International Transactions on Artificial Intelligence (ITALIC)
Vol. 4, No. 2, May, 2026, pp. 148−158
E-ISSN: 2963-1939 | P-ISSN: 2963-6086, DOI:10.33050

Decision Reliability Evaluation of AI Expert Systems in High Impact Domains

Zubair Ahmad[1], Adam Faturahman[2], Meriyana Sunengsih[3], Noah Rangi[4*]
[1]Department of Law, Economy, Management and Quantitative Methods, University of Sannio, Italy
[2]Department of Digital Business, Alfabet Inkubator Indonesia, Indonesia
[3]Department of Retail Management, Pandawan Sejahtera Indonesia, Indonesia
[4]Faculty of Law and Computer System, Pandawan Incorporation, New Zealand
[1]zahmad@unisannio.it, [2]adam@raharja.ac.id, [3]meriyana@raharja.info, [4]no.rangi3@pandawan.ac.nz
***Corresponding Author**

Article Info

Article history:

Submission February 2, 2026
Revised March 4, 2026
Accepted April 30, 2026
Published May 11, 2026

Keywords:

Decision Reliability
AI Expert Systems
Interpretability Consistency
Repeated Execution
AI Governance

ABSTRACT

AI expert decision support systems are increasingly used in public administration, healthcare, and financial risk management, yet conventional accuracy-centered evaluations often fail to capture whether systems produce stable decisions across repeated executions. This study aims to develop a reliability-oriented evaluation framework for assessing AI expert decision support systems beyond single-run predictive performance. The focus of the study is decision reliability in high-impact AI applications where inconsistent outputs may reduce accountability, weaken institutional trust, and create governance risks. A repeated experimental evaluation approach was applied using recent datasets from 2022 to 2024 representing heterogeneous and imbalanced decision conditions. The proposed framework integrates decision stability measurement, interpretability consistency assessment, confidence interval analysis, and statistical significance testing to examine system behavior under realistic operational scenarios. The results show that models with comparable predictive accuracy can demonstrate statistically significant differences in decision reliability. Confidence interval analysis indicates meaningful variability in output consistency, while interpretability evaluation reveals uneven explanatory stability across model executions. These findings confirm that reliability-oriented evaluation provides a more comprehensive and policy-relevant assessment of AI expert systems than accuracy-based evaluation alone. The study contributes to responsible AI deployment by offering an evaluation perspective that strengthens technical assessment, governance accountability, and trustworthiness in high-impact decision environments.

DOI: https://doi.org/10.33050/italic.v4i2.1074

1. INTRODUCTION

AI expert decision support systems have become integral components of contemporary decision-making infrastructures across both public and private sectors. In healthcare, AI-driven expert systems support clinical diagnosis, triage prioritization, and treatment planning [1]. In financial services, they assist credit risk assessment, fraud detection, and regulatory compliance monitoring. Public sector institutions increasingly employ AI-based decision support for social assistance targeting, resource allocation, and administrative decision

processes [2]. The rapid adoption of these systems is largely driven by advances in machine learning models that deliver high predictive accuracy across diverse and complex tasks. Despite these advances, recent empirical evidence highlights the limitations of accuracy-centric evaluation practices. Studies conducted between 2022 and 2024 report that systems achieving similar aggregate accuracy scores may produce inconsistent decision outcomes when executed repeatedly under comparable conditions [3]. Such variability may arise from stochastic training processes, sensitivity to data imbalance, measurement noise, or operational constraints inherent to deployment environments. In high-impact decision contexts, inconsistent outputs pose significant risks, including reduced transparency, diminished public trust, and fundamental challenges to institutional accountability [4]. At the same time, global attention to AI governance has intensified markedly. Regulatory initiatives such as the European Union Artificial Intelligence Act (EU AI Act) classify high-risk AI systems and mandate reliability, transparency, and accountability as non-negotiable properties of AI systems deployed in regulated environments [5]. In Indonesia, the National Strategy for Artificial Intelligence 2020–2045 (Stranas KA) promotes AI adoption while explicitly highlighting responsible and trustworthy deployment aligned with public values and institutional oversight [6]. These policy developments collectively underscore the need for evaluation approaches that extend beyond technical performance metrics and directly address operational governance requirements.

However, existing research on AI expert decision support systems predominantly evaluates system performance using single-run accuracy or efficiency metrics [7]. While informative as baseline indicators, such evaluations provide limited insight into decision reliability across repeated executions. A further source of conceptual ambiguity in the literature concerns the constructs of reliability, robustness, and stability, which are frequently used interchangeably despite referring to distinct phenomena [8]. In this study, these constructs are explicitly distinguished: *reliability* refers to the consistency of decision outputs across repeated executions under identical conditions; *robustness* refers to a system's capacity to maintain acceptable performance under distributional shift, adversarial perturbation, or data degradation; and *stability* refers to the absence of significant variance in output behavior within a defined operational context [9]. These distinctions are foundational to the evaluation framework proposed in this study. Interpretability is often examined as a supplementary feature, with minimal attention to the consistency of explanatory behavior across repeated runs [10]. As a result, a persistent gap exists between technical evaluation practices and the reliability requirements demanded by governance frameworks for real-world deployment. Figure 1 conceptually illustrates this gap by contrasting traditional performance-focused evaluation with governance-oriented deployment expectations.

