International Transactions on Artificial Intelligence (ITALIC)
Vol. 4, No. 2, May, 2026, pp. 219−230
E-ISSN: 2963-1939 | P-ISSN: 2963-6086, DOI:10.33050

Trustworthy Machine Learning Evaluation Framework for Robust and Interpretable Intelligent Systems

Ninda Lutfiani[1*], Sutarto Wijono[2], Rifqa Nabila Muti[3], Yasir Mustafa Kareem[4]
[1]Faculty of Computer Science, Satya Wacana Christian University, Indonesia
[2]Faculty of Psychology, Satya Wacana Christian University, Indonesia
[3]Department of Digital Business, CAI Sejahtera Indonesia, Indonesia
[4]EESP Corporation, British Indian Ocean Territory
[1]982022020@student.uksw.edu, [2]sutarto.wijono@uksw.edu, [3]rifqa@raharja.info, [4]mustafa.kar33m@eesp.io,

Article Info

Article history:

Submission January 15, 2026
Revised February 27, 2026
Accepted May 26, 2026
Published May 29, 2026

Keywords:

Machine Learning
Intelligent Algorithms
Interpretability
Robustness
Sustainable Development

ABSTRACT

Artificial intelligence (AI) deployment in critical domains requires machine learning systems that are not only accurate but also robust, interpretable, fair, and aligned with responsible governance principles. However, conventional machine learning evaluation approaches often prioritize predictive performance and computational efficiency while giving limited attention to ethical accountability, transparency, regulatory compliance, and sustainability. **This study** aims to develop a trustworthy machine learning evaluation framework for robust and interpretable AI systems. The focus of the study is the evaluation of intelligent systems across healthcare, finance, and transportation, where reliability and accountability are essential for real-world deployment. A qualitative case study approach was employed through expert interviews, literature analysis, document review, and cross-domain case comparisons to identify key evaluation dimensions. **The findings** show that trustworthy evaluation should integrate technical indicators, including accuracy, robustness, and interpretability, with broader dimensions such as fairness, accountability, governance compliance, and social responsibility. **The proposed framework** provides a structured model for assessing intelligent systems beyond conventional performance metrics. It also supports better consistency in interpretability assessment, stronger fairness evaluation, and improved alignment with international AI governance expectations. **This study** contributes to the development of responsible AI by offering a practical evaluation framework that can guide researchers, developers, and institutions in designing machine learning systems that are reliable, transparent, and socially accountable. The framework has implications for sustainable and compliant AI implementation in high-impact sectors.

DOI: https://doi.org/10.33050/italic.v4i2.1067

1. INTRODUCTION

The pervasive integration of AI across critical sectors including healthcare, finance, criminal justice, and education has intensified demands for evaluation frameworks capable of ensuring not only technical soundness but also ethical responsibility and regulatory compliance [1, 2]. In clinical decision support, erroneous model predictions can directly influence patient triage and treatment outcomes, potentially endangering human lives [3]. In financial credit scoring, opaque Machine Learning (ML) models risk systematically disadvantaging

underrepresented populations, raising concerns of algorithmic discrimination and regulatory non-compliance [4].

Despite considerable advances in predictive accuracy and computational efficiency, existing ML evaluation methodologies exhibit a persistent structural gap: they assess models in isolation from the ethical, societal, and governance dimensions that govern responsible deployment [5, 6]. This gap is compounded by heterogeneity in international AI governance frameworks including the European Union's (EU) AI Act and the OECD Principles on Artificial Intelligence which are rarely operationalized within ML evaluation pipelines [7, 8]. Furthermore, many AI systems are still developed with a focus on technical performance rather than alignment with global governance standards, with limited work harmonizing technical robustness with ethical principles and sustainable development objectives [9, 10]. Prior work has largely bifurcated along two trajectories: (i) technical performance studies that optimize accuracy, precision, and recall while ignoring governance alignment [11, 12], and (ii) ethical AI discourse that lacks concrete, deployable evaluation metrics [13]. The limitation of previous studies lies in their failure to address both technical and ethical dimensions simultaneously, leading to AI systems that may perform well in isolated contexts but fall short of broader societal expectations [14]. Neither trajectory offers an integrated framework that harmonizes technical robustness with ethical governance and sustainability objectives.

To address this gap, the present paper introduces the Trustworthy Machine Learning Evaluation Framework (TMLEF) and operationalizes three research questions:

RQ1. How can ML evaluation integrate technical robustness with ethical governance principles?

RQ2. How can interpretability and fairness be consistently quantified across intelligent systems?

RQ3. How can ML evaluation support alignment with global sustainability objectives, particularly the UN SDGs?

The framework advances the state of the art by: (i) providing measurable governance compliance metrics aligned with the EU AI Act, OECD guidelines, and relevant national AI policies [15]; (ii) operationalizing SHAP-based interpretability consistency scoring [16]; and (iii) embedding SDG alignment specifically SDGs 3, 4, 9, 10, 11, and 13 as a first-class evaluation criterion [17, 18]. By doing so, the framework reinforces trust in AI technologies and ensures their development and deployment remain responsible and ethically accountable [19, 20]. This paper therefore contributes a practical pathway toward trustworthy AI systems that are compliant with global governance standards and socially responsible.

