13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

A Formal Methodological Framework for Auditing Robustness and Fidelity in Explainable AI: From Application to Trust Certification

Rosa Elysabeth Ralinirina[a*], Jean Christian Ralaivao[a†], Niaiko Michaël Ralaivao[a‡], Alain Josué Ratovondrahona[a§], Thomas Mahatody[a¶]

[a]Doctoral School Modeling-Computer Science, University of Fianarantsoa, Madagascar
ralinirinarosa7@gmail.com

Abstract

While post-hoc interpretability methods have gained traction in high-stakes predictive modeling, their lack of stability and technical truthfulness remains a significant scientific barrier. Building upon our previous research on XAI application in food security [8], this paper moves beyond descriptive interpretation to propose a formal methodological framework for evaluating model robustness and fidelity. We introduce a meta-evaluation protocol that quantifies the sensitivity of explanations to stochastic noise and their adherence to the underlying model's decision boundaries. By implementing a systematic "Feature Influence Verification" layer, we provide a verifiable "Trust Score" designed to audit black-box architectures. Validated using institutional and socio-economic data from Madagascar, the framework demonstrates that high predictive accuracy does not inherently imply explanatory reliability. This contribution establishes a rigorous auditing standard essential for certifying responsible AI systems in applied informatics.

*ORCID: 0009-0003-3048-1765
†ORCID: 0009-0002-5157-1826
‡ORCID: 0009-0003-6659-8535
§ORCID: 0009-0004-3749-4400
¶ORCID: 0009-0003-8060-601X

Keywords: Explainable AI (XAI), Robustness, Fidelity, Methodological Framework, Trustworthy AI, Food Security.

1. Introduction

The transition from black-box predictive performance to transparent decision-making is no longer an option but a requirement in critical infrastructures. In the Malagasy context, where AI models guide humanitarian and agricultural interventions, the reliability of a "justification" is as vital as the prediction itself. Previous works, including our study on food security indicators [8], highlighted the potential of XAI but also identified significant variance in explanation outputs. As noted by Adadi, Berrada [1] and Guidotti et al. [4], the absence of formal auditing metrics leads to an "illusion of transparency." Recent surveys in specific fields like healthcare [10] and natural language processing [3] further emphasize that interpretability requires domain-specific verification. This research addresses this gap by shifting from a simple application of XAI tools toward the conception of a robust evaluative framework.

2. Methodological Framework Architecture

Our innovation lies in a modular auditing pipeline that evaluates the explainer-model interaction through two rigorous dimensions. This approach aligns with the broader goals of DARPA's Explainable AI program [5].

2.1. Robustness and Local Stability

We define robustness as the resilience of an explanation $E(x)$ against infinitesimal input perturbations. We formalize this using a local stability index, ensuring that similar environmental conditions yield consistent feature importance rankings. As demonstrated by Slack et al. [9] and Alvarez-Melis, Jaakkola [2], an unstable explanation indicates a model that has learned noise rather than causal drivers.

2.2. Fidelity through Feature Ablation

Fidelity (faithfulness) measures if the explanation truthfully reflects the model's logic. We propose a systematic "In-and-Out" protocol where features identified as crucial are masked to observe the actual decay in model performance. This prevents "plausible but false" interpretations, a risk particularly high in sensitive areas like predictive modeling for nutrition [6] and general agriculture [7].

3. Case Study: Madagascar Food Security

The framework was validated using a multi-sectoral dataset consolidating official reports from the Ministry of Agriculture and Livestock, the Ministry of Public Health, and the National Development Plan (PND). Our study incorporates variables such as rainfall patterns, cyclonic events, and logistical constraints related to transport infrastructure. Furthermore, the model integrates international indicators from the FAO, nutrition status surveys, and trade dynamics, specifically the import and export of staple foods. Preliminary experiments reveal that our framework successfully flags models that rely on spurious correlations, providing a standardized metric for trust certification in complex socio-economic environments.

4. Conclusion

This methodological framework provides formal tools to audit XAI reliability. By quantifying robustness and fidelity, we bridge the gap between algorithmic complexity and responsible action. Future work will focus on integrating these metrics into decision-support systems for Malagasy national institutions.

References

[1] A. Adadi, M. Berrada: *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, IEEE Access 6 (2018), pp. 52138–52160, DOI: 10.1109/ACCESS.2018.2870052.

[2] D. Alvarez-Melis, T. S. Jaakkola: *On the Robustness of Interpretability Methods* (2018), DOI: 10.48550/arXiv.1806.08049.

[3] M. Danilevsky et al.: *A Survey of the State of Explainable AI for Natural Language Processing* (2020), pp. 447–459, DOI: 10.18653/v1/2020.aacl-main.46.

[4] R. Guidotti et al.: *A Survey of Methods for Explaining Black Box Models*, ACM Computing Surveys 51.5 (2018), pp. 1–42, DOI: 10.1145/3236009.

[5] D. Gunning, D. W. Aha: *DARPA's Explainable Artificial Intelligence (XAI) Program*, AI Magazine 40.2 (2019), pp. 44–58, DOI: 10.1609/aimag.v40i2.2850.

[6] T. Islam et al.: *Explainable AI for Predictive Modeling in Healthcare*, Smart Health 32 (2024), p. 100465, DOI: 10.1016/j.smhl.2024.100465.

[7] E. S. M. Linheiro et al.: *Explainable AI (XAI) for Agriculture*, in: Industry 4.0 Convergence with AI, IoT, Big Data and Cloud Computing, Bentham Science Publishers, 2023, pp. 161–176, DOI: 10.2174/9789815179187123010013.

[8] R. E. Ralinirina, J. C. Ralaivao, N. M. Ralaivao, A. J. Ratovondrahona, T. Mahatody: *Unveiling the Potential of Explainable Artificial Intelligence in Predictive Modeling, Exploring Food Security and Nutrition in Madagascar*, Communications in Computer and Information Science 2391 (2025), DOI: 10.1007/978-3-031-84093-7_3.

[9] D. Slack et al.: *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods* (2020), pp. 180–186, DOI: 10.1145/3375627.3375830.

[10] G. Stiglic et al.: *Interpretability of machine learning-based prediction models in healthcare*, WIREs Data Mining and Knowledge Discovery 10.5 (2020), e1379, DOI: 10.1002/widm.1379.