Towards Transparent, Fair, and Trustworthy AI for Sustainable Decisions in High-Stakes Sectors: Research Methodology and Experimental Design

Author 1 Gaurav Kumar Arora[1], Guide 1 Dr. Vishal Jain [2], Supervisor 1 Shashi Kant Gupta [3]

[1] Affiliation Author: Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan , Malaysia.
Email ID: gaurav@gaurav-arora.com.
ORCID: 0009-0002-8855-3525
[2] Affiliation Guide: Sharda University, Greater Noida, India Email ID: drvishaljain83@gmail.com
[3] Affiliation Supervisor: Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan , Malaysia.
raj2008enator@gmail.com.
shashigupta@lincoln.edu.my
Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology. Chitkara University, Rajpura, 140401, Punjab, Índia.
raj2008enator@gmail.com
ORCID: 0000-0001-6587-5607

Abstract: This paper presents the research methodology and experimental design for a Postdoc research program investigating transparent, fair, and trustworthy AI across healthcare, finance, and environmental sustainability. Our first conference literature review identified five critical gaps in the existing research. These gaps led to four hypotheses (H1–H4), which we tested through five pre-registered experiments. The experiments used three benchmark datasets: MIMIC-III for healthcare, Home Credit Default Risk of finance, and UCI Air Quality for environmental research. The Trustworthy AI Assessment Composite (TAAC = $\alpha \cdot \bar{I} + \beta \cdot \bar{F} + \gamma \cdot \bar{R}$) combines normalized scores for interpretability, fairness, and robustness into a validated composite measure. The evaluation used SHAP, LIME, integrated gradients, adversarial debiasing, and perturbation-based robustness testing and was validated by 24 domain experts. A rigorous statistical plan, including Bonferroni correction and bootstrap confidence intervals, governs all tests.

Keywords: Explainable AI; Algorithmic Fairness; TAAC; SHAP; Trustworthy AI; High-Stakes Sectors; Composite Metrics; Health care AI.

Introduction

AI systems are deployed in high-stakes sectors such as clinical diagnosis, credit scoring, and environmental monitoring. These systems must demonstrate more than predictive accuracy. Clinicians, loan applicants, and regulators require interpretable decisions, fairness, and robustness. These properties are well framed as trustworthy AI in The EU AI Act (2024) and IEEE EAD1e (2019), yet research has addressed them in isolation [1].

In the First Conference, we identified five gaps after reviewing more than 150 publications (2015–2024): fragmentation of interpretability, fairness, and trust research; absence of composite metrics; single-domain validation; lack of expert human evaluation; and weak SDG alignment. In this paper, we operationalize those gaps into the Trustworthy AI Assessment Composite (TAAC) framework [3], [7], [9].

Related work

SHAP [7] and LIME [11] provide locally faithful post-hoc explanations but do not guarantee equitable outcomes [2]. Integrated Gradients [12] satisfies completeness for neural networks. Rudin (2019) advocates inherently interpretable models, noting that post-hoc methods cannot fully represent model reasoning [10]. Fairness definitions demographic parity, equalized odds, and individual fairness address bias, but Barocas et al. (2023) show that they are mathematically incompatible [2]. Three-stage mitigation spans pre-processing [5], in-processing [15], and post-processing [8]. Table 1 Map of the gaps each approach leaves.

Table 1. ***Existing Trustworthy AI Approaches: Gap Analysis***

Study	Method	Domain	Gap Left
Lundberg & Lee (2017)	SHAP	Finance / HC	No fairness metrics
Ribeiro et al. (2016)	LIME	General ML	No composite score
Barocas et al. (2023)	Fairness	Hiring/Justice	No interpretability
EU AI Act (2024)	Regulatory	Cross-sector	No computation

Key Contribution

This paper contributes: (1) TAAC, the first composite metric integrating interpretability ($\bar{I}$), fairness ($\bar{F}$), and robustness ($\bar{R}$); (2) four pre-registered hypotheses (H1–H4) with precise statistical thresholds; (3) a five-experiment cross-domain design across healthcare, finance, and environment; and (4) a 24-expert validation panel (clinicians, financial analysts, and environmental scientists) providing concurrent validity [4], [6].

Method, Experiments and Results

In the first conference, we conceptualized a research methodology.

Research Design: Explanatory sequential mixed-methods design. Phase 1 (quantitative) evaluates TAAC through controlled ML experiments. Phase 2 (qualitative) validates scores against expert assessments across all three domains.

Datasets: MIMIC-III 46,520 ICU admissions, 30-day mortality prediction, Sensitive: Age/Gender/Ethnicity [6]. Home Credit: 307,511 applications, loan default classification, Sensitive: Gender/Family status. UCI Air Quality: 9,358 instances, 15 sensors, pollutant regression [3].