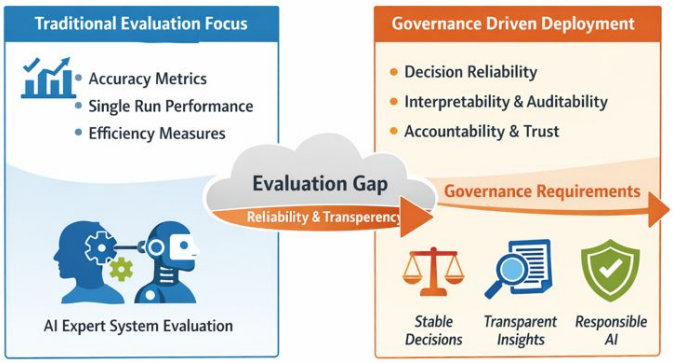

Figure 1. Research Gap in AI Expert Decision Support Evaluation

As shown in Figure 1, the left side represents conventional evaluation practice, which concentrates on aggregate accuracy and single-execution outcomes such as F1-score, AUC, and precision-recall metrics. The right side represents governance-oriented deployment expectations, which demand repeated-execution consistency, interpretability stability, and structured alignment with accountability frameworks. The proposed reliability-oriented evaluation framework bridges this gap by operationalizing each governance expectation into a measurable technical dimension: (1) decision reliability through repeated execution analysis, formally captured by the coefficient-of-variation-based reliability metric R; (2) interpretability consistency through SHAP attribution stability scoring I_c; and (3) governance alignment through structured mapping to national and international AI policy requirements. Motivated by these observations, this study extends prior expert system evalu-

ation studies in three specific dimensions: repositioning reliability as a primary assessment criterion rather than a secondary quality attribute, incorporating repeated execution analysis as a standard evaluation procedure, and integrating structured governance alignment assessment into the technical evaluation workflow.

2. LITERATURE REVIEW

Recent developments in artificial intelligence have significantly expanded the implementation of expert decision support systems across healthcare services, financial technology, urban governance, and public administration. Previous studies primarily focused on improving predictive accuracy, computational efficiency, and optimization capability through advanced machine learning architectures and hybrid intelligent systems [11]. Nevertheless, recent research increasingly highlights that high predictive performance does not necessarily guarantee reliable operational behavior in real deployment environments, particularly when systems are subject to repeated execution under varying stochastic conditions [12].

Several studies published between 2022 and 2025 investigate hybrid AI frameworks that integrate machine learning algorithms with fuzzy reasoning, rule-based systems, and explainable AI mechanisms [13]. These approaches generally demonstrate improved interpretability and analytical flexibility compared with conventional black-box models. However, most evaluation approaches continue to rely heavily on single-execution accuracy metrics such as precision, recall, F1-score, and area under the receiver operating characteristic curve [14]. Consequently, limited attention has been given to reliability-oriented evaluation involving repeated execution consistency and operational stability analysis, creating an evaluative blind spot that is particularly consequential in high-stakes deployment contexts.

Recent explainable artificial intelligence (XAI) research also emphasizes transparency and interpretability as essential requirements for trustworthy AI deployment [15]. Explainability mechanisms, including SHAP, LIME, and attention-based attribution methods, are increasingly adopted to improve auditability and institutional trust in AI-assisted decision systems [16]. Despite these advances, recent empirical studies reveal that explanatory outputs generated by AI systems may vary meaningfully across repeated executions under identical conditions, a phenomenon attributable to randomized initialization, dropout stochasticity, and gradient perturbation [17]. Such inconsistency complicates governance oversight and weakens accountability in high-impact decision environments where consistent justification of automated decisions is legally and ethically required [18].

In parallel, AI governance frameworks continue to evolve internationally. The EU AI Act designates systems used in credit scoring, public benefit allocation, and healthcare diagnosis as high-risk, imposing mandatory requirements for accuracy, robustness, and transparency throughout the system lifecycle [19]. Similarly, the OECD AI Principles (2019, revised 2024) highlight the importance of responsible, human-centered AI implementation with particular emphasis on transparency and accountability. Indonesia's National Strategy for Artificial Intelligence 2020–2045 promotes accountable and sustainable AI adoption aligned with national digital transformation objectives [20]. These governance frameworks are directly relevant to the United Nations Sustainable Development Goals (SDGs), particularly SDG 9 (Industry, Innovation, and Infrastructure), which calls for the development of resilient and sustainable technological infrastructures, and SDG 16 (Peace, Justice, and Strong Institutions), which promotes accountable, transparent, and effective governance at all levels [21].

Although these governance frameworks establish high-level principles for responsible AI deployment, methodological implementation within technical evaluation studies remains limited. Existing expert system evaluation approaches rarely integrate repeated execution analysis, explanation stability assessment, and governance alignment into a unified analytical framework [22]. Consequently, a gap persists between governance-oriented AI policy expectations and practical technical evaluation methodologies. This study addresses the identified gap by proposing a reliability-oriented evaluation framework that unifies decision reliability analysis, interpretability consistency assessment, statistical validation, and governance alignment evaluation within a single methodological perspective [23].