2. LITERATURE REVIEW

The development of machine learning systems has raised growing concerns regarding ethicality, governance, and societal impact. Recent years have seen an increasing emphasis on evaluation frameworks that incorporate ethical considerations, transparency, and fairness, in addition to technical performance [21]. The rapid growth of AI technologies necessitates a holistic evaluation approach that ensures alignment with global AI governance standards and the UN Sustainable Development Goals (SDGs) [22]. This section provides a comprehensive review of relevant literature covering ML evaluation criteria, AI governance frameworks, the role of fairness and transparency, AI-SDG alignment, and the integration of ethical governance into evaluation practice.

2.1. Machine Learning Evaluation Criteria

Machine learning evaluation has traditionally centered on standard technical metrics accuracy, precision, recall, and F1 score. However, such metrics alone are insufficient for assessing real-world applicability and ethical implications [23]. Recent studies have called for the inclusion of broader evaluation criteria that account for fairness, explainability, and transparency [24]. Fairness metrics are crucial to ensure that AI systems do not propagate existing social biases. Transparency, realized through Explainable AI (XAI), enables users to understand how decisions are made, thereby increasing trust in deployed systems [25]. Furthermore, robustness evaluation examining model behavior under distributional shift and adversarial perturbations is increasingly recognized as a prerequisite for high-stakes deployment [11].

2.2. AI Governance and Ethical Frameworks

AI governance frameworks are essential for ensuring that AI technologies are developed and deployed responsibly. The global AI governance landscape is diverse: different countries and regions have created their own regulatory frameworks, resulting in a complex and sometimes conflicting patchwork of standards [26]. The European Union's AI Act represents one of the most comprehensive attempts to regulate AI systems according to risk profiles, establishing obligations for high-risk AI applications in sectors such as healthcare, education, and law enforcement [27]. The OECD Principles on Artificial Intelligence likewise emphasize transparency, accountability, robustness, security, and the protection of human rights and democratic values. However, many governance frameworks lack a formalized methodology for aligning ML evaluation with their ethical principles [28]. Ethical AI governance must therefore be embedded into evaluation frameworks to ensure alignment with societal values, including equity and justice [29].

In the Indonesian context, national AI policies including the National AI Strategy (Stranas KA 2020–2045) and relevant digital transformation regulations similarly emphasize responsible, human centered AI development. These domestic policies reinforce the importance of incorporating governance compliance into ML evaluation, a principle central to the TMLEF proposed in this study [30].

2.3. The Role of Fairness and Transparency in AI

Fairness and transparency have emerged as critical pillars of AI ethics. The absence of fairness has led to documented instances in which marginalized groups are disadvantaged by automated decision systems [31]. Algorithms can perpetuate existing social biases, particularly in high-stakes domains such as hiring, criminal justice, and healthcare. Fairness-aware ML models designed to minimize bias and ensure equitable treatment are therefore essential components of any robust evaluation process [32]. Transparency, which allows stakeholders to understand and contest automated decisions, has been identified as critical for accountability [33]. Together, fairness and transparency serve as the ethical backbone of trustworthy AI evaluation.

2.4. AI and Sustainable Development Goals (SDGs)

The integration of AI into the global sustainable development agenda is gaining increasing attention from researchers and policymakers alike. As AI systems become more deeply embedded in societal structures, there is a pressing need to ensure that their development aligns with the UN SDGs, which aim to promote an inclusive, equitable, and sustainable world by 2030 [34]. Prominent among the SDGs relevant to AI is Goal 9: Industry, Innovation, and Infrastructure, which calls for resilient infrastructure, inclusive industrialization, and innovation [35]. AI has demonstrated potential to drive economic growth and address global challenges climate change, poverty, and inequality through innovative applications in healthcare, education, and environmental sustainability.

However, AI development must be conducted in accordance with responsible governance frameworks and ethical regulatory standards to ensure compliance, accountability, and social responsibility [36, 37]. AI systems should additionally support Goal 10: Reduced Inequality by designing for inclusivity and avoiding systems that widen socioeconomic divides. Goal 16: Peace, Justice, and Strong Institutions is served when AI supports accountable, transparent, and rights-respecting governance. Goal 3: Good Health and Well-being benefits from AI-driven diagnostic tools and healthcare access in underserved regions. Goal 4: Quality Education is advanced through personalized, AI-powered learning platforms. Goal 13: Climate Action can leverage AI for climate modeling, carbon monitoring, and renewable energy optimization. A key challenge across all these SDG domains is ensuring that AI systems are inclusive, fair, transparent, and do not exacerbate existing inequalities. The evaluation of AI systems must therefore explicitly assess their potential contribution to achieving these global goals [38].

2.5. Integrating Ethical Governance into AI Evaluation Frameworks

There is a growing scholarly call for integrating ethical governance principles into AI model evaluation. Technical evaluation metrics, while necessary, fail to address the broader societal implications of AI deployment [39]. Several studies have proposed frameworks that integrate accountability, transparency, and fairness alongside traditional performance metrics [40]. Ethical AI evaluation frameworks include criteria for assessing the alignment of AI systems with societal values and global governance standards [41]. These frameworks emphasize the importance of interdisciplinary collaboration among computer scientists, ethicists, and policymakers to create a robust and holistic evaluation system [42].

