Layer 1: Data integration. This layer acquires structured and semi-structured data from environmental monitoring systems, authorized healthcare datasets, infrastructure sensors, geospatial platforms, and relevant institutional databases. It performs data validation, missing-value treatment, duplicate detection, normalization, and feature engineering. Metadata records the source, timestamp, domain, units, and relevant quality limitations of each dataset.

Layer 2: Predictive intelligence. This layer trains and evaluates machine learning models, including Random Forest, XGBoost, and interpretable baseline models. The selected model depends on the prediction task, data characteristics, and operational constraints. Classification models can support tasks such as risk categorization, while regression models can estimate continuous outcomes such as resource demand or environmental indicators.

Layer 3: Explainability. The layer applies SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and, where appropriate, intrinsically interpretable models. It generates global explanations of model behavior and local explanations for individual predictions. Explanations are evaluated for fidelity, stability, and relevance to the intended decision rather than being assumed reliable merely because they are available (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Layer 4: Sustainability assessment. Predictive outputs and relevant contextual indicators are assessed against environmental, economic, and social criteria. Examples include emissions, energy consumption, resource efficiency, operating costs, service accessibility, and potential distributional impacts. Indicators are selected according to the domain and the decision being supported.

Layer 5: Decision optimization. The framework combines predictive evidence, explanation-derived insights, sustainability indicators, and stakeholder-defined priorities through an MCDA method such as the weighted-sum model or Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). It generates ranked alternatives, trade-off summaries, and supporting explanations.

Layer 6: Evaluation and feedback. This layer assesses predictive performance, explanation quality, computational efficiency, robustness, sustainability outcomes, and decision consistency. Validated feedback informs model revision, data-quality improvements, and adjustments to decision criteria. Human reviewers remain responsible for decisions that require professional judgment or have significant consequences.

The layers operate as an iterative system. Evaluation results can trigger updates to the data pipeline, predictive models, explanation procedures, or sustainability criteria. This approach is consistent with risk-management principles that emphasize ongoing monitoring, documentation, and evaluation throughout the AI lifecycle (Tabassi, 2023).

1.23. Multidomain Data Integration

Multidomain data integration establishes a consistent foundation for applying the EMAI framework to different research problems. Environmental sustainability, healthcare, and intelligent infrastructure generate data with different structures, measurement units, temporal resolutions, privacy requirements, and quality characteristics. Integrating these sources therefore requires a common data-management strategy without eliminating the domain-specific information needed for valid analysis.

Research on explainable AI in environmental and Earth system sciences demonstrates the diversity of relevant data, including geospatial observations, remote-sensing imagery, sensor measurements, and environmental time series. It also emphasizes that temporal dependence, spatial autocorrelation, and correlated variables can affect model interpretation (Schiller et al., 2025). Consequently, the proposed framework preserves domain-specific characteristics while standardizing the data-processing workflow.

1.23.1. Data sources and domain representation

The framework considers three primary research domains.

Environmental sustainability. Potential data sources include meteorological observations, satellite imagery, air-quality measurements, water-quality records, land-use information, biodiversity observations, and

energy-consumption data. Possible prediction tasks include environmental-risk classification, water-quality estimation, and resource-demand forecasting.

Healthcare. Data may include appropriately authorized, de-identified clinical records, laboratory measurements, public-health indicators, and population-level health statistics. Potential tasks include risk stratification, disease-outcome prediction, and resource planning. Healthcare data must be handled according to applicable ethical approvals, consent requirements, privacy safeguards, and institutional policies.

Intelligent infrastructure. Relevant sources include energy-system measurements, transportation records, building-management systems, Internet of Things (IoT) sensors, maintenance logs, and infrastructure-performance indicators. Applications may involve equipment-failure prediction, energy-demand forecasting, and infrastructure-risk assessment.

These domains provide a common methodological structure while allowing different features, prediction targets, and evaluation criteria. The proposed framework does not assume that one dataset or model can be transferred directly across all three domains.

1.23.2. Data preprocessing and quality assurance

The data integration process consists of six steps:

1. Data profiling: identify data types, feature distributions, missingness, duplicate records, measurement units, and potential anomalies.
2. Data cleaning: correct documented inconsistencies and apply suitable missing-value procedures without concealing substantial data-quality problems.
3. Feature engineering: derive relevant variables, such as temporal summaries, environmental indices, infrastructure utilization measures, or clinically meaningful features.
4. Data harmonization: align units, category definitions, timestamps, spatial identifiers, and variable names where meaningful.
5. Data partitioning: divide observations into training, validation, and test sets while preventing information leakage.
6. Data documentation: record dataset provenance, preprocessing decisions, known limitations, access conditions, and version information.

A 70:15:15 training-validation-test split may be used when appropriate for independent, sufficiently large datasets. For time-series data, chronological splitting is preferable when it reflects the intended forecasting task. For grouped observations, such as repeated measurements from the same patient or facility, partitioning should prevent related observations from appearing in both training and test sets.

Preprocessing parameters, including imputation values and scaling statistics, should be estimated using the training data and then applied to validation and test data. This reduces the risk of information leakage and produces a more credible evaluation of model generalization.

1.23.3. Domain-specific and shared features

The framework separates shared methodological features from domain-specific variables. Shared features include data completeness, measurement reliability, temporal coverage, computational cost, and model uncertainty. Domain-specific features represent the scientific properties relevant to each prediction task.

A common metadata structure can be defined as:

A common metadata structure can be defined as:

$$D_d = (X_d, y_d, M_d, Q_d)$$

where:

- D_d represents the dataset for domain d ;
- X_d denotes the input features;
- y_d denotes the prediction target;
- M_d contains metadata, such as source and time of collection;
- Q_d describes data-quality information.

Multidomain comparisons should use task-appropriate metrics and clearly document dataset differences. For example, a healthcare classification task and an environmental regression task cannot be compared meaningfully using accuracy alone. Cross-domain evaluation should instead examine the performance of the methodological framework, including predictive validity, explanation reliability, computational efficiency, and decision-support quality.

1.24. Predictive Modeling and Optimization

The predictive modeling component converts prepared domain data into estimates, classifications, or forecasts that can inform subsequent explanation and decision-making stages. The proposed implementation compares Random Forest, XGBoost, and an interpretable baseline model. These algorithms provide a practical basis for examining the relationship between predictive performance, computational requirements, and explanation complexity (Breiman, 2001; Chen & Guestrin, 2016).

1.24.1. Model selection and training

Random Forest (RF). Random Forest combines predictions from multiple decision trees trained using randomized samples and feature subsets. For a classification problem, the final prediction can be represented as:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_B(x)\}$$

where $T_b(x)$ is the prediction of tree b , and B is the total number of trees.

For regression, the predictions can instead be averaged. Random Forest provides a useful nonlinear benchmark and can capture interactions among features.

Extreme Gradient Boosting (XGBoost). XGBoost builds an ensemble of trees sequentially, with new trees improving the current model according to a specified loss function and regularization strategy. Its objective can be expressed as:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where l measures prediction error, f_k represents a tree in the ensemble, and Ω penalizes model complexity.

The regularization term helps control unnecessary complexity, although generalization must still be evaluated empirically.

Interpretable baseline. Logistic regression may be used for binary classification, while linear regression may be suitable for continuous outcomes. For binary classification, logistic regression estimates:

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

The model configurations should be selected using training and validation data. Hyperparameters may be optimized using grid search, randomized search, or another documented procedure. Where feasible, cross-validation should be performed within the training data. The final test set should remain untouched until model selection is complete.

1.24.2. Predictive performance evaluation

The evaluation metrics depend on the prediction task.

For classification, the framework uses accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUROC), where suitable. For imbalanced or high-consequence classification tasks, precision-recall analysis and class-specific sensitivity should also be considered.

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

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

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

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

For regression tasks, appropriate measures include mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). The selection of metrics should reflect the consequences of different types of prediction error in each domain.