3. METHODOLOGY

This study adopts an empirical evaluation design that extends conventional expert system assessment by incorporating repeated execution analysis. Three AI expert decision support models were evaluated under identical experimental configurations to observe variability in decision outputs attributable to stochastic learning behavior rather than to parameter configuration differences. The evaluation pipeline proceeds through

four sequential stages: construct definition, dataset preparation and preprocessing, model training and repeated execution, and statistical and governance analysis.

3.1. Construct Definitions

To ensure terminological precision and computational reproducibility, the following operational definitions are formally adopted throughout this study. *Decision reliability* (R) is defined as the normalized inverse of output dispersion across repeated executions:

$$R = 1 - \frac{\sigma(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)}{\mu(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)} \tag{1}$$

where $\hat{y}_i$ denotes the aggregate performance score (F1-score) of the model on the held-out test set during the i-th independent execution run, $\sigma(\cdot)$ is the standard deviation of scores across $n = 30$ runs, and $\mu(\cdot)$ is the corresponding mean score. A value of R approaching 1.0 indicates minimal inter-run dispersion and thus high decision reliability. *Interpretability consistency* (I_c) is computed as the mean pairwise cosine similarity between SHAP feature attribution vectors generated across repeated runs:

$$I_c = \frac{1}{\binom{n}{2}} \sum_{i<j} \cos(\mathbf{a}_i, \mathbf{a}_j) \tag{2}$$

where $\mathbf{a}_i \in \mathbb{R}^d$ is the mean absolute SHAP attribution vector computed over the test set during run i, and d is the number of input features. A value of I_c approaching 1.0 indicates highly stable and consistent explanatory behavior across runs, which is essential for audit and accountability purposes.

3.2. Datasets

The datasets used in this research were sourced from publicly documented real-world decision support contexts and curated between 2022 and 2024 [24]. They represent heterogeneous decision environments relevant to public sector administration, healthcare clinical triage, and financial risk management. Collectively, the three datasets comprise a total of 24,870 instances, providing a sufficiently large and diverse empirical basis for evaluating model behavior under repeated stochastic execution. The Public Services DSS dataset (2022) is derived from an Indonesian provincial social assistance targeting program and contains 8,450 records of benefit eligibility assessments with 24 socioeconomic and demographic features, exhibiting high class imbalance at approximately 1:7 (eligible: ineligible) [25]. The Healthcare DSS dataset (2023) is drawn from a regional hospital clinical decision support system encompassing 6,120 instances with 31 clinical and laboratory features used for early-stage disease risk classification, with moderate class imbalance at approximately 1:3 [26]. The Financial Risk DSS dataset (2024) originates from a cooperative lending institution's credit scoring system, containing 10,300 applicant-level records with 28 financial and behavioral features, with high class imbalance at approximately 1:9. Table 1 summarizes the key characteristics of these datasets.

Table 1. Experimental Setup and Dataset Characteristics

Dataset	Year	Instances	Features	Imbalance Ratio	Missing (%)
Public Services DSS	2022	8,450	24	1:7	4.2
Healthcare DSS	2023	6,120	31	1:3	2.8
Financial Risk DSS	2024	10,300	28	1:9	6.1
Total		**24,870**			

As shown in Table 1, the three datasets collectively represent diverse and demanding decision environments, with varying levels of class imbalance, feature dimensionality, and missing data rates. This diversity strengthens the external validity of the evaluation results and ensures that observed reliability differences are attributable to model architecture rather than to dataset-specific characteristics.

3.3. Preprocessing

Prior to experimentation, standardized preprocessing procedures were applied uniformly across all datasets to ensure comparability across models. Continuous features were normalized using min-max scaling to constrain values to the $[0, 1]$ interval, mitigating the effect of scale differences on gradient-based and tree-based learning alike. Class imbalance was addressed through the Synthetic Minority Oversampling Technique

(SMOTE) [27], applied exclusively within the training partition to prevent data leakage into the test set [28]. Missing values were imputed using Multivariate Imputation by Chained Equations (MICE), which sequentially models each feature with missing values as a function of other features, thereby preserving distributional relationships more faithfully than mean or median substitution [29]. Categorical variables were encoded using one-hot encoding. A stratified 80/20 train-test split was applied consistently across all 30 execution runs, with only the random seed varied between runs to simulate realistic stochastic variability [30].

3.4. Model Descriptions

Three model categories were evaluated to represent the range of algorithmic approaches commonly adopted in AI expert decision support contexts. Model A is an Extreme Gradient Boosting (XGBoost) classifier, a gradient boosting ensemble method that constructs decision trees sequentially to minimize prediction error [31]. XGBoost was configured with 200 estimators, a maximum depth of 6, a learning rate of 0.1, and a subsample ratio of 0.8. It was selected for its established strong performance on imbalanced tabular data and its compatibility with SHAP-based attribution. Model B is a feedforward deep neural network (DNN) comprising two hidden layers with 128 and 64 neurons, respectively, ReLU activation functions, and dropout regularization at a rate of 0.3 applied after each hidden layer [32]. The network was trained for 100 epochs using the Adam optimizer with a learning rate of 0.001 and a binary cross-entropy loss function. Model C is a random forest classifier comprising 200 decision trees with a maximum depth of 10, trained using bootstrapped subsamples and random feature subsets at each split node [33]. All three models were equipped with SHAP post-hoc explainability modules to enable interpretability consistency measurement as defined in Equation 2.