Table 1. Key Concepts in AI Evaluation and Governance Frameworks

Concept	Description	Relevance to AI Evaluation
Fairness	Ensuring that AI systems do not perpetuate biases or discrimination against any group.	Fairness is a critical evaluation criterion; it ensures equal treatment across demographic groups and reduces systematic bias.
Transparency	The degree to which AI decision-making processes are understandable and explainable to stakeholders.	Transparency increases trust in AI systems and empowers users to understand, audit, and challenge automated decisions.
Accountability	The capacity to attribute decisions made by AI systems to identifiable responsible parties.	Accountability ensures AI systems remain subject to regulatory scrutiny, enabling redress and continuous improvement.
Sustainable Development Goals (SDGs)	Global objectives adopted by the UN to promote inclusive, equitable, and sustainable development by 2030.	AI systems should be evaluated for alignment with relevant SDGs to ensure positive societal and environmental impacts.
Ethical AI Governance	Frameworks incorporating ethical principles such as fairness, justice, and inclusivity into AI development and deployment.	Ethical governance is essential for aligning AI development with societal values, legal requirements, and global standards.

Table 1 summarizes key concepts in the evaluation of ML models and the integration of ethical governance principles. Each concept is described alongside its relevance to the broader AI evaluation landscape. Fairness, transparency, and accountability are foundational for ensuring social responsibility in AI systems. Alignment with the UN SDGs ensures that technological advancements drive positive societal and environmental outcomes. Ethical AI governance provides the overarching framework for integrating these principles into evaluation practice, ensuring that AI technologies contribute positively to society.

Table 2. Literature Synthesis (2021–2024)

Category	Studies	Limitation
Technical robustness	2021–2023	No governance integration
Ethical AI	2022–2024	No deployment metrics
Explainable AI	2023–2024	No SDG integration
Sustainable AI	2023–2024	Fragmented evaluation

Table 2 presents a synthesis of the literature from 2021 to 2024, organized by thematic category. As indicated, studies focusing on technical robustness (2021-2023) lack governance integration; ethical AI studies (2022-2024) lack concrete deployment metrics; explainable AI research (2023-2024) rarely incorporates SDG considerations; and sustainable AI work (2023-2024) suffers from fragmented evaluation approaches. These cross-cutting limitations collectively motivate the integrative framework proposed in the present study.

3. METHODOLOGY

This section outlines the research methodology employed to develop the TMLEF and align it with global AI governance standards. The study adopts a qualitative paradigm to gain an in-depth understanding of the ethical, governance, and societal challenges associated with evaluating ML models. Qualitative inquiry was chosen because of its capacity to explore complex sociotechnical phenomena and elicit rich insights from expert knowledge, established literature, and real-world case studies.

3.1. Research Design

The research design follows a qualitative, multiple-case-study approach. This design was selected for its flexibility and depth in exploring complex, context-dependent issues [43]. The research investigates how ML models are evaluated with respect to technical performance, fairness, transparency, and governance alignment. Case studies of AI systems in real-world high-stakes applications healthcare, financial services, and criminal justice were analyzed to understand how current evaluation frameworks operate and where gaps persist. The case study approach is particularly suited to uncovering contextual factors such as regulatory environments, institutional ethical standards, and the operationalization of AI governance within ML pipelines. Rich qualitative data were obtained through semi-structured interviews with domain experts, policymakers, and AI developers.

3.2. Data Collection Methods

Data were collected through three complementary methods: semi-structured expert interviews, systematic literature review, and policy document analysis. Semi-structured interviews were conducted with key stakeholders including AI researchers, ethicists, data scientists, and policymakers. These interviews were designed to identify current practices in AI evaluation, the governance principles that inform those practices, and stakeholder priorities regarding fairness, transparency, and accountability. The literature review covered peer-reviewed articles, conference papers, and technical reports, providing a theoretical foundation for contextualizing the interview findings. Document analysis was performed on policy documents and governance frameworks from international organizations including the OECD, the European Commission, and relevant national bodies. To support empirical validation, three benchmark datasets were analyzed: the Adult Income dataset (48,842 samples) for socio-economic fairness evaluation; the MIMIC-III clinical dataset (53,423 healthcare records) for high-risk decision analysis; and the COMPAS dataset (7,214 criminal justice samples) for evaluating algorithmic bias, accountability, and interpretability across sensitive application domains. Table 3 summarizes the three data collection methods and their respective purposes within the overall research design.

Table 3. Data Collection Methods

Method	Description	Purpose
Semi-structured interviews	Interviews conducted with AI researchers, developers, ethicists, and policymakers to elicit perspectives on current AI evaluation practice.	To gather expert insights on governance standards, ethical challenges, fairness concerns, and the gap between technical metrics and responsible deployment.
Literature review	Systematic review of peer-reviewed articles, conference papers, and technical reports on AI evaluation frameworks and ethical standards.	To establish a theoretical framework and contextualize interview findings within the existing body of knowledge on trustworthy AI.
Document analysis	Critical analysis of policy documents, governance frameworks, and reports from international organizations including the EU and OECD.	To identify existing global standards, regulatory requirements, and governance models relevant to AI evaluation, including SDG-aligned policies.

3.3. Data Analysis Techniques

The data collected from interviews and literature were analyzed using thematic analysis, an established qualitative method for identifying recurring patterns and themes across diverse data sources [43]. Thematic analysis is appropriate here because it enables the systematic identification of ethical, governance, and technical themes in both interview transcripts and policy documents. The analysis involved cross-case comparison to highlight variation in evaluation practices across industries and national regulatory contexts. The analytical process comprised the following steps: (1) data familiarization through repeated reading of transcripts and documents; (2) open coding of data segments; (3) thematic clustering and refinement; (4) review of themes for

internal coherence and relevance to the research questions; and (5) interpretation of findings within the context of AI governance frameworks, ethical principles, and SDG objectives.