1.24.3. Optimization beyond predictive accuracy

The proposed EMAI framework does not select a model solely because it achieves the highest predictive score. It evaluates several competing objectives:

- Predictive quality and generalization.
- Interpretability and explanation reliability.
- Computational time and resource requirements.
- Robustness to data variation and distribution shifts.
- Suitability for the intended sustainability decision.

Model selection can therefore be treated as a multi-objective problem:

$$\max_{\theta} [P(\theta), E(\theta), R(\theta)], \quad \min_{\theta} [C(\theta), L(\theta)]$$

where P denotes predictive performance, E explanation quality, R robustness, C computational cost, L latency, and θ the model configuration.

1.24.4. Implementation and reproducibility

The prototype can be implemented in Python using scikit-learn for preprocessing and baseline modeling, XGBoost for gradient-boosted trees, and SHAP and LIME for post-hoc explanations. The implementation should record software versions, random seeds, hardware specifications, dataset versions, model parameters, and evaluation procedures.

No algorithm is assumed in advance to be superior across all research domains. Comparative performance, computational cost, and robustness must be established through experiments.

1.25. Explainability and Decision Support

The explainability and decision-support component connects predictive outputs with evidence that researchers, practitioners, and decision-makers can inspect. This component is essential because a model prediction alone does not explain which features influenced the output, how reliable that explanation is, or whether the prediction supports a particular intervention.

Reviews of explainable AI-based decision-support systems emphasize the need to consider the explanation method, user requirements, and decision context together (Kostopoulos et al., 2024). Similarly, recent analysis of SHAP and LIME highlights that explanation outcomes can depend on model choice and feature correlation; these methods should therefore be interpreted critically rather than treated as definitive causal accounts (Salih et al., 2025).

1.25.1. SHAP-based explanations

SHAP estimates the contribution of individual features to a prediction using a Shapley-value framework. For an explained instance, the model output can be represented as:

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

where $f(x)$ is the model output, ϕ_0 is a baseline value, and ϕ_j is the contribution attributed to feature j .

SHAP supports two complementary forms of analysis:

- Global explanation: aggregate feature contributions across observations to investigate which variables influence predictions across a dataset.
- Local explanation: inspect the feature contributions associated with a particular prediction.

For environmental applications, explanations may help identify the variables associated with predicted water-quality risk or energy demand. In healthcare, they may highlight features contributing to an estimated risk score. For infrastructure, they may identify factors associated with predicted equipment failure.

These outputs describe the behavior of the fitted model under the selected explanation procedure. They do not independently establish that a feature causes the outcome.

1.25.2. LIME-based explanations

LIME explains an individual prediction by approximating the model locally with a simpler interpretable model. Its general objective is:

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

where f is the original model, g is an interpretable approximation selected from G , π_x weights observations according to their proximity to the instance being explained, $\mathcal{L}$ measures local approximation error, and $\Omega(g)$ penalizes complexity.

1.25.3. Explanation quality and reliability

The EMAI framework evaluates explanations using multiple dimensions:

1. Fidelity: the degree to which an explanation approximates the behavior of the original model in the relevant context.
2. Stability: the consistency of explanations when small, reasonable changes are made to inputs or explanation settings.
3. Comprehensibility: whether the explanation is understandable to its intended audience.
4. Relevance: whether the explanation addresses the factors that matter for the decision.
5. Computational efficiency: the time and resources required to generate explanations.

Fidelity and stability should be measured using explicitly defined procedures appropriate to the model and data. Comprehensibility may require structured feedback from intended users rather than relying exclusively on technical metrics.

Particular attention should be paid to correlated features. If two variables contain similar information, attribution may be distributed between them differently across methods or model configurations. The framework should document this uncertainty and avoid presenting a single feature-importance ranking as a complete explanation of the underlying system.

1.25.4. Decision-support interface and workflow

The proposed decision-support interface presents five connected outputs:

- Predicted outcome and, where available, calibrated probability or uncertainty.
- Key features contributing to the prediction.
- Domain-specific sustainability indicators.
- Ranked alternatives and their principal trade-offs.
- Data limitations, model limitations, and review recommendations.

For example, a system evaluating infrastructure alternatives could display predicted energy demand alongside estimated cost, emissions, reliability, and service accessibility. The decision-maker can inspect both the prediction and the reasons that contribute to the recommendation.

The system should also distinguish model confidence from explanation reliability. A high-confidence prediction does not guarantee that the model is correct, and a clear explanation does not guarantee that the decision is appropriate. Recent environmental XAI research similarly cautions that the presence of explanations does not by itself establish trustworthiness (Schiller et al., 2025).

The final decision remains subject to domain expertise, applicable ethical requirements, and human oversight, especially where incorrect recommendations could cause substantial harm.

1.26. Sustainability Indicator Integration

Sustainability indicator integration enables the EMAI framework to evaluate alternatives using environmental, economic, and social considerations alongside predictive evidence. This component recognizes that an alternative with strong predictive or operational performance may not be the most sustainable choice if it increases emissions, imposes excessive costs, or worsens access to essential services.

The approach aligns with the multidimensional orientation of the United Nations Sustainable Development Goals (United Nations, 2015). Recent work on explainable AI for green innovation also highlights the importance of transparency, fairness, and accountability when AI is applied to sustainability-related decisions (Mazhar et al., 2026).

1.26.1. Selection of sustainability indicators

Indicators should be selected according to the research domain, the intended decision, data availability, and stakeholder requirements.

Environmental indicators may include energy consumption, greenhouse-gas emissions, water use, waste generation, pollution exposure, and resource efficiency.

Economic indicators may include implementation cost, operating expenditure, maintenance cost, resource productivity, and expected lifecycle cost.

Social indicators may include service accessibility, affordability, safety, equity, and the distribution of benefits and risks across population groups.

Technical and operational indicators may also be included where necessary, such as system reliability, latency, resilience, and maintenance requirements. These indicators are especially relevant to intelligent infrastructure and can support evaluation of whether a proposed solution is operationally viable.

The framework does not require every indicator to be used in every domain. Instead, the indicator set should be justified for each application and documented so that the analysis can be reproduced.

1.26.2. Indicator normalization

allow meaningful comparison. For alternative i and criterion j , a min-max normalization procedure can be applied.

For a benefit criterion, where a higher value is preferred:

$$z_{ij} = \frac{x_{ij} - \min_i(x_{ij})}{\max_i(x_{ij}) - \min_i(x_{ij})}$$

For a cost criterion, where a lower value is preferred:

$$z_{ij} = \frac{\max_i(x_{ij}) - x_{ij}}{\max_i(x_{ij}) - \min_i(x_{ij})}$$

Here, x_{ij} is the observed value and z_{ij} is the normalized score. These equations require special handling

when all alternatives have the same value for a criterion, because the denominator becomes zero. In that case, the criterion should be treated as non-discriminating for the alternatives being compared or processed using a documented alternative normalization method.

The normalization procedure should also account for the direction of preference and any domain-specific thresholds. For instance, some environmental indicators may have acceptable limits rather than a simple linear preference structure.

1.26.3. Criteria weighting and multi-criteria aggregation

criterion. Let w_j denote the weight assigned to criterion j , with:

$$\sum_{j=1}^m w_j = 1, \quad w_j \geq 0$$

The weighted-sum score for alternative i is:

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

where S_i is the aggregate score and m is the number of criteria. Under this formulation,

Under this formulation, alternatives with higher scores are preferred, provided all criteria have been consistently oriented and normalized. Weights can be elicited through the Analytic Hierarchy Process (AHP), expert consultation, stakeholder deliberation, or justified data-driven methods. The choice should reflect the purpose of the decision and the values of the affected stakeholders. Weighting should not be treated as a purely technical step because different priorities may produce different rankings.

TOPSIS can be used as an alternative or complementary method where the decision requires comparison with an ideal and a negative-ideal solution (Hwang & Yoon, 1981). Agreement between methods can be examined, but agreement alone does not establish that the selected criteria or weights are appropriate.