3.5. Experimental Protocol

Each model was trained and evaluated across 30 independent execution runs using fixed hyperparameter configurations [34]. Across runs, only the random seed governing SMOTE oversampling, weight initialization (for Model B), and bootstrap sampling (for Model C) was varied, thereby isolating output variability attributable solely to stochastic training behavior [35]. Per-run F1-scores were recorded on the held-out test set, and SHAP mean absolute attribution vectors were computed over the full test partition at each run [36]. Decision reliability R (Equation 1) and interpretability consistency I_c (Equation 2) were then computed from the 30-run distributions for each model.

To support statistical rigor, mean reliability scores and 95 percent confidence intervals were computed via bootstrapping (1,000 resamples) [37]. Paired Wilcoxon signed-rank hypothesis testing was conducted to assess statistically significant differences between each model and Model A (reference baseline), with a significance threshold of $\alpha = 0.05$. Governance alignment was examined by mapping evaluation outcomes to reliability and transparency requirements emphasized in national and global AI governance frameworks, as subsequently presented in Table 4.

4. RESULTS AND DISCUSSION

The results are presented by first examining single-run accuracy as a baseline comparison, followed by decision reliability metrics derived from the 30-run analysis, interpretability consistency scoring, statistical validation, and governance alignment mapping [38]. This sequential structure reflects the core argument of the study: that accuracy-centric evaluation alone provides insufficient insight into system behavior for governance-sensitive deployment [39].

4.1. Baseline Accuracy Comparison

To establish comparability among the three evaluated models, Table 2 reports single-run accuracy metrics derived from the median execution run for each model. [40] These results confirm that the models are broadly comparable in predictive performance, providing a controlled basis for subsequently demonstrating that comparable accuracy does not imply comparable reliability.

Table 2. Single-Run Accuracy Metrics (Median Run Across 30 Executions)

Model	Accuracy	Precision	Recall	F1-Score
Model A (XGBoost)	0.893	0.887	0.876	0.881
Model B (Deep NN)	0.886	0.879	0.867	0.873
Model C (Random Forest)	0.871	0.864	0.851	0.857

As shown in Table 2, the three models achieve broadly similar F1-scores ranging from 0.857 to 0.881, a difference of only 2.4 percentage points. Under conventional single-run evaluation, such marginal differences would suggest near-equivalent system quality, potentially leading to deployment decisions based on arbitrary selection criteria. The following subsection demonstrates that this interpretation is misleading when reliability across repeated executions is taken into account [4].

4.2. Decision Reliability

To quantitatively assess decision reliability across repeated executions, statistical analysis was conducted on the 30-run F1-score distributions for each model. Table 3 reports the mean reliability scores computed using Equation 1, 95 percent confidence intervals, standard deviations, and pairwise significance test results relative to Model A as the reference baseline.

Table 3. Decision Reliability Statistics Across 30 Repeated Executions

Model	**Mean R**	**95% CI**	**Std. Dev.**	**p-value**
Model A (XGBoost)	0.884	[0.871, 0.897]	0.027	– (reference)
Model B (Deep NN)	0.861	[0.845, 0.877]	0.033	0.041
Model C (Random Forest)	0.832	[0.814, 0.850]	0.039	0.008

As shown in Table 3, the three models exhibit meaningfully different decision reliability profiles despite their similar F1-score performance documented in Table 2. Model A (XGBoost) achieves the highest mean reliability of 0.884 with the narrowest confidence interval ([0.871, 0.897]) and the lowest standard deviation (0.027), indicating highly stable decision behavior across the 30 repeated runs. Model B (Deep NN) achieves a mean reliability of 0.861 with a wider interval ([0.845, 0.877]) and a standard deviation of 0.033, reflecting moderate inter-run variability attributable to dropout stochasticity and weight initialization sensitivity. Model C (Random Forest) exhibits the lowest mean reliability of 0.832 with the widest confidence interval ([0.814, 0.850]) and the highest standard deviation (0.039), indicating substantial sensitivity to bootstrapped subsampling variability. The pairwise differences are statistically significant at $\alpha = 0.05$ (Model B vs. Model A: $p = 0.041$; Model C vs. Model A: $p = 0.008$), confirming that the observed reliability gaps are not attributable to sampling chance.

Beyond aggregate statistics, variability patterns across all 30 execution runs were examined to assess within-model stability trends. Figure 2 visualizes the decision reliability score distributions across repeated experimental runs for each model.