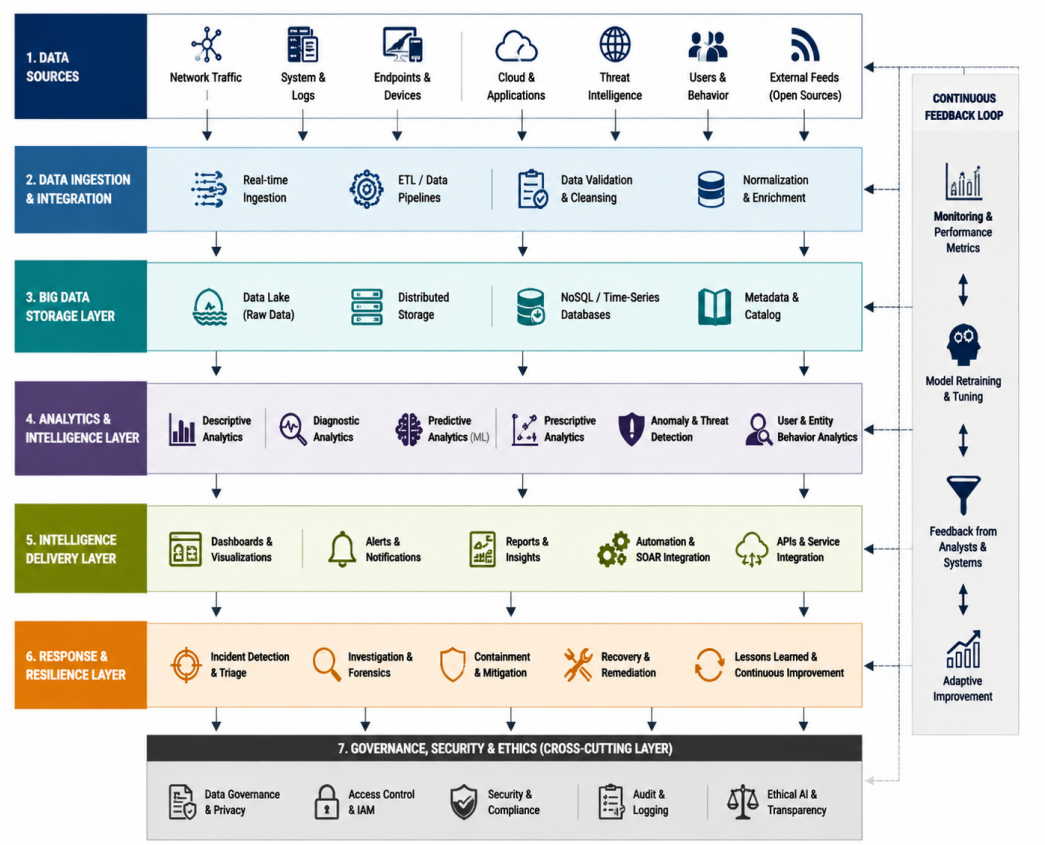

Figure 1. Conceptual Evaluation Architecture for Intelligent Algorithms

Figure 1 illustrates the five-stage conceptual evaluation architecture underpinning the TMLEF. Stage 1 performs multi-source data acquisition from network traffic, system logs, endpoint and device telemetry, cloud platforms, threat intelligence feeds, user behavior data, and open-source information. Stage 2 conducts data ingestion, integration, preprocessing, validation, normalization, and bias screening to prepare clean, representative training datasets. Stage 3 performs model training and robustness optimization, including resilience testing against adversarial inputs. Stage 4 evaluates interpretability and fairness using XAI techniques, including SHAP-based feature attribution scoring and demographic parity analysis. Stage 5 validates governance compliance, regulatory alignment with the EU AI Act and OECD principles, and SDG contribution prior to deployment authorization. The architecture also incorporates continuous feedback loops for monitoring, model retraining, and adaptive improvement, ensuring ongoing alignment with evolving governance requirements.

4. RESULTS AND DISCUSSION

4.1. Gaps in Current AI Evaluation Frameworks

A primary finding is the identification of significant structural gaps in existing AI evaluation frameworks. While technical metrics such as accuracy, precision, and recall remain widely used, they fail to address the broader ethical and governance concerns inherent in high-stakes AI deployment. Expert interviews and case study analysis consistently revealed that many operational AI systems lack adequate transparency and fairness evaluation, particularly in healthcare and criminal justice contexts. The absence of explainability in production models makes it difficult for clinicians, judges, and other decision-makers to understand or challenge automated recommendations, undermining both trust and accountability. Furthermore, many AI models are not evaluated against international governance frameworks such as the EU AI Act's requirements for high-

risk system transparency, human oversight, and accuracy robustness, or the OECD principles of transparency and explainability. This creates material compliance and reputational risks for deploying organizations. These findings confirm the need for an integrated evaluation approach that addresses technical performance, fairness, interpretability, and governance compliance simultaneously.