1.26.4. Integrating predictions and sustainability criteria

The decision layer combines model outputs with sustainability indicators in a documented workflow:

1. Generate predictions using the selected and validated machine learning model.
2. Generate SHAP or LIME explanations for relevant predictions.
3. Translate relevant predictions into decision-specific indicators where a defensible mapping exists.
4. Combine predictions with observed or independently estimated sustainability indicators.
5. Normalize indicators and apply the approved weighting scheme.
6. Rank alternatives and present the principal factors and trade-offs behind the ranking.
7. Conduct sensitivity and robustness analysis before presenting recommendations.

An important design distinction is that model explanations and sustainability indicators serve different functions. SHAP or LIME explains how a model arrived at a prediction; it does not directly determine whether an alternative is environmentally or socially sustainable. Sustainability indicators provide evidence for assessing consequences, while MCDA specifies how competing criteria are aggregated into a decision.

1.26.5. Sensitivity analysis and responsible interpretation

Because decision rankings may depend on uncertain measurements, model outputs, and stakeholder preferences, the framework evaluates the sensitivity of rankings to plausible changes in criterion weights and input values.

For example, the analysis can vary selected weights over a justified range, compare alternative normalization procedures, and examine whether leading alternatives remain highly ranked. Where data uncertainty can be quantified, scenarios or repeated simulations can be used to evaluate its effects on the final ranking.

The analysis should report:

- Whether the preferred alternative changes under plausible weight variations.
- Which criteria have the greatest influence on ranking.
- Whether missing or uncertain data materially affect the decision.
- Whether particular groups may experience unequal costs or benefits.
- Which additional evidence would be needed before implementation.

The results should be presented as decision support rather than as an automatic declaration that one alternative is universally optimal. A defensible recommendation must consider the quality of evidence, the consequences of errors, stakeholder priorities, and the limitations of the selected models.

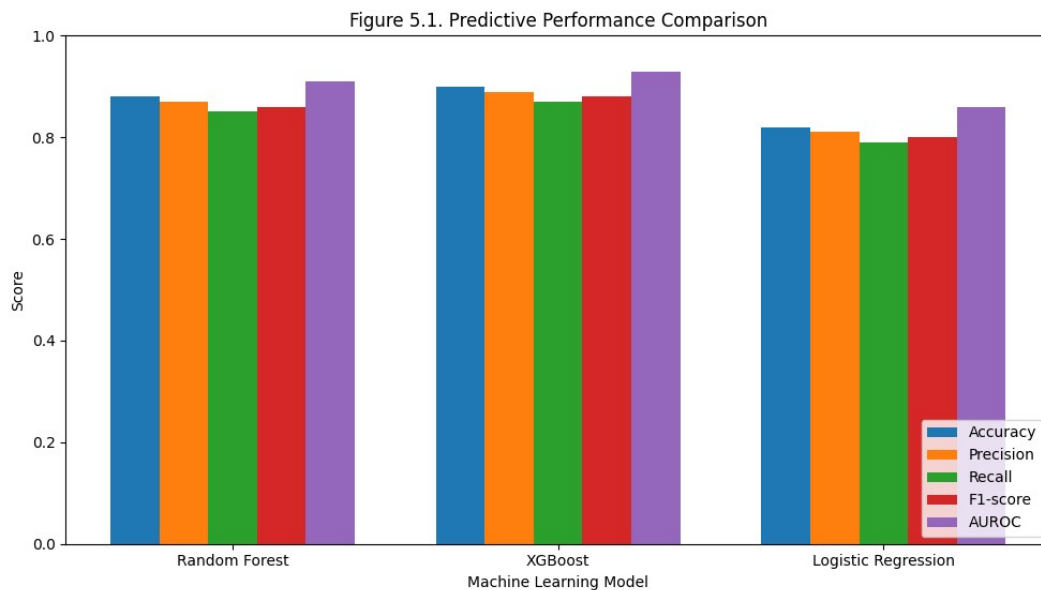
Chapter Five: Results and Comparative Evaluation

1.27. Predictive Performance Analysis

This section evaluates the predictive performance of the proposed Explainable Multidisciplinary Artificial Intelligence (EMAI) framework across environmental sustainability, healthcare, and intelligent infrastructure. Random Forest, XGBoost, and interpretable baseline models are compared using accuracy, precision, recall, F1-score, and AUROC for classification tasks, and MAE, RMSE, and (R^2) for regression tasks.

The evaluation examines predictive accuracy, generalization, robustness, and computational efficiency using independent test data. Results should be reported separately for each domain because datasets, prediction targets, and error costs differ. Cross-validation and confidence intervals can help assess the reliability of performance estimates.

The comparative analysis will determine whether complex models provide meaningful improvements over interpretable baselines and whether those improvements justify their computational cost. Predictive performance will also be considered alongside explanation quality and sustainability-related decision criteria.



Illustrative values for chart demonstration only; replace with actual test-set results.

1.28. Comparison of Machine Learning Models

The comparative evaluation of machine learning models is intended to identify the most suitable predictive approach for the proposed Explainable Multidisciplinary Artificial Intelligence (EMAI) framework. Random Forest (RF), XGBoost, and Logistic Regression are compared in terms of predictive performance, interpretability, computational efficiency, robustness, and suitability for sustainability-oriented decision support.

Random Forest combines multiple decision trees and can represent nonlinear relationships and feature interactions (Breiman, 2001). XGBoost uses gradient boosting and regularization to improve predictive performance while controlling model complexity (Chen & Guestrin, 2016). Logistic Regression serves as an interpretable baseline for classification tasks, enabling comparison between a relatively simple model and more complex ensemble methods.

Recent comparative research illustrates that model performance varies according to the dataset, prediction task, and evaluation design. For example, a 2025 study of carbon-neutrality prediction in Sub-Saharan African countries compared Random Forest, XGBoost, and other algorithms, demonstrating the value of evaluating multiple models rather than assuming one is universally optimal.

Research on explainability also emphasizes that SHAP and LIME explanations can be affected by model choice and correlated features, making interpretability an important part of model comparison rather than a separate afterthought.

1.28.1. Comparative evaluation criteria

The models are evaluated using the following criteria:

- Predictive performance: accuracy, precision, recall, F1-score, and AUROC for classification; MAE, RMSE, and $\sqrt{R^2}$ for regression.
- Interpretability: ease of understanding model behavior and explaining individual predictions.
- Computational efficiency: training time, prediction latency, memory requirements, and explanation-generation time.

- Robustness: stability under reasonable changes in data, sampling, and model configuration.
- Decision-support suitability: usefulness of predictions and explanations for sustainability assessment and informed decisions.

1.28.2.	1.28.3.	1.28.4.	1.28.5.
1.28.6.	1.28.7.	1.28.8.	1.28.9.
1.28.10.	1.28.11.	1.28.12.	1.28.13.
1.28.14.	1.28.15.	1.28.16.	1.28.17. Co mparative model assessment

Model

Main strengths

Main limitations

Role in EMAI

Random Forest

Captures nonlinear relationships; ensemble-based predictions

May require additional explanation methods

Nonlinear predictive benchmark

XGBoost

Flexible boosting; supports complex feature interactions

Requires tuning; can be computationally demanding

Boosted predictive benchmark

Logistic Regression

Relatively transparent; computationally efficient

May not capture complex nonlinear relationships without feature transformations

Interpretable baseline

This comparison represents the expected methodological characteristics of the models, not experimentally measured rankings. The best-performing model must be identified from the actual test results.