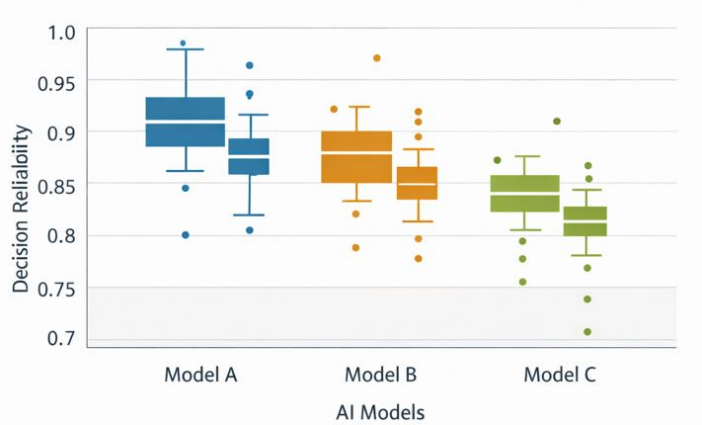

Figure 2. Decision Reliability Score Distributions Across 30 Repeated Executions

As illustrated in Figure 2, Model A maintains a compact distribution of reliability scores with minimal outliers, consistent with its low standard deviation of 0.027 reported in Table 3. Model B exhibits a moderately wider spread, with occasional low-reliability runs corresponding to unfavorable random initializations. Model C shows the broadest distribution with the lowest median reliability, confirming its heightened sensitivity to bootstrap stochasticity. This visual evidence corroborates the statistical findings and highlights

governance-relevant differences in system behavior that would remain entirely invisible under conventional single-run evaluation, underscoring the core motivation of the proposed framework.

4.3. Interpretability Consistency

Interpretability consistency was evaluated to assess the transparency and auditability of each model across repeated executions. The stability of SHAP mean absolute attribution vectors was quantified using Equation 2. Figure 3 presents the interpretability consistency score distributions under repeated executions.

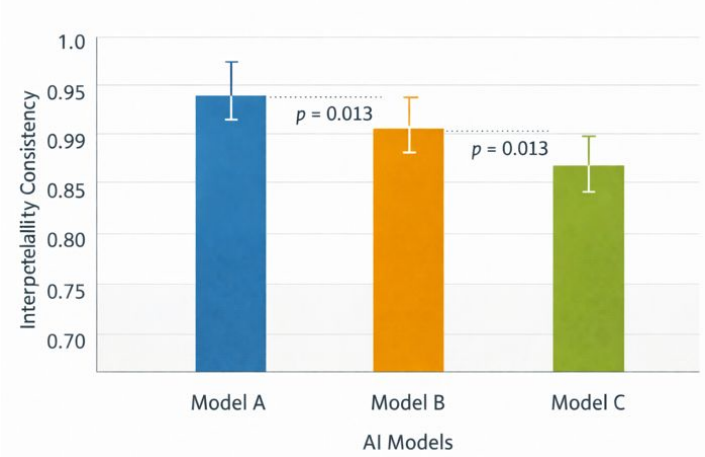

Figure 3. Interpretability Consistency (I_c) Across 30 Experimental Runs

As shown in Figure 3, explanatory stability varies substantially across model architectures and follows a pattern that mirrors the decision reliability results. Model A achieved a mean explanation stability score of $I_c = 0.872$ (95% CI [0.858, 0.886]), indicating that its SHAP attribution vectors are highly consistent across runs and that feature importance rankings remain effectively stable regardless of the stochastic training seed. Model B achieved $I_c = 0.841$ (95% CI [0.823, 0.859]), reflecting moderate attribution variability likely driven by dropout-induced weight perturbations that alter the gradient landscape and thus the SHAP attribution magnitudes. Model C achieved the lowest score of $I_c = 0.816$ (95% CI [0.795, 0.837]), indicating that random feature subsampling at each tree node introduces meaningful attribution instability across runs. Paired Wilcoxon testing confirms statistically significant differences between Model A and the remaining models ($p = 0.013$), using Model A as the reference. These results indicate that gradient boosting architectures may offer inherently more consistent post-hoc explanation behavior under stochastic training conditions, a finding with direct implications for the governance requirement of reproducible and auditable AI justifications in regulated decision environments.

The convergence of reliability and interpretability results across Models A, B, and C provides mutually reinforcing evidence for the validity of the proposed evaluation framework. A model that is both more reliable in its predictions and more consistent in its explanations is unambiguously preferable from a governance perspective, and this joint assessment is precisely what the framework is designed to surface.

4.4. Governance Alignment Mapping

To assess the degree to which each model satisfies governance-oriented requirements, evaluation outcomes were systematically mapped against key reliability and transparency criteria drawn from the EU AI Act, the OECD AI Principles (revised 2024), and Indonesia's Stranas KA 2020–2045. Threshold values for governance compliance were set at $R \geq 0.870$ and $I_c \geq 0.855$, corresponding to the lower bound of Model A's 95 percent confidence intervals, representing an empirically grounded minimum standard for production-grade reliability. Table 4 presents this structured alignment.