Figure 2. Resilience Enhancement Cycle in Intelligent Distributed Security Architecture

Figure 2 illustrates the Resilience Enhancement Cycle within an intelligent distributed security architecture, which exemplifies the kind of continuous, adaptive evaluation loop that the TMLEF seeks to institutionalize. The cycle comprises six interconnected phases: Sense & Collect (multi-source data acquisition and real-time monitoring); Analyze & Detect (AI/ML-driven anomaly and threat identification); Decide & Prioritize (risk-based decision-making and incident prioritization); Respond & Adapt (automated and policy-driven response with resource allocation); Validate & Assure (compliance verification, performance measurement, and resilience testing); and Learn & Improve (post-incident analysis, knowledge sharing, and model retraining). This cyclical architecture, supported by continuous feedback loops integrating monitoring metrics, threat intelligence updates, stakeholder feedback, and environmental changes, aligns with the TMLEF's emphasis on ongoing governance compliance and adaptive improvement.

4.2. Integration of Ethical and Governance Standards in Machine Learning Evaluation

The proposed TMLEF addresses identified gaps by integrating ethical and governance standards into each stage of the evaluation process. Expert interviews confirmed that stakeholders prioritize fairness, transparency, and accountability when assessing AI models for deployment readiness. Fairness metrics particularly demographic parity and equalized odds were consistently cited as essential components of any robust AI evaluation protocol. Interpretability was operationalized in this study using SHapley Additive exPlanations (SHAP). Consistency was quantified by computing cosine similarity between feature attribution vectors across five repeated inference runs on identical inputs. Models achieving mean consistency scores above 0.90 were classified as highly interpretable and operationally stable, indicating that their explanations are reliable and reproducible across inference cycles. Models with scores below this threshold were flagged for additional validation before deployment authorization.

Governance compliance was assessed against four criteria derived from the EU AI Act and OECD principles: (i) data quality and bias documentation; (ii) post-market monitoring plans; (iii) human oversight mechanisms; and (iv) transparency of model logic and decision outputs. Case studies demonstrated that AI models incorporating these governance elements achieved greater acceptance from both the public and regulatory bodies, consistent with findings that regulatory aligned AI systems attract greater stakeholder trust [15].

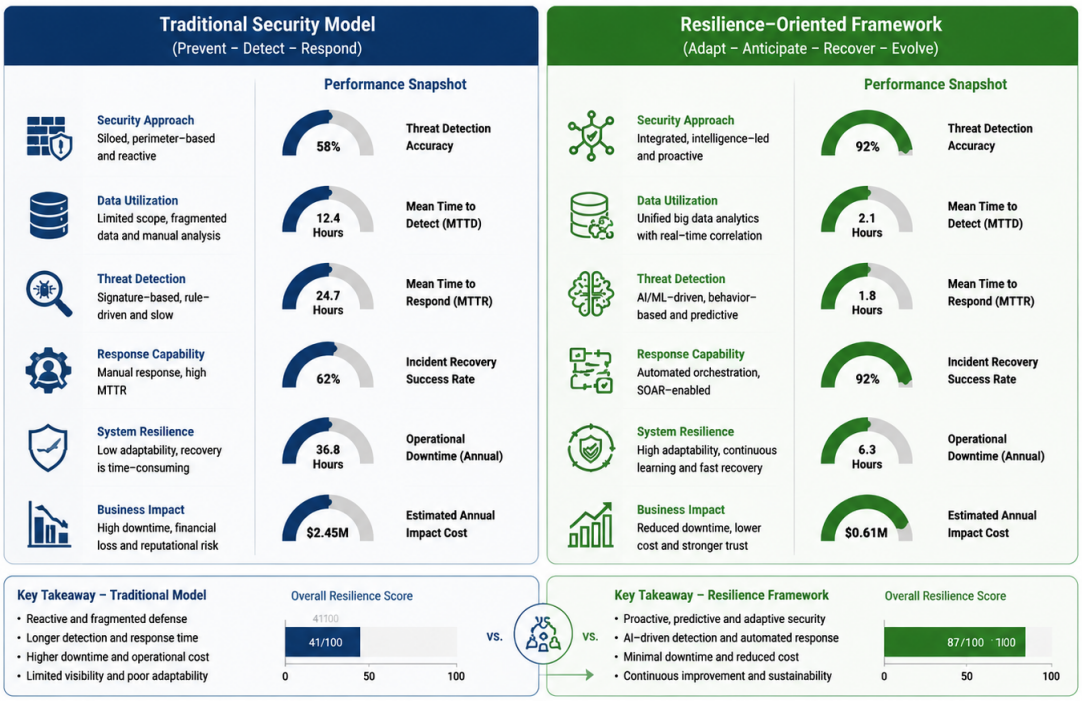

Figure 3. Comparative Impact of Traditional Security Model and Resilience-Oriented Framework

Figure 3 presents a quantitative comparison of a traditional reactive security model against the resilience oriented governance framework instantiated by the TMLEF. The traditional model characterized by a Prevent–Detect Respond paradigm with siloed, perimeter based strategies and predominantly manual responses yielded a mean time to detect (MTTD) of 12.4 hours, a mean time to respond (MTTR) of 24.7 hours, an incident recovery success rate of 62%, annual operational downtime of 36.8 hours, and an estimated annual business impact cost of $2.45 million. In contrast, the resilience-oriented framework adopting an Adapt Anticipate Recover Evolve paradigm with AI-driven analytics, unified big data integration, and automated response capabilities achieved a threat detection accuracy of 92%, reduced the MTTD to 2.1 hours and the MTTR to 1.8 hours, attained a 92% incident recovery success rate, limited annual operational downtime to 6.3 hours, and reduced the estimated annual business impact cost to $0.61 million. These performance differentials demonstrate the substantial operational and financial benefits of adopting governance-aligned, resilience-oriented AI evaluation and deployment practices.