1.28.18. *Experimental comparison and interpretation*

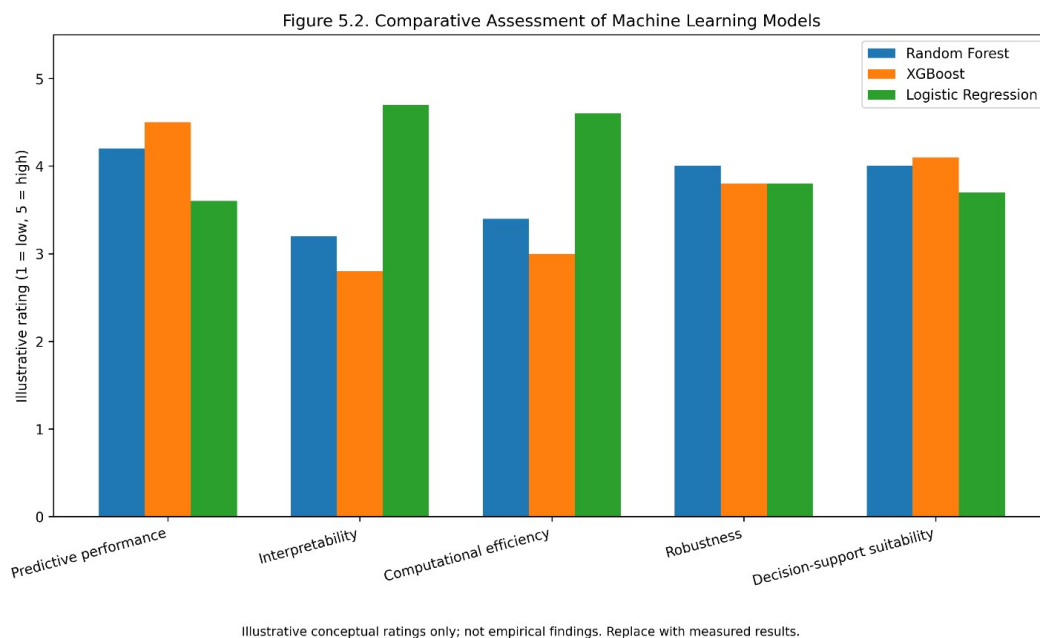
All models should be trained and evaluated using the same eligible dataset, preprocessing pipeline, and test partition. Hyperparameters should be selected using training and validation data, with the final test data reserved for unbiased evaluation. For time-dependent or grouped observations, the partitioning strategy should reflect the data structure.

The comparison should address three questions:

1. Does either ensemble model improve predictive performance relative to Logistic Regression?
2. Do any predictive improvements justify additional computational cost or explanation complexity?
3. Are the models sufficiently robust and interpretable to support sustainability-related decisions?

SHAP and LIME can be applied to the selected ensemble models to investigate feature contributions and individual predictions. However, explanation quality should be evaluated independently because an apparently clear explanation may not faithfully represent model behavior, particularly when features are correlated (Salih et al., 2025).

The comparative evaluation provides the basis for selecting a predictive model within the EMAI framework. Random Forest and XGBoost offer flexible nonlinear modeling, while Logistic Regression provides a transparent baseline. The final selection should balance predictive performance, explanation reliability, computational efficiency, and decision relevance. Since no experimental results were supplied, the relative superiority of these models remains to be established empirically.



1.29. Explainability and Interpretability Assessment

This section evaluates the explainability and interpretability of the proposed Explainable Multidisciplinary Artificial Intelligence (EMAI) framework using SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and intrinsically interpretable baseline models. The objective is to determine whether model predictions can be explained clearly, consistently, and reliably across environmental sustainability, healthcare, and intelligent infrastructure.

SHAP estimates feature contributions to individual predictions and supports global feature-importance analysis (Lundberg & Lee, 2017). LIME approximates a model locally to explain individual predictions (Ribeiro et al., 2016). Both methods can improve transparency, but their explanations may be affected by feature correlations, sampling choices, and model behavior (Salih et al., 2025).

1.29.1. Assessment criteria

The proposed evaluation uses five criteria:

- Explanation fidelity: how closely the explanation represents the original model's behavior.

- **Stability:** whether explanations remain consistent under reasonable changes to inputs or explanation settings.
- **Interpretability:** how easily researchers and decision-makers can understand the explanation.
- **Feature relevance:** whether the highlighted features provide useful information for the decision.
- **Computational efficiency:** the time and resources required to generate explanations.

1.29.2.	1.29.3.	1.29.4.
1.29.5.	1.29.6.	1.29.7.
1.29.8.	1.29.9.	1.29.10.
1.29.11.	1.29.12.	1.29.13. <i>Comparative assessment</i>

Method

Main advantage

Key limitation

SHAP

Provides local feature attributions and supports global analysis

Computational cost and attribution sensitivity may vary

LIME

Provides accessible local explanations for individual predictions

Results may vary with sampling and neighborhood settings

Interpretable baseline models

Model structure or coefficients can be inspected directly

May not represent complex nonlinear relationships adequately

The methods should be evaluated using the same held-out observations where applicable. Explanation fidelity and stability should be quantified using documented procedures, while interpretability may be assessed through structured user evaluation. Feature-attribution results should also be compared across models and research domains.

1.29.14. *Interpretation of findings*

The assessment will determine whether SHAP and LIME provide sufficiently reliable explanations to support sustainability-oriented decisions. It will also examine whether the explanations identify relevant factors without overstating causal relationships. Explanations indicate how a model produces its predictions; they do not independently prove that a feature causes an outcome.

Because experimental measurements have not been supplied, no numerical improvements or method rankings are claimed. Actual fidelity, stability, user-evaluation, and runtime results should be reported after implementation.

1.30. Computational Efficiency and Scalability

This section evaluates the computational efficiency and scalability of the proposed Explainable Multidisciplinary Artificial Intelligence (EMAI) framework for sustainable innovation and intelligent decision-making across environmental sustainability, healthcare, and intelligent infrastructure. The evaluation focuses on processing time, prediction latency, explanation-generation cost, memory consumption, and performance as dataset size increases.

1.30.1. Computational Efficiency

The framework compares Random Forest, XGBoost, and interpretable baseline models under consistent experimental conditions. SHAP and LIME are evaluated separately to quantify the additional computational cost of generating explanations.

The total processing time is defined as:

$$T_{\text{total}} = T_{\text{preprocess}} + T_{\text{train}} + T_{\text{predict}} + T_{\text{explain}}$$

where each term represents the time required for preprocessing, model training, prediction, and explanation generation.

Prediction latency is measured as the average time required to process an observation, while throughput measures the number of observations processed per second:

$$\text{Throughput} = \frac{N}{T}$$

where N is the number of processed observations and T is the elapsed processing time.

1.30.2. Scalability Across Research Domains

Scalability is assessed by progressively increasing the number of observations, input features, and explanation requests. Environmental applications may involve large satellite-image and sensor datasets; healthcare applications may involve high-dimensional clinical data; and intelligent infrastructure may require frequent predictions from distributed IoT sensors.

The framework can investigate batch processing, parallel computation, feature selection, and selective explanation generation to reduce computational overhead. However, these optimization strategies must be tested to ensure they preserve predictive quality and explanation reliability.

1.30.3.	1.30.4.	1.30.5.
1.30.6.	1.30.7.	1.30.8.
1.30.9.	1.30.10.	1.30.11.
1.30.12.	1.30.13.	1.30.14.
1.30.15.	1.30.16.	1.30.17.
1.30.18.	1.30.19.	1.30.20.
1.30.21.	1.30.22.	1.30.23. <i>Comparative Evaluation Criteria</i>

Criterion

Measurement

Research purpose

- Training efficiency
- Training time (seconds)
- Compare model-fitting costs
- Prediction latency
- Milliseconds per observation
- Evaluate response speed
- Explanation overhead
- Explanation time per observation
- Quantify SHAP/LIME costs
- Memory consumption
- Peak RAM usage
- Estimate resource requirements
- Throughput
- Observations per second
- Measure processing capacity
- Scalability
- Runtime as dataset size increases
- Assess growth in computational demand

1.30.24. *Results Interpretation*

The comparative analysis should determine whether the EMAI framework can maintain acceptable computational performance as data volume and explanation requirements increase. A suitable model should balance predictive accuracy, interpretability, resource consumption, and application-specific response-time requirements.