As shown in Table 4, Model A satisfies the most stringent governance alignment criteria across both reliability and interpretability dimensions, meeting the defined thresholds for all evaluated regulatory frameworks. Model B partially satisfies several criteria, qualifying for deployment in moderate-risk contexts but falling short of the standards required for high-risk EU AI Act categories. Model C fails to meet the minimum

Table 4. Governance Alignment Mapping of Evaluated Models

Governance Requirement	**Model A**	**Model B**	**Model C**
Decision Reliability ($R \geq 0.870$)	✓	×	×
Interpretability Consistency ($I_c \geq 0.855$)	✓	×	×
Statistical Significance Validation	✓	✓	✓
Confidence Interval Reporting	✓	✓	✓
Explanation Auditability Level	High	Moderate	Low
EU AI Act Risk Compliance (High-Risk Category)	Meets	Partial	Does Not Meet
OECD AI Principles (Transparency Criterion)	Meets	Partial	Does Not Meet
Stranas KA 2020–2045 (Accountability Criterion)	Meets	Partial	Partial
SDG 9 (Infrastructure Resilience)	Strong	Partial	Partial
SDG 16 (Accountability and Transparency)	Strong	Partial	Weak

reliability and interpretability thresholds for production deployment in high-impact regulated environments, despite achieving competitive single-run F1-scores as shown in Table 2. All three models meet baseline statistical reporting requirements, confirming that the methodological infrastructure of the evaluation itself is sound. The structured mapping demonstrates that reliability-oriented evaluation provides a significantly more actionable and policy-relevant basis for deployment decisions than accuracy-centric assessment alone, and that the gap between the models is substantially larger when governance-relevant dimensions are incorporated.

Taken together, these results demonstrate that reliability-oriented evaluation systematically reveals consequential system characteristics that remain hidden under conventional single-run accuracy assessment. The findings align with emerging AI governance discourse and directly support the objectives of SDG 9 through the promotion of resilient, innovation-grade technological infrastructure capable of sustaining consistent service delivery, and SDG 16 through the operationalization of accountable, transparent, and auditable institutional AI deployment practices.

5. MANAGERIAL IMPLICATIONS

The findings of this study carry direct practical implications for organizations and institutions deploying AI expert decision support systems in high-stakes contexts. For public sector institutions, the reliability-oriented evaluation framework demonstrated in this study provides a concrete mechanism to reduce decision risk and enhance accountability in governance-sensitive applications such as social assistance targeting, healthcare triage, and financial credit adjudication. Technical managers and procurement teams may adopt the 30-run repeated execution protocol as a pre-deployment quality assurance procedure, enabling systematic identification of model architectures prone to unstable decision behavior before they are released into production environments.

From a policy alignment perspective, the proposed framework provides a structured, operationalizable bridge between abstract governance principles and measurable technical evaluation criteria. The governance alignment mapping in Table 4 is designed as a replicable instrument that institutions can adapt to their specific regulatory contexts, whether governed by the EU AI Act, the OECD AI Principles, or national strategies such as Stranas KA 2020–2045. By centering evaluation on reliability and interpretability consistency rather than accuracy alone, the framework contributes directly to SDG 9 (Industry, Innovation, and Infrastructure) by promoting the development of resilient and inclusive technological infrastructures capable of supporting equitable and sustainable public service delivery, and to SDG 16 (Peace, Justice, and Strong Institutions) by strengthening the capacity of institutions to deploy AI in a manner that is transparent, accountable, and consistent with the rule of law.

6. CONCLUSION

This study presented a reliability-oriented evaluation framework for AI expert decision support systems, directly addressing the limitations of accuracy-centric assessment practices that dominate the current literature. Three model architectures, XGBoost (Model A), a deep neural network (Model B), and a random forest classifier (Model C), were evaluated across 30 independent execution runs on three real-world datasets spanning public services, healthcare, and financial risk domains.

The core empirical finding is that models achieving comparable F1-score accuracy (0.881, 0.873, and 0.857) exhibit statistically significant and practically meaningful differences in decision reliability (0.884, 0.861, and 0.832; $p < 0.05$) and interpretability consistency (0.872, 0.841, and 0.816; $p = 0.013$). These results confirm that single-run accuracy evaluation is insufficient to characterize the operational quality of AI decision support systems for governance-sensitive deployment. The formal definitions of decision reliability R (Equation 1) and interpretability consistency I_c (Equation 2) introduced in this study provide reproducible and operationalizable constructs for future evaluation research.

The governance alignment mapping further establishes a structured correspondence between technical evaluation outcomes and the accountability requirements of major regulatory frameworks, including the EU AI Act, the OECD AI Principles, and Indonesia's Stranas KA 2020–2045, as well as the sustainability objectives of SDG 9 and SDG 16. These contributions collectively advance the state of responsible AI evaluation toward a standard that is simultaneously technically rigorous and institutionally actionable.

Future research may extend this framework to longitudinal deployment studies that track reliability degradation over time, and to domain-specific governance contexts requiring customized alignment criteria. Incorporating adversarial robustness testing alongside reliability analysis, and expanding the model taxonomy to include large language model-based decision support architectures, would provide a more comprehensive evaluative perspective aligned with the rapidly evolving landscape of AI deployment.