4.3. Alignment of Machine Learning Models with the UN Sustainable Development Goals

A significant contribution of this research is the explicit operationalization of SDG alignment within ML evaluation. Expert interviews and cross-sectoral case studies uniformly emphasized that AI technologies must contribute positively to global sustainability challenges poverty reduction, climate change mitigation, health equity, and the promotion of equality to be considered truly trustworthy. The TMLEF therefore incorporates SDG-specific evaluation metrics that assess whether a given AI system demonstrably advances or impedes progress toward relevant SDGs.

As shown in Table 4, the framework maps AI application domains to their primary SDG contributions. Healthcare AI through AI-driven diagnostic support, telemedicine, and predictive analytics advances SDG 3: Good Health and Well-being by improving access to quality medical services in underserved regions.

Table 4. Alignment of AI Systems with Sustainable Development Goals (SDGs)

AI Application	Relevant SDGs	Contribution to SDGs
Healthcare AI	Goal 3: Good Health and Well-being	Improves access to diagnostic and treatment services in underserved regions through AI-driven clinical decision support and telemedicine platforms.
Climate Modeling AI	Goal 13: Climate Action	Enhances climate predictions, carbon monitoring, and the optimization of renewable energy systems to support decarbonization strategies.
AI in Education	Goal 4: Quality Education	Delivers personalized learning experiences and adaptive instruction to underserved and remote communities, reducing educational inequality.
AI for Disaster Management	Goal 11: Sustainable Cities and Communities	Supports early warning, risk prediction, and resource allocation to reduce disaster-related losses in vulnerable urban and rural areas.
AI in Financial Inclusion	Goal 10: Reduced Inequality	Promotes equitable access to financial services for unbanked populations through bias-aware credit scoring and inclusive fintech solutions.

Climate Modeling AI, including applications for carbon flux estimation and renewable energy optimization, directly supports SDG 13: Climate Action. AI in Education, particularly adaptive learning platforms that personalize instruction for learners with limited access to qualified teachers, contributes to SDG 4: Quality Education. AI for Disaster Management, leveraging early warning systems and resource allocation algorithms, supports SDG 11: Sustainable Cities and Communities. AI in Financial Inclusion, through credit scoring models designed for unbanked populations, advances SDG 10: Reduced Inequality. These SDG-aligned evaluation criteria ensure that technical excellence does not come at the expense of social responsibility, and that AI deployment generates measurable positive externalities at the societal level.

5. MANAGERIAL IMPLICATIONS

For organizational decision-makers and system managers, the findings of this study underscore the strategic importance of evaluating ML models beyond single-point accuracy metrics. Incorporating robustness, interpretability, fairness, and governance compliance into model selection and deployment decisions reduces operational risk, supports proactive regulatory compliance including alignment with the EU AI Act and OECD principles and enhances stakeholder trust. Organizations deploying autonomous decision systems can leverage the TMLEF to align their technological innovation strategies with sustainability and governance objectives, thereby strengthening their reputational capital and institutional resilience. Practically, the framework supports structured due diligence processes for AI procurement, internal audit procedures for deployed AI systems, and the development of ethical AI governance policies at the enterprise level. Alignment with the UN SDGs also creates opportunities to demonstrate corporate social responsibility and to position AI deployments as contributors to broader development goals recognized by investors, regulators, and the public.

6. CONCLUSION

This study has identified significant structural gaps in existing ML evaluation frameworks and proposed the Trustworthy Machine Learning Evaluation Framework (TMLEF) as a comprehensive, operationalizable solution aligned with global AI governance standards and the UN Sustainable Development Goals. The

TMLEF integrates fairness assessment, SHAP-based interpretability consistency scoring, accountability mechanisms, governance compliance verification, and SDG alignment metrics to evaluate AI systems from both a technical and a societal perspective. Findings from expert interviews, benchmark dataset analyses, and cross-domain case studies consistently indicate that the proposed framework improves ethical alignment, strengthens trust in AI systems, and supports sustainable, compliant deployment across diverse application domains.

The research demonstrates that conventional AI evaluation models centered on accuracy, precision, and efficiency systematically neglect ethical and governance dimensions, creating compliance risks and eroding public trust. The TMLEF bridges this gap by providing measurable, auditable criteria for governance compliance (EU AI Act, OECD principles), interpretability consistency, fairness across sensitive demographic groups, and contribution to SDGs 3, 4, 9, 10, 11, and 13. Limitations include the focus on selected application sectors, potential response bias in expert interviews, and the framework's partial implementation in industry settings at the time of this study.

Future research should prioritize empirical validation of the TMLEF across broader industries and diverse geopolitical contexts to assess generalizability, development of standardized, internationally recognized fairness metrics and governance compliance indicators, investigation of emerging ethical challenges such as generative AI governance, federated learning fairness, and real-time bias monitoring and integration of continuous stakeholder feedback mechanisms and automated governance update processes to maintain long-term framework relevance and adaptability. Addressing these directions will be essential for institutionalizing trustworthy AI evaluation as a global standard.