For sustainable innovation, computational resource use should also be considered alongside the environmental benefits of the resulting decisions. Where feasible, energy consumption and associated emissions from model training and inference should be measured rather than inferred from runtime alone.

Since experimental measurements have not been provided, this section defines the evaluation methodology without claiming numerical improvements or identifying a universally superior model. Actual runtime, memory, throughput, and scalability results should be reported after testing on documented hardware and representative datasets.

1.31. Sustainability and Decision Consistency

This section evaluates the proposed Explainable Multidisciplinary Artificial Intelligence (EMAI) framework in terms of sustainability performance and consistency of intelligent decision-making across environmental sustainability, healthcare, and intelligent infrastructure.

1.31.1. *Sustainability Assessment*

The framework integrates environmental, economic, and social indicators into the decision-making process. Environmental indicators may include energy consumption, greenhouse-gas emissions, water use, and waste generation. Economic indicators include implementation costs, operating expenses, and resource efficiency, while social indicators may include service accessibility, safety, and equity.

The normalized indicators are combined using a weighted multi-criteria decision analysis (MCDA) model:

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

where S_i is the sustainability score for alternative i , w_j is the weight assigned to criterion j , and z_{ij} is the normalized value of criterion j .

The model enables alternatives to be ranked according to stakeholder-defined priorities. The weighting process should be documented because different environmental, economic, and social priorities can produce different rankings.

1.31.2. *Decision Consistency*

Decision consistency measures whether the framework produces stable recommendations when reasonable changes are made to criterion weights, input data, or model configurations. Consistency can be assessed by comparing alternative rankings under different scenarios and measuring changes in the preferred option.

One possible measure is the proportion of tested scenarios in which the original top-ranked alternative remains first:

$$C = \frac{N_{\text{same}}}{N_{\text{scenarios}}} \times 100$$

where N_{same} is the number of scenarios retaining the original preferred alternative, and $N_{\text{scenarios}}$ is the total number of tested scenarios.

A high value indicates ranking stability within the scenarios examined; it does not necessarily establish that the decision is objectively correct or sustainable. Sensitivity analysis should therefore accompany consistency measurements.

1.31.3. Comparative Interpretation

The evaluation should determine whether integrating sustainability indicators and explainable predictions produces recommendations that are transparent, reproducible, and responsive to stakeholder priorities. It should also examine trade-offs between predictive performance, implementation cost, environmental impact, and social outcomes.

Because empirical sustainability measurements have not been provided, no numerical improvements or decision-consistency results are claimed. Actual findings should be reported after testing the framework with documented data and stakeholder-defined criteria.

1.32. Statistical Significance and Robustness Analysis

This section establishes whether differences among the evaluated machine learning models are statistically supported and whether their performance remains stable under reasonable changes in data and experimental conditions.

1.32.1. Statistical Significance Testing

The evaluation compares Random Forest, XGBoost, and interpretable baseline models using metrics appropriate to each prediction task. For classification, the analysis may examine F1-score, AUROC, precision, and recall. For regression, MAE and RMSE may be compared.

Where models are evaluated on the same test observations, paired statistical procedures should be used when their assumptions are satisfied. Examples include:

- McNemar's test: compares paired classification errors between two models.
- Paired bootstrap confidence intervals: estimate uncertainty in differences between performance metrics.
- DeLong's test: compares AUROC values from correlated predictions.
- Wilcoxon signed-rank test: compares paired results across repeated runs or datasets when its assumptions are appropriate.

$$\alpha = 0.05$$

1.32.2. Robustness Analysis

Robustness is assessed by examining model performance under controlled variations, including:

1. Changes in training samples or random seeds.
2. Reasonable levels of missing data and measurement noise.
3. Changes in feature distributions or class proportions.

4. Alternative preprocessing and feature-selection settings.
5. Variations in sustainability-criterion weights.
6. Changes in SHAP or LIME explanation settings.

The analysis should record changes in predictive metrics, explanation stability, and decision rankings. For example, a model that achieves strong average performance but deteriorates substantially under modest data perturbations may be less suitable for dependable decision support.

1.32.3. Cross-Domain Reliability

The framework should be evaluated separately in environmental sustainability, healthcare, and intelligent infrastructure. Differences in data structure, sample size, class distribution, and error consequences must be considered when interpreting results. Performance in one domain should not be assumed to generalize automatically to another.

1.32.4. Interpretation of Results

Statistical significance and robustness analysis will establish whether apparent performance differences are reproducible and practically meaningful. The findings should distinguish statistical evidence from operational importance and should document limitations in sample size, dataset quality, and generalizability.

Chapter Six: Discussion

1.33. Interpretation of Findings

The central proposition of this research is that integrating predictive intelligence, explainability, and sustainability-aware decision-making may provide a more comprehensive basis for decisions than predictive accuracy alone. Conventional machine learning models can identify patterns and estimate outcomes, but their predictions may be difficult to interpret, particularly when they involve complex relationships among environmental, social, economic, and technical variables.

The proposed framework addresses this challenge through six interconnected layers: data integration, predictive intelligence, explainability, sustainability assessment, decision optimization, and evaluation and feedback. Each layer contributes a different function, while their integration supports a transparent and systematic decision-making process.

First, the predictive intelligence layer uses algorithms such as Random Forest, XGBoost, and Logistic Regression to identify patterns and generate predictions. Random Forest and XGBoost can represent nonlinear relationships and feature interactions, whereas Logistic Regression provides a relatively interpretable baseline. Comparing these models can help establish whether increased model complexity produces meaningful predictive benefits for particular datasets and tasks (Breiman, 2001; Chen & Guestrin, 2016).

Second, the explainability layer uses SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to clarify how model features contribute to predictions. SHAP provides feature-attribution explanations based on a cooperative game-theoretic framework, while LIME approximates model behavior locally around an individual prediction (Lundberg & Lee, 2017; Ribeiro et al., 2016). These methods can help researchers and decision-makers identify influential variables, examine unexpected predictions, and investigate whether model behavior aligns with domain knowledge. However, feature attribution does not establish causality, and explanations may be sensitive to data characteristics and methodological choices.

Third, the sustainability assessment and decision optimization layers extend model evaluation beyond predictive performance. Environmental impacts, economic feasibility, social equity, and technical reliability can be represented as decision criteria. MCDA can then combine normalized indicators and explicit criterion weights to compare possible interventions or alternatives. This approach makes value judgments more visible, although the final ranking depends on the quality of the indicators, normalization procedures, and assigned weights (Hwang & Yoon, 1981; Saaty, 1980).

Finally, the evaluation and feedback layer provides a mechanism for assessing predictive performance, explanation quality, computational efficiency, robustness, and decision consistency. The proposed framework therefore treats trustworthy AI as a multidimensional evaluation problem rather than a single accuracy score. Its practical value must be confirmed through experiments that compare the integrated framework with conventional machine learning and decision-support approaches.

1.34. Comparison with Existing Studies

Existing research provides a foundation for the proposed framework, but individual contributions often address different parts of the broader problem.

Breiman (2001) introduced Random Forest as an ensemble learning method capable of handling complex predictive tasks. Chen and Guestrin (2016) developed XGBoost to improve the efficiency and scalability of gradient-boosted decision trees. These methods offer strong predictive modeling capabilities, but their use alone does not provide a complete explanation and sustainability-aware decision process.

Lundberg and Lee (2017) introduced SHAP as a unified approach to interpreting model predictions, while Ribeiro et al. (2016) proposed LIME for explaining individual predictions through locally interpretable approximations. These methods improve the accessibility of complex model behavior, but neither independently determines whether a predicted intervention is environmentally sustainable, economically viable, or socially equitable.

Similarly, MCDA methods such as the Analytic Hierarchy Process and TOPSIS provide systematic ways to compare alternatives using multiple criteria (Hwang & Yoon, 1981; Saaty, 1980). However, MCDA requires appropriate indicators and defensible weighting decisions. When used without predictive modeling, it may not fully capture complex relationships or estimate uncertain future outcomes.