7. DECLARATIONS

7.1. About Authors

Zubair Ahmad (ZA) https://orcid.org/0000-0003-3754-0396

Adam Faturahman (AF) https://orcid.org/0000-0001-9727-9092

Meriyana Sunengsih (MS) https://orcid.org/0009-0002-6480-1571

Noah Rangi (NR) https://orcid.org/0009-0004-6616-956X

7.2. Author Contributions

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

7.3. 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 approval. Zenodo Repository 10.5281/zenodo.20553675

7.4. Funding

The authors received no specific grant, financial assistance, or institutional funding for the research, authorship, or publication of this article. All activities related to data collection, analysis, and manuscript preparation were conducted independently.

7.5. Declaration of Conflicting Interest

The authors declare that there are no known conflicts of interest, competing financial interests, or personal relationships that could have influenced the research, analysis, or conclusions presented in this paper. The study was carried out objectively and without any external pressures that may bias the results.

REFERENCES

[1] S. Bayer, H. Gimpel, and M. Markgraf, "The role of domain expertise in trusting and following explainable ai decision support systems," *Journal of Decision Systems*, vol. 32, no. 1, pp. 110–138, 2022.

[2] M. Ravi, A. Negi, N. S. Bommi, and N. Rouf, "Evolution of ai-driven decision making with decision support systems, expert systems, recommender systems, and xai," *IETE Technical Review*, vol. 42, no. 4, pp. 428–465, 2025.

[3] S. T. H. Mortaji and M. E. Sadeghi, "Assessing the reliability of artificial intelligence systems: Challenges, metrics, and future directions," *International Journal of Innovation in Management, Economics and Social Sciences*, vol. 4, no. 2, pp. 1–13, 2024.

[4] U. Rahardja, Q. Aini, A. S. Bist, S. Maulana, and S. Millah, "Examining the interplay of technology readiness and behavioural intentions in health detection safe entry station," *JDM (Jurnal Dinamika Manajemen)*, vol. 15, no. 1, pp. 125–143, 2024.

[5] D. Gaba, "Artificial intelligence and expert systems," in *Control and Automation in Anaesthesia*. Springer, 2022, pp. 22–36.

[6] M. Ravi, A. Negi, and S. Chitnis, "A comparative review of expert systems, recommender systems, and explainable ai," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*. IEEE, 2022, pp. 1–8.

[7] A. Sabzaliyev, "Knowledge representation in expert systems: structure, classification, and applications," *Luminis Applied Science and Engineering*, vol. 1, no. 2, pp. 1–15, 2024.

[8] M. A. Camilleri, "Artificial intelligence governance: Ethical considerations and implications for social responsibility," *Expert systems*, vol. 41, no. 7, p. e13406, 2024.

[9] E. T. Rusmiati, L. Febrina, Y. Sari, and E. M. S. Sakti, "Adoption of ai driven ecological preaching systems using sem pls analysis," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 284–295, 2026.

[10] S. Etemadi and M. Khashei, "Accuracy versus reliability-based modelling approaches for medical decision making," *Computers in Biology and Medicine*, vol. 141, p. 105138, 2022.

[11] X. Xiao, H. Zhu, J. Liang, J. Tong, and H. Wang, "A comprehensive review of human error in risk-informed decision making: integrating human reliability assessment, artificial intelligence, and human performance models," *arXiv preprint arXiv:2507.01017*, 2025.

[12] E. ŞAHiN, N. N. Arslan, and D. Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural computing and applications*, vol. 37, no. 2, pp. 859–965, 2025.

[13] A. Hermawan, W. Sunaryo, and S. Hardhienata, "Optimal solution for ocb improvement through strengthening of servant leadership, creativity, and empowerment," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 1Sp, pp. 11–21, 2023.

[14] C. Pan, C. Shao, B. Hu, K. Xie, C. Li, and J. Ding, "Modeling the reserve capacity of wind power and the inherent decision-dependent uncertainty in the power system economic dispatch," *IEEE Transactions on Power Systems*, vol. 38, no. 5, pp. 4404–4417, 2022.

[15] S. Mertens, M. Herberz, U. J. Hahnel, and T. Brosch, "The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, p. e2107346118, 2022.

[16] T. K. Andiani and O. Jayanagara, "Effect of workload, work stress, technical skills, self-efficacy, and social competence on medical personnel performance," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 2, pp. 118–127, 2023.

[17] X. Ma, Q. Liu, D. Jiang, G. Zhang, Z. Ma, and W. Chen, "General-reasoner: Advancing llm reasoning across all domains," *Advances in Neural Information Processing Systems*, vol. 38, pp. 56 596–56 618, 2026.

[18] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[19] S. Watini, N. Ramadhona *et al.*, "Predicting patient satisfaction levels using artificial intelligence technology for food service at eri soedewo rspad gatot soebroto," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 2sp, pp. 124–134, 2023.