Models: Three classes; Logistic Regression (inherently interpretable baseline), XGBoost (post-hoc SHAP), and MLP 2-layer (DeepSHAP + Integrated Gradients). Optuna optimization (100 trials); 5-fold × 5 seeds = 25 evaluations per condition.

$$\mathbf{TAAC(m) = \alpha \cdot \bar{I}(m) + \beta \cdot \bar{F}(m) + \gamma \cdot \bar{R}(m)}$$

$\alpha+\beta+\gamma=1$ | $\bar{I}$=interpretability, $\bar{F}$=fairness, $\bar{R}$=robustness

Interpretability: SHAP [7], LIME [11], and integrated gradients [12]. Three quality metrics: Fidelity(e) = $1-\text{MSE}(f(x),g(e,x))/\text{Var}(f(x))$; Stability(e) = $1-(1/N)\Sigma\|e(x_i+\varepsilon)-e(x_i)\|/\|e(x_i)\|$; Complexity(e) = $|\{j:|\text{SHAP}_j|>0.01\}|$.

Fairness Metrics: ΔDP, ΔEOD, Brier calibration, Lipschitz individual fairness. Three-stage mitigation: Reweighing [5] (pre), Adversarial Debiasing [15] + Exponentiated Gradient [1] (in), and Calibrated equalized Odds [8] (post). Tools: IBM AIF360, Fairlearn 0.10.

Robustness Gaussian noise ($\sigma\in\{0.05,0.10,0.20\}$), top-k SHAP feature removal, temporal distribution shift. Robustness(m,σ) = AUC(D_clean) − AUC(D_perturbed(σ)). MMD^2 quantifies domain covariate shift.

Research Hypotheses

H1: SHAP-regularized XGBoost: $|\Delta\text{AUC}| \leq 0.03$ vs. unconstrained AND Fidelity(constrained) > Fidelity(baseline).

H2: Adversarial debiasing: $\Delta\text{DP} \leq 0.05$ AND $\Delta\text{EOD} \leq 0.05$ with accuracy reduction < 5% on MIMIC-III and Home Credit.

H3: TAAC concurrent validity: Pearson $r \geq 0.70$ with expert trustworthiness ratings (N = 24 experts × 5 reports = 120 ratings).

H4: $|\text{TAAC}_{\text{healthcare}} - \text{TAAC}_{\text{domain\_k}}| \leq 0.10$ for $k \in \{\text{finance, environment}\}$ without domain-specific retraining.

Table 2. ***Five-Experiment Pre-Registered Design***

Experiment (E)	Focus	Datasets	Hypohesses (H)
1	Baseline benchmarking	All 3 × 3 models	H1 baseline
2	SHAP regularization	All 3 datasets	H1
3	Fairness optimization	MIMIC, Home Credit	H2
4	Expert validation panel	All 3 domains	H3

Statistical Plan. Pre-registered. $\alpha = 0.05$, Bonferroni-corrected ($\alpha_{adj} = 0.017$, three domains). Cohen's d ≥ 0.40 (power = 0.80). Bootstrap CI (B = 1,000) for fairness metrics. Wilcoxon signed-rank if normality rejected (Shapiro-Wilk $p < 0.05$). TAAC weight sensitivity: ±20% perturbation.

Discussions

The TAAC framework integrates three trust dimensions that prior work has addressed separately [1],[2],[7]. Pre-registration ensures rigor and reproducibility. Expert validation (E4) addresses the human-grounded evaluation gap [4]. Cross-domain design (E5) uses MMD^2 to quantify the distributional shift between the ICU, credit, and sensor domains. The equal-weight baseline ($\alpha=\beta=\gamma=1/3$) will be refined through AHP elicitation, enabling sector-specific TAAC profiles aligned with EU AI Act Articles 9, 13, and 15 [3],[9].

Conclusions

This paper operationalizes five literature gaps into a pre-registered methodology centered on TAAC ($\alpha\cdot\bar{I} + \beta\cdot\bar{F} + \gamma\cdot\bar{R}$). Four hypotheses with precise thresholds ($\Delta\text{AUC} \leq 0.03$, $\Delta\text{DP} \leq 0.05$, $r \geq 0.70$, deviation ≤ 0.10) were tested across five experiments and three high-stakes domains. The design yields the first cross-

domain empirical evidence on interpretability–accuracy and fairness–accuracy trade-offs in high-stakes AI. Results and reviewer refinements will be presented at the Third Conference.

References

[1] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). *A Reductions Approach to Fair Classification.* Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80, 60–69. https://proceedings.mlr.press/v80/agarwal18a.html

[2] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press. https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/

[3] European Commission. (2024). Regulation (EU) 2024/1689 AI Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[4] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*. arXiv:1812.04608 https://doi.org/10.48550/arXiv.1812.04608

[5] Kamiran, F., & Calders, T. (2012). Data preprocessing for classification without discrimination. Knowledge and Information Systems, 33(1). https://doi.org/10.1007/s10115-011-0463-8

[6] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, 160035. https://doi.org/10.1038/sdata.2016.35

[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[8] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems, 30*. https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html

[9] 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 (KDD '16)*, 1135–1144. https://doi.org/10.1145/2939672.2939778

[10] Rudin, C. (2019). Stop explaining black box ML models. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

[11] Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., 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), 2400304. https://doi.org/10.1002/aisy.202400304

[12] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

[13] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

[14] IEEE. (2019). Ethically Aligned Design (EAD1e). IEEE. https://standards.ieee.org/industry-connections/ec/autonomous-systems/

[15] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*, 335–340. https://doi.org/10.1145/3278721.3278779