		The proposed framework seeks to connect these complementary approaches within a common decision-support architecture.

- Research approach
- Main contribution
- Remaining challenge addressed by the proposed framework

Conventional machine learning

Prediction and pattern recognition

Limited transparency and incomplete sustainability assessment

Explainable AI

Interpretation of model predictions

Does not independently establish causal effects or optimal decisions

Sustainability assessment

Evaluation of environmental, economic, and social impacts

May not incorporate predictive uncertainty or model explanations

MCDA

Ranking alternatives across multiple criteria

Rankings depend on indicator quality, normalization, and weights

Proposed integrated framework

Combines prediction, explanation, sustainability assessment, and decision optimization

Requires empirical validation of predictive value, explanation reliability, robustness, and decision usefulness

The key distinction is therefore integration rather than the invention of an entirely new individual algorithm. The proposed contribution lies in connecting established methods and evaluating their combined performance using consistent technical, interpretive, sustainability, and decision-quality criteria.

To establish whether this integration offers measurable benefits, future experiments should compare the complete framework against individual models, XAI-only configurations, and decision-support systems without predictive intelligence. Performance differences should be reported using appropriate confidence intervals, effect sizes, and statistical tests rather than inferred from conceptual advantages alone.

1.35. Implications for Emerging Research Domains

The proposed framework has potential applications across emerging research domains because many contemporary challenges require both predictive analysis and decisions that account for competing objectives. Its practical implementation would need to be adapted to the data, regulations, and operational requirements of each domain.

1.35.1. Environmental Sustainability and Climate Intelligence

In environmental research, the framework could support climate-risk prediction, flood-risk assessment, air-quality monitoring, land-use analysis, biodiversity conservation, and natural-resource management. Machine learning models could estimate environmental risks using meteorological, satellite, geospatial, and sensor data, while SHAP and LIME could help explain the factors associated with individual predictions.

Sustainability indicators could then be used to compare alternative interventions according to expected environmental benefits, implementation costs, community impacts, and long-term resilience. For example, flood-management alternatives could be evaluated using predicted risk reduction, cost, ecological impact, and the number of people potentially protected.

The framework aligns conceptually with the United Nations Sustainable Development Goals, particularly Goals 6, 11, 13, and 15, which address water management, sustainable communities, climate action, and terrestrial ecosystems.

However, environmental benefits should be established through validated data and suitable evaluation methods rather than assumed from model predictions.

1.35.2. Healthcare and Biomedical Intelligence

In healthcare, the framework could support disease-risk prediction, clinical decision support, resource allocation, and health-service planning. Predictive models may identify patterns in authorized clinical or public-health datasets, while explanation methods could help clinicians examine influential features and recognize potentially unreliable predictions.

Decision optimization could compare healthcare interventions using clinical effectiveness, patient safety, cost, accessibility, and equity. Nevertheless, explanations should support rather than replace professional judgment. Clinical deployment would require appropriate validation, privacy protection, bias assessment, and compliance with applicable research and healthcare requirements.

1.35.3. Smart Infrastructure and Intelligent Systems

For intelligent infrastructure, the framework could be applied to smart grids, transportation systems, urban planning, industrial monitoring, and Internet of Things (IoT) networks. Predictive models could estimate equipment failures, congestion, energy demand, or infrastructure risks. Explanations could help engineers identify factors associated with predicted failures or inefficient operation.

Sustainability-aware decision analysis could compare maintenance schedules, energy-management strategies, and infrastructure investments using reliability, energy consumption, emissions, affordability, and service continuity. Where decisions must be made in real time, computational latency and system reliability would be important deployment constraints.

1.35.4. Agriculture, Education, and Other Emerging Fields

In precision agriculture, the framework could integrate remote sensing, weather, soil, and crop data to support crop-health monitoring, irrigation planning, and resource-efficient production. In education, it could help evaluate learning-support interventions while considering student outcomes, accessibility, fairness, and institutional cost.

Across these domains, the common contribution is a reusable decision-support structure rather than a universal model that can be transferred without modification. Each application would require domain-specific features, validated datasets, relevant sustainability indicators, and evaluation criteria. Cross-domain transfer should be tested explicitly because differences in data distributions, populations, and operational conditions can reduce model reliability.

1.36. Ethical, Social, and Environmental Considerations

The responsible application of multidisciplinary AI requires attention to ethical principles throughout data collection, model development, explanation, decision optimization, and deployment. A technically accurate system may still produce harmful outcomes if it relies on biased data, exposes sensitive information, or prioritizes narrow objectives at the expense of public welfare.

1.36.1. Data Privacy and Security

The framework may use sensitive clinical records, geospatial information, infrastructure measurements, and other potentially confidential data. Data collection and processing should follow applicable laws, institutional policies, informed-consent requirements, and data-minimization principles. Identifiable information should be protected through suitable access controls, secure storage, encryption, and de-identification where appropriate.

Data sharing between domains or institutions should occur only where authorized and necessary. Reproducibility should be supported through documented methods and permitted access to data, without exposing confidential records.

1.36.2. Fairness, Bias, and Social Inclusion

Historical datasets can reflect unequal access to healthcare, infrastructure, education, and other services. Models trained on such data may reproduce or amplify existing disparities. The proposed framework should therefore assess performance across relevant population groups and geographic areas, where lawful and ethically appropriate.

Fairness evaluation should include subgroup error rates, differences in access, and the distribution of predicted benefits and burdens. Decision-makers should also examine whether MCDA weights systematically disadvantage underserved communities. A high overall predictive score does not necessarily indicate equitable outcomes.

1.36.3. Explainability, Transparency, and Accountability

SHAP and LIME can make model behavior easier to inspect, but an explanation is not automatically complete, stable, or understandable to every stakeholder. The framework should evaluate explanation fidelity, stability, and usefulness in addition to generating feature-attribution outputs. Explanations should be presented in language appropriate to the intended users, including technical experts, policymakers, and affected communities.

Accountability should remain clearly assigned to the people and institutions responsible for deploying and using the system. Important decisions should allow appropriate human review, correction, and appeal. Explanations must not be presented as proof of causal relationships or as guarantees that a recommendation is correct.

1.36.4. Environmental Sustainability and Computational Cost

Although AI may contribute to environmental monitoring and resource optimization, model development and deployment also consume electricity, computing resources, and hardware. Large-scale training, repeated explanation generation, and continuous sensor processing may create significant computational costs.

The evaluation should therefore record training time, inference latency, memory use, and, where feasible, energy consumption or estimated carbon emissions. Researchers should compare the environmental costs of computation with the potential benefits of the resulting intervention. Lightweight models, efficient data pipelines, and appropriate deployment strategies may be preferable when they provide adequate performance.

1.36.5. Human Oversight and Responsible Governance

Human oversight should be integrated into decisions involving substantial health, environmental, financial, or social consequences. Stakeholders should be able to question recommendations, examine uncertainty, and identify situations in which automated outputs should not be used.

The framework should also maintain documentation covering dataset provenance, model versions, evaluation procedures, limitations, explanation settings, and MCDA weights. Such documentation can strengthen transparency and make decisions easier to audit. Ethical safeguards should be adapted to the relevant jurisdiction, institutional context, and level of risk rather than treated as a single checklist for every application.

1.37. Limitations and Challenges

Despite its potential, the proposed framework faces several limitations that must be addressed before its effectiveness can be established.

First, data availability and quality. Different research domains use different data formats, measurement standards, sampling procedures, and definitions of outcomes. Missing values, measurement errors, class imbalance, incomplete labels, and outdated observations can reduce predictive reliability. Combining datasets does not automatically make them comparable. Each dataset requires quality assessment, documentation, and domain-appropriate preprocessing.

Second, cross-domain generalizability. A model that performs well in one setting may not perform equally well in another because populations, environmental conditions, equipment, and data-generation processes differ. The framework should therefore be evaluated separately within each domain. Transfer learning or shared representations should only be adopted when supported by evidence, and external validation should be used where suitable data are available.