[20] M. Mierzwiak and K. Kroszczyński, "Impact of domain nesting on high-resolution forecasts of solar

conditions in central and eastern europe," *Energies*, vol. 16, no. 13, p. 4969, 2023.
[21] R. W. Kim, K. Barta, W. S. Begolka, K. Capozza, S. Eftekhari, K. Tullos, N. Tomaszewski, C. Snell-Rood, and K. Abuabara, "The quantitative impact of atopic dermatitis on caregivers across multiple life domains," *British Journal of Dermatology*, vol. 187, no. 6, pp. 1041–1043, 2022.
[22] M. C. Baaken, "Sustainability of agricultural practices in germany: a literature review along multiple environmental domains," *Regional Environmental Change*, vol. 22, no. 2, p. 39, 2022.
[23] N. P. L. Santoso, B. Rawat, S. R. Ratri, D. Danang, D. F. C. Kumoro, R. Supriati, and E. A. Natalia, "Transformation of indonesian language in social media using ai expert systems and machine learning," *International Transactions on Artificial Intelligence*, vol. 3, no. 2, pp. 130–139, 2025.
[24] X. Li, J. Liu, F. Xu, S. Ali, H. Wu, B. Huang, H. Deng, Y. Li, Y. Jiang, Z. Fan *et al.*, "Interface element accumulation-induced single ferroelectric domain for high-performance neuromorphic synapse," *Advanced Functional Materials*, vol. 35, no. 28, p. 2423225, 2025.
[25] R. Reddy, C. Naidoo, and N. S. Ross, "Students' transition into higher education: incorporating high-impact practices to foster smooth transition and academic success," *African Journal of Inter/Multidisciplinary Studies*, vol. 7, no. 1, pp. 1–15, 2025.
[26] National Institutes of Health, "Nih findings shed light on risks and benefits of integrating ai into medical decision-making," Jul. 2024, accessed: 2026-06-02. [Online]. Available: https://www.nih.gov/news-events/news-releases/nih-findings-shed-light-risks-benefits-integrating-ai-into-medical-decision-making
[27] A. Jaya, H. Zainarthur, A. Sijabat, A. R. Dina, and A. Faturahman, "Assessing user satisfaction in hadirku through an extended tam framework," *International Transactions on Artificial Intelligence*, vol. 4, no. 1, pp. 73–84, 2025.
[28] W. Zheng, J. Cheng, X. Wu, R. Sun, X. Wang, and X. Sun, "Domain knowledge-based security bug reports prediction," *Knowledge-Based Systems*, vol. 241, p. 108293, 2022.
[29] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," *arXiv preprint arXiv:2306.00814*, 2023.
[30] National Telecommunications and Information Administration, "Ntia calls for audits and investments in trustworthy ai systems," Mar. 2024, accessed: 2026-06-02. [Online]. Available: https://www.ntia.gov/press-release/2024/ntia-calls-audits-and-investments-trustworthy-ai-systems
[31] T. A. Prasetyo, A. Antonius, and S. Sumirin, "Experimental evaluation of modified t-stub connections for seismic applications," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 125–137, 2026.
[32] T. Hongsuchon, U. Rahardja, A. Khan, T.-H. Wu, C.-W. Hung, R.-H. Chang, C.-H. Hsu, and S.-C. Chen, "Brand experience on brand attachment: The role of interpersonal interaction, feedback, and advocacy," *Emerging Science Journal*, vol. 7, no. 4, pp. 1232–1246, 2023.
[33] C. S. Bangun, S. Purnama, and A. S. Panjaitan, "Analysis of new business opportunities from online informal education mediamorphosis through digital platforms," *International Transactions on Education Technology*, vol. 1, no. 1, pp. 42–52, 2022.
[34] R. J. Kiran, J. Sanil, and S. Asharaf, "A novel approach for model interpretability and domain aware fine-tuning in adaboost," *Human-Centric Intelligent Systems*, vol. 4, no. 4, pp. 610–632, 2024.
[35] T. Shimoda, K. Tomida, C. Nakajima, A. Kawakami, K. Tsutsumimoto, and H. Shimada, "Prevalence and prognostic impact of multiple frailty domain in japanese older adults," *Journal of the American Medical Directors Association*, vol. 25, no. 11, p. 105238, 2024.
[36] P. A. Sunarya, "The impact of gamification on idu (ilearning instruction) in expanding understudy learning inspiration," *International Transactions on Education Technology*, vol. 1, no. 1, pp. 59–67, 2022.
[37] Y. Matsuzaka and R. Yashiro, "Ai-based computer vision techniques and expert systems," *Ai*, vol. 4, no. 1, pp. 289–302, 2023.
[38] P. Guleria and M. Sood, "Explainable ai and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling," *Education and Information Technologies*, vol. 28, no. 1, pp. 1081–1116, 2023.
[39] L. M. Putri Mulyaningsih, "The impact of product quality and brand image on repurchase intention through customer satisfaction," *APTISI Transactions on Management*, vol. 8, no. 1, pp. 1–13, 2024.
[40] N. Lutfiani, Q. Aini, U. Rahardja, N. Septiani, and I. K. Gunawan, "Desain aplikasi software as a service sebagai layanan perbelanjaan online," *ANDHARUPA: Jurnal Desain Komunikasi Visual & Multimedia*, vol. 9, no. 02, pp. 181–194, 2023.