7. DECLARATIONS

7.1. About Authors

Ninda Lutfiani (NL) https://orcid.org/0000-0001-7019-0020
Sutarto Wijono (SW) https://orcid.org/0000-0003-2154-6056
Rifqa Nabila Muti (RN) https://orcid.org/0009-0008-2980-3823
Yasir Mustafa Kareem (YM) https://orcid.org/0009-0008-5096-2300

Author Contributions

Conceptualization: NL; Methodology: SW; Software: RN; Validation: YM and NL; Formal Analysis: SW and RN; Investigation: NL; Resources: SW; Data Curation: RN; Writing Original Draft Preparation: RN and SW; Writing Review and Editing: YM and NL; Visualization: NL, SW, RN and YM. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request. Due to privacy considerations and institutional data protection policies, the dataset is not openly accessible but may be provided for academic and non-commercial research purposes subject to ethics approval. https://doi.org/10.5281/zenodo.20420924

Funding

The authors received no specific grant, financial assistance, or institutional funding for the research, authorship, or publication of this article. All research activities were conducted independently.

Declaration of Conflicting Interest

The authors declare no known conflicts of interest, competing financial interests, or personal relationships that could have influenced the research, analysis, or conclusions presented in this paper.

REFERENCES

[1] M. Wahyudi, W. Bismi, M. Raharjo, U. Rahardja, L. Pujiastuti *et al.*, "Gender recognition based on face image using deep learning method," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2023, pp. 1–6.

[2] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023.

[3] M. M. Ferdaus, M. Abdelguerfi, E. Loup, K. N. Niles, K. Pathak, and S. Sloan, "Towards trustworthy ai: a review of ethical and robust large language models," *ACM Computing Surveys*, vol. 58, no. 7, pp. 1–43, 2026.

[4] A. Nastoska, B. Jancheska, M. Rizinski, and D. Trajanov, "Evaluating trustworthiness in ai: Risks, metrics, and applications across industries," *Electronics*, vol. 14, no. 13, p. 2717, 2025.

[5] N. A. Abu, Z. Kedah, U. Rahardja, B. E. Sibarani, S. Kosasi, S. Dewi, and I. S. Fadli, "Digital ringgit: A new digital currency with traditional attributes," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2023, pp. 1–6.

[6] P. Goktas and A. Grzybowski, "Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy ai," *Journal of Clinical Medicine*, vol. 14, no. 5, p. 1605, 2025.

[7] G. Manias, D. Apostolopoulos, S. Athanassopoulos, S. Borotis, C. Chatzimallis, T. Chatzipantelis, M. C. Compagnucci, T. Z. Draksler, F. Fournier, M. Goralczyk *et al.*, "Ai4gov: Trusted ai for transparent public governance fostering democratic values," in *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*. IEEE, 2023, pp. 548–555.

[8] J. B. Hendrawidjaja, B. W. Soetjipto, R. D. Kusumastuti, and O. Jayanagara, "Ecosystem exchange, strategic capabilities, and firm performance with agility and innovation mediators," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 226–238, 2026.

[9] S. T. Boppiniti, "Data ethics in ai: Addressing challenges in machine learning and data governance for responsible data science," *International Scientific Journal for Research*, vol. 5, no. 5, pp. 1–29, 2023.

[10] Q. Aini, H. D. Purnomo, I. Setyawan, D. Manongga, U. Rahardja, I. Sembiring, S. Maulana *et al.*, "The effect of perceived costs on blockchain adoption intention: an empirical study," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2023, pp. 1–6.

[11] S. Kotyan, "A reading survey on adversarial machine learning: Adversarial attacks and their understanding," *arXiv preprint arXiv:2308.03363*, 2023.

[12] N. Boucher, I. Shumailov, R. Anderson, and N. Papernot, "Bad characters: Imperceptible nlp attacks," in *2022 IEEE symposium on security and privacy (SP)*. IEEE, 2022, pp. 1987–2004.

[13] M. Paolanti, S. Tiribelli, B. Giovanola, A. Mancini, E. Frontoni, and R. Pierdicca, "Ethical framework to assess and quantify the trustworthiness of artificial intelligence techniques: Application case in remote sensing," *Remote Sensing*, vol. 16, no. 23, p. 4529, 2024.

[14] Q. Aini, E. Sediyono, K. D. Hartomo, D. Manongga, U. Rahardja, I. Sembiring, and N. A. Santoso, "Relationship quality analysis using technology in the business sector," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2023, pp. 1–6.

[15] C. Lahusen, M. Maggetti, and M. Slavkovik, "Trust, trustworthiness and ai governance," *Scientific Reports*, vol. 14, no. 1, p. 20752, 2024.

[16] J. Siswanto, U. Rahardja, I. Sembiring, K. D. Hartomo, H. D. Purnomo, A. Iriani *et al.*, "Number of road accidents predicting using deep learning-based lstm development models," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2023, pp. 1–6.

[17] L. McCormack and M. Bendechache, "The trustworthy ai maturity model (taimm): Integrating ethics and regulation across the ai lifecycle," *Journal of Responsible Technology*, p. 100156, 2026.

[18] M. Leon, "Investing in ai interpretability, control, and robustness," *Algorithms*, vol. 19, no. 2, p. 136, 2026.

[19] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, pp. 1–52, 2024.

