

A Comprehensive Audit Framework for Responsible AI Systems

Arnav Vaidya, Arya Thanekar, Shruti Mone, Sampada Tagalpallewar, Prof. Manish Khodaskar

Department of Information Technology, Pune Institute of Computer Technology, Pune, India, arnavaidya@gmail.com, asthanekar2004@gmail.com, shrutimone7@gmail.com, tagalpallewarsampada10@gmail.com, mrkhodaskar@pict.edu

Abstract.

With the increasing use of AI systems in critical applications, there is a need for structured audits over multiple dimensions of Responsible AI. This paper proposes a holistic AI auditing framework combining energy efficiency, fairness, explainability, and data governance into one pipeline. Experiments show a 72% cumulative energy saving via early stopping, equalized odds reduced from 0.083 to 0.031, ~90% SHAP–LIME agreement, and correct identification of 589 unmasked PII values of 1,000, showing multi-dimensional auditing to be feasible, structured, and scalable.

Keywords. AI Auditing; Fairness; Explainable AI; Energy Efficiency; Data Governance; Compliance

1. INTRODUCTION

AI models are now deployed across finance, healthcare, manufacturing, and cybersecurity, yet high accuracy often lacks explainability, undermining trust and fairness. Fairness, sustainability [1], and compliance automation [2] have been addressed separately, though deployment requires evaluating AI governance together. This paper presents an AI auditing system with modules for fairness, explainability, energy efficiency, and compliance, inspired by Model Cards [3]. This work builds on our earlier survey of AI auditing frameworks [4], translating its comparative findings into a working implementation.

2. RELATED WORK

CodeCarbon has managed to quantify the carbon footprint of AI [1]. There have been developments in fairness toolkits such as Fairlearn [5] and Model Cards [3]. SHAP and LIME methods have been used in malware detection [6]. For PII Detection, ML-based NER [7] and NLP-aligned GDPR [2] are used.

Approach	Fairness	Explainability	Energy	Compliance
Fairlearn [5]	Yes	No	No	No

CodeCarbon [1]	No	No	Yes	No
Presidio [8]	No	No	No	Yes

Table 2.1. Comparison of Prior Auditing Approaches

3. SYSTEM ARCHITECTURE AND IMPLEMENTATION

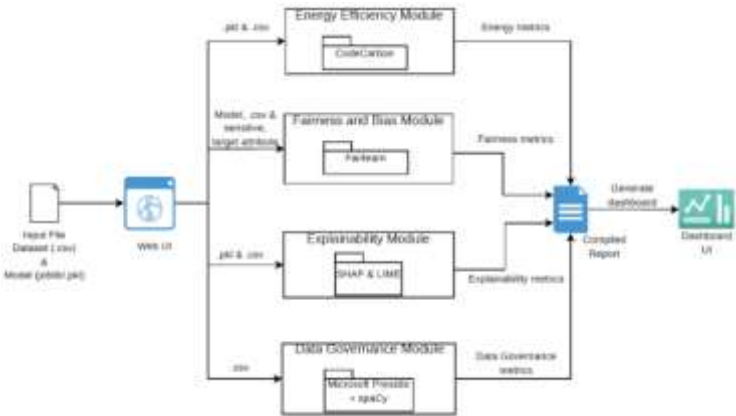


Figure 3.1. System Architecture of the Audit Framework