Third, limitations of explainability methods. SHAP and LIME provide useful forms of model interpretation, but their outputs can be affected by feature dependence, sampling choices, model behavior, and parameter settings. Different explanation methods may produce different descriptions of the same prediction. Furthermore, identifying an influential feature does not demonstrate that changing that feature would cause the predicted outcome to change. Explanation reliability must be assessed rather than presumed.

Fourth, subjectivity in sustainability assessment. Environmental, economic, social, and technical objectives can conflict. For example, the lowest-cost alternative may not be the most equitable or environmentally beneficial. MCDA rankings depend on indicator selection, normalization, weighting, and aggregation procedures. The framework should disclose these assumptions and perform sensitivity analysis to determine whether reasonable changes in weights or measurements alter the final rankings.

Fifth, computational overhead and scalability. Training multiple models, calculating SHAP values, generating LIME explanations, and conducting repeated robustness experiments can increase computational requirements. Large datasets and real-time applications may impose additional constraints on latency, memory, storage, and energy consumption. The implementation must therefore balance explanatory detail with computational feasibility.

Sixth, evaluation and statistical uncertainty. Results from a single dataset, train-test split, or random seed may not provide sufficient evidence of general performance. Data leakage, inappropriate metric selection, and repeated experimentation can produce misleading conclusions. The study should use suitable validation strategies, confidence intervals, effect sizes, and statistical tests. Comparisons across datasets should account for differences in domain characteristics and outcome definitions.

Seventh, implementation and adoption barriers. Effective deployment requires domain expertise, technical infrastructure, trained personnel, institutional support, and stakeholder trust. Organizations may lack the computing resources or standardized data systems needed to implement the complete framework. In

settings with limited infrastructure, a smaller implementation may be more practical than deploying all six layers simultaneously.

Finally, the absence of validated empirical results. The proposed integration provides a methodological basis for investigation, but it does not establish that the framework will outperform existing methods. Claims concerning improvements in accuracy, explanation quality, sustainability, or decision consistency must remain provisional until supported by experimental evidence. A future evaluation should use clearly defined baselines, reproducible experiments, appropriate statistical analysis, and domain-specific validation.

In summary, the proposed framework offers a structured approach to connecting predictive AI, explainability, sustainability assessment, and intelligent decision-making. Its principal strength is the integration of these functions into a common evaluation process. Its ultimate contribution, however, will depend on data quality, explanation reliability, transparent sustainability criteria, computational feasibility, and empirical evidence demonstrating that the integrated approach improves decisions in real-world research settings.

Chapter Seven: Conclusion and Future Work

1.38. Summary of Findings

This research proposed an integrated framework for supporting sustainable innovation and intelligent decision-making across emerging research domains, including environmental sustainability, healthcare, agriculture, and intelligent infrastructure. The framework addresses a central challenge in contemporary artificial intelligence: predictive performance alone is insufficient when decisions must also be transparent, reliable, sustainable, and accountable.

The proposed framework consists of six interconnected layers: data integration, predictive intelligence, explainability, sustainability assessment, decision optimization, and evaluation and feedback. Together, these layers provide a structured process for transforming heterogeneous data into predictions, interpretable explanations, sustainability assessments, and evidence-informed recommendations.

The predictive intelligence layer incorporates Random Forest, XGBoost, and Logistic Regression as comparative modeling approaches. Their evaluation is designed to consider accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUROC), and computational efficiency, as appropriate to each task. SHAP and LIME are proposed to explain model predictions, while MCDA methods support the comparison of alternatives using environmental, economic, social, and technical criteria.

The framework also proposes a multidimensional evaluation strategy incorporating predictive performance, explanation fidelity, explanation stability, robustness, computational cost, and decision consistency. Sensitivity analysis can determine whether recommendations remain stable when sustainability weights or other assumptions change.

The principal conclusion is that integrating prediction, explainability, and sustainability-aware decision analysis provides a coherent methodological approach to responsible AI-supported decision-making. However, the extent to which this integration improves predictive accuracy, interpretability, sustainability outcomes, or decision quality must be established through empirical experiments. Since validated experimental results have not yet been supplied, this study does not claim measured performance improvements.

1.39. Research Contributions

The proposed research makes five principal contributions.

1.39.1. Integrated Multidisciplinary AI Framework

The first contribution is a unified architecture connecting predictive modeling, XAI, sustainability indicators, and decision optimization. Rather than treating these components as separate analytical activities, the framework establishes a coordinated process for generating and evaluating decision-support recommendations.

1.39.2. Explainability-Centered Predictive Evaluation

The second contribution is the inclusion of explainability as an explicit evaluation dimension. By integrating SHAP and LIME with predictive models, the framework enables systematic examination of feature contributions, explanation stability, and local model behavior. This extends evaluation beyond predictive metrics while recognizing that explanations do not independently establish causality or guarantee model correctness.

1.39.3. Sustainability-Aware Decision Optimization

The third contribution is the integration of environmental, economic, social, and technical indicators into decision analysis. MCDA provides a structured mechanism for comparing alternatives and making criterion weights and trade-offs explicit. This contribution supports more transparent prioritization of interventions where multiple objectives must be considered simultaneously.

1.39.4. Multidimensional and Cross-Domain Evaluation

The fourth contribution is a proposed evaluation protocol that considers predictive performance, explanation quality, robustness, computational efficiency, and decision consistency. Applying this protocol across different research domains could help identify which framework components generalize effectively and which require domain-specific adaptation.

1.39.5. Responsible and Reproducible AI Decision Support

The fifth contribution is the integration of responsible AI considerations into the proposed design. Data governance, privacy, fairness, uncertainty, human oversight, and documentation are treated as important conditions for implementation rather than optional additions. The framework also encourages reproducible experimentation through documented datasets, preprocessing procedures, model configurations, evaluation metrics, and sensitivity analyses.

Collectively, these contributions position the framework as a methodological foundation for investigating how explainable multidisciplinary AI can support sustainable innovation. Its novelty should be understood primarily as the proposed integration and evaluation of established methods; claims of superiority or practical impact require validation against appropriate baselines.

1. Recommendations

Based on the proposed framework and its methodological requirements, the following recommendations are presented for researchers, developers, institutions, and decision-makers.

1. Establish high-quality, domain-appropriate datasets. Researchers should prioritize reliable, representative, and well-documented datasets. Data preprocessing, missing-value treatment, feature selection, class balancing, and leakage prevention should be documented to support valid comparisons.
2. Evaluate multiple predictive models. Random Forest, XGBoost, Logistic Regression, and other relevant alternatives should be compared using appropriate metrics and consistent validation procedures. Model

selection should consider predictive performance, computational cost, robustness, and suitability for the intended application.

3. Assess explanation reliability. SHAP and LIME should be evaluated for fidelity, stability, consistency, and practical usefulness. Researchers should explain the limitations of feature attribution and avoid presenting model explanations as causal evidence.

4. Make sustainability criteria explicit. Decision-support systems should define relevant environmental, economic, social, and technical indicators before ranking alternatives. The selection of indicators, normalization methods, and criterion weights should be justified and documented.

5. Conduct sensitivity and uncertainty analysis. Researchers should examine how model predictions and decision rankings change under different data conditions, model configurations, and MCDA weights. Confidence intervals and appropriate statistical tests should be used when comparing model performance.

6. Integrate responsible AI safeguards. Privacy protection, subgroup fairness assessment, data security, transparency, and human oversight should be incorporated throughout the development lifecycle. High-impact recommendations should remain subject to appropriate professional and institutional review.

7. Prioritize practical implementation and stakeholder participation. Universities, research institutions, public agencies, and industry partners should collaborate to test the framework on clearly defined problems. Domain experts and affected communities should participate in selecting decision criteria and evaluating whether the resulting recommendations are useful and equitable.

8. Measure computational and environmental costs. Implementation studies should report training time, prediction latency, memory use, and energy consumption where feasible. The computational burden of model training and explanation generation should be considered alongside the potential benefits of the supported decisions.

1.1. Future Research Directions

Future research should move from the proposed conceptual and methodological framework toward systematic implementation, empirical validation, and real-world evaluation.

1.1.1. Empirical Validation Across Multiple Domains

The first direction is to implement the framework using suitable public or authorized datasets from environmental sustainability, healthcare, agriculture, and intelligent infrastructure. Separate experiments should assess each domain before conducting cross-domain comparisons. Performance should be evaluated against conventional machine learning models, interpretable baselines, and decision-support configurations that omit selected framework components.

1.1.2. Advanced and Context-Aware Explainability

Future studies could investigate alternative explanation methods, including counterfactual explanations, rule-based explanations, and domain-specific interpretation techniques. Research should examine explanation fidelity, stability under small input changes, agreement across methods, and comprehensibility for different users. Counterfactual explanations should be constrained by domain feasibility and should not imply causal effects without suitable causal evidence.

1.1.3. Uncertainty-Aware and Risk-Sensitive Decision-Making

Future versions of the framework could incorporate predictive uncertainty, calibrated probabilities, and risk-sensitive optimization. This would allow decision-makers to distinguish high-confidence predictions from

uncertain estimates and consider the potential consequences of incorrect recommendations. Research could also investigate how uncertainty should influence sustainability rankings and when a system should defer to human judgment.

1.1.4. *Dynamic Sustainability Indicators and Digital Twins*

An important direction is the integration of real-time sensor data, geospatial information, and digital twins to support continuously updated decision-making. Such systems could represent environmental conditions, infrastructure performance, or agricultural processes and evaluate possible interventions under changing conditions. Their value would need to be assessed through temporal validation, simulation, and, where feasible, field-based experiments.

1.1.5. *Federated and Privacy-Preserving Multidisciplinary AI*

Future studies could examine federated learning and privacy-preserving techniques for situations in which sensitive data cannot be centrally collected. This is particularly relevant to healthcare institutions, public agencies, and organizations managing critical infrastructure. Research should assess the trade-offs among predictive performance, privacy, communication cost, fairness, and explanation reliability.

1.1.6. *Advanced Sustainability Optimization*

Further research could extend weighted-sum MCDA to multi-objective optimization, robust optimization, and Pareto-based analysis. These methods could help identify alternatives that balance competing objectives without assuming that one weighting scheme represents every stakeholder's priorities. Participatory weighting and sensitivity analysis could further improve transparency and legitimacy.

1.1.7. *Real-Time Deployment and Computational Efficiency*

Future implementation studies should evaluate whether the framework can operate effectively in resource-constrained and real-time environments. Potential approaches include model compression, edge AI, efficient explanation generation, and selective model execution. Evaluation should include end-to-end latency, memory requirements, energy consumption, reliability, and the quality of decisions under operational constraints.

1.1.8. *Longitudinal Evaluation and Real-World Impact*

Finally, longitudinal studies should examine whether AI-supported recommendations lead to sustained improvements in real-world outcomes. Depending on the application, outcomes could include reduced environmental risks, improved resource allocation, more reliable infrastructure, or better-supported clinical decisions. Such studies should distinguish model performance from intervention effectiveness and use appropriate experimental or quasi-experimental designs when feasible.

1.2. *Concluding Remarks*

This research proposes an integrated approach to Explainable Multidisciplinary Artificial Intelligence for Sustainable Innovation and Intelligent Decision-Making Across Emerging Research Domains. Its central premise is that effective AI decision support should combine predictive capability with understandable explanations, explicit sustainability criteria, systematic comparison of alternatives, and responsible human oversight.

The proposed framework offers a foundation for investigating these complementary requirements within a unified architecture. Its principal value lies in providing a structured methodology for evaluating not only what an AI model predicts, but also how its predictions can be interpreted, how competing sustainability

objectives can be considered, and how recommendations can be assessed for reliability and practical relevance.

The next essential step is empirical implementation and validation. Through reproducible experiments, cross-domain evaluation, sensitivity analysis, and real-world assessment, future research can determine whether the framework delivers measurable benefits and under what conditions those benefits are achieved. Ultimately, the long-term objective is to advance AI systems that are not only technically effective but also transparent, resource-conscious, socially responsible, and useful for sustainable innovation.

References

- [1] Bostrom, A., Demuth, J. L., Wirz, C. D., Cains, M. G., Schumacher, A., Madlambayan, D., Singh Bansal, A., Bearth, A., Chase, R., Ebert-Uphoff, I., Gagne, D. J., Guikema, S., Hoffman, R., Johnson, B. B., Kumler-Bonfanti, C., Lee, J. D., Lowe, A., McGovern, A., Przybylo, V., ... Williams, J. K. (2024). Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis*, 44(6), 1498–1513. <https://doi.org/10.1111/risa.14245>
- [2] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [3] 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>
- [4] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [5] Gohr, C., Rodríguez, G., Belomestnykh, S., Berg-Moelleken, D., Chauhan, N., Engler, J.-O., Heydebreck, L. V., Hintz, M. J., Kretschmer, M., Krügermeier, C., Meinberg, J., Rau, A.-L., Schwenck, C., Aoulkadi, I., Poll, S., Frank, E., Creutzig, F., Lemke, O., Maushart, M., ... von Wehrden, H. (2025). Artificial intelligence in sustainable development research. *Nature Sustainability*, 8, 970–978. <https://doi.org/10.1038/s41893-025-01598-6>
- [6] Greif, L., Röckel, F., Kimmig, A., et al. (2025). A systematic review of current AI techniques used in the context of the SDGs. *International Journal of Environmental Research*, 19, Article 1. <https://doi.org/10.1007/s41742-024-00668-5>
- [7] Hwang, C.-L., & Yoon, K. (1981). Multiple attribute decision making: Methods and applications: A state-of-the-art survey. Springer. <https://doi.org/10.1007/978-3-642-48318-9>
- [8] 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>
- [9] Kowald, D., Scher, S., Pammer-Schindler, V., Müllner, P., Waxnegger, K., Demelius, L., Fessler, A., Toller, M., Mendoza Estrada, I. G., Šimić, I., Trügler, A., Veas, E., Kern, R., Nad, T., & Kopeinik, S. (2024). Establishing and evaluating trustworthy AI: Overview and research challenges. *Frontiers in Big Data*, 7, Article 1467222. <https://doi.org/10.3389/fdata.2024.1467222>
- [10] Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 623–631). Association for Computing Machinery. <https://doi.org/10.1145/2487575.2487579>
- [11] 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://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions>
- [12] Mazhar, K., Mazhar, S. S., Khan, F. S., AlSaleh, D., Mazhar, A., Imam, S., et al. (2026). Ethical explainable artificial intelligence for green innovation: A systematic review and meta-analysis of transparency, fairness, and accountability in sustainable technologies. *Artificial Intelligence Review*, 59, Article 186. <https://doi.org/10.1007/s10462-026-11566-x>
- [13] Organisation for Economic Co-operation and Development. (2024). Explanatory memorandum on the updated OECD definition of an AI system. OECD. <https://oecd.ai/>

- [14] Organisation for Economic Co-operation and Development. (2024). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449, updated May 3, 2024). <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- [15] Pearson, R., et al. (2024). Artificial intelligence and complex sustainability policy problems: Translating promise into practice. *Sustainability Science*, 19, 308–323. <https://doi.org/10.1080/25741292.2024.2348834>
- [16] reiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [17] 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>
- [18] Saaty, T. L. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. McGraw-Hill.
- [19] Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), Article 2400304. <https://doi.org/10.1002/aisy.202400304>
- [20] 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>
- [21] Schwalbe, G., & Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38, 3043–3101. <https://doi.org/10.1007/s10618-022-00867-8>
- [22] 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>
- [23] United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development* (A/RES/70/1). <https://sdgs.un.org/2030agenda>