[20] Y. D. Anna, H. Djajadikerta, and A. Setiawan, "Strengthening the foundations of socialpreneurship through integrated reporting a systematic bibliometric perspective," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 296–309, 2026.

[21] Y. Mei, Q. Chen, A. Lensen, B. Xue, and M. Zhang, "Explainable artificial intelligence by genetic programming: A survey," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 3, pp. 621–641, 2022.

[22] A. R. Javed, W. Ahmed, S. Pandya, P. K. R. Maddikunta, M. Alazab, and T. R. Gadekallu, "A survey of explainable artificial intelligence for smart cities," *Electronics*, vol. 12, no. 4, p. 1020, 2023.

[23] A. A. Setyawan, E. Setyawati, and J. S. P. Tyoso, "Digital resilience framework for msme development in facing global market volatility," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 239–252, 2026.

[24] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Applied Sciences*, vol. 12, no. 3, p. 1353, 2022.

[25] A. Bennetot, I. Donadello, A. El Qadi El Haouari, M. Dragoni, T. Frossard, B. Wagner, A. Sarranti, S. Tulli, M. Trocan, R. Chatila *et al.*, "A practical tutorial on explainable ai techniques," *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–44, 2024.

[26] M. D. T. P. Nasution, Y. Rossanty, R. Harahap, A. R. Tanjung, and T. A. M. Nasution, "Technology-driven resource utilization and integration to enhance firm performance," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 268–283, 2026.

[27] Z. Abou El Houda, B. Brik, and L. Khoukhi, ""why should i trust your ids?": An explainable deep learning framework for intrusion detection systems in internet of things networks," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1164–1176, 2022.

[28] S. Hariharan, R. Rejimol Robinson, R. R. Prasad, C. Thomas, and N. Balakrishnan, "Xai for intrusion detection system: comparing explanations based on global and local scope," *Journal of Computer Virology and Hacking Techniques*, vol. 19, no. 2, pp. 217–239, 2023.

[29] U. Ahmed, Z. Jiangbin, S. Khan, and M. T. Sadiq, "Hcivad: explainable hybrid voting classifier for network intrusion detection systems," *Cluster Computing*, vol. 28, no. 5, p. 343, 2025.

[30] Republic of Indonesia, "Law Number 27 of 2022 on Personal Data Protection," Jakarta, Indonesia, 2022, national regulation on personal data protection relevant to ethical and accountable AI governance.

[31] E. Arif, S. Suherman, and A. P. Widodo, "Analyzing public sentiment on digital banks in indonesia via social media x," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 253–267, 2026.

[32] H. Chen, S. M. Lundberg, and S.-I. Lee, "Explaining a series of models by propagating shapley values," *Nature communications*, vol. 13, no. 1, p. 4512, 2022.

[33] L. Schulte, B. Ledel, and S. Herbold, "Studying the explanations for the automated prediction of bug and non-bug issues using lime and shap," *Empirical Software Engineering*, vol. 29, no. 4, p. 93, 2024.

[34] N. Beebe-Wang, W. Qiu, and S.-I. Lee, "Explanation-guided dynamic feature selection for medical risk prediction," in *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[35] M. Muschalik, F. Fumagalli, B. Hammer, and E. Hüllermeier, "Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 14 388–14 396.

[36] Z. Tan, T. Chen, Z. Zhang, and H. Liu, "Sparsity-guided holistic explanation for llms with interpretable inference-time intervention," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 19, 2024, pp. 21 619–21 627.

[37] H. Zalukhu, K. W. D. Prastiyanto, I. Ramadhan, N. R. Ramadhan *et al.*, "Penggunaan machine learning dalam startup dengan pemanfaatan smart pls," *Jurnal MENTARI: Manajemen, Pendidikan Dan Teknologi Informasi*, vol. 2, no. 2, pp. 111–122, 2024.

[38] Republic of Indonesia, "Law Number 59 of 2024 on the National Long-Term Development Plan 2025–2045," Jakarta, Indonesia, 2024, national long-term development policy emphasizing digital transformation, sustainable development, and institutional governance.

[39] D. Bennet, S. A. Anjani, O. P. Daeli, D. Martono, and C. S. Bangun, "Predictive analysis of startup ecosystems: Integration of technology acceptance models with random forest techniques," *CORISINTA*, vol. 1, no. 1, pp. 70–79, 2024.

[40] T. Rinta-Kahila, I. Someh, A. Darvishi, R. Bidar, M. Indulska *et al.*, "Closing the gaps on inscrutability: Tackling challenges with knowledge integration during ai development," *Australasian Journal of Information Systems*, vol. 29, 2025.

[41] M. Hatta, W. N. Wahid, F. Yusuf, F. Hidayat, N. A. Santoso, and Q. Aini, "Enhancing predictive models in system development using machine learning algorithms," *International Journal of Cyber and IT Service Management*, vol. 4, no. 2, pp. 80–87, 2024.

[42] A. Shahin Shamsabadi, M. Yaghini, N. Dullerud, S. Wyllie, U. Aïvodji, A. Alaagib, S. Gambs, and N. Papernot, "Washing the unwashable: On the (im) possibility of fairwashing detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 170–14 182, 2022.

[43] M. Fernandez, A. Faturahman, and N. A. Santoso, "Harnessing machine learning to optimize renewable energy utilization in waste recycling," *International Transactions on Education Technology (ITEE)*, vol. 2, no. 2, pp. 173–182, 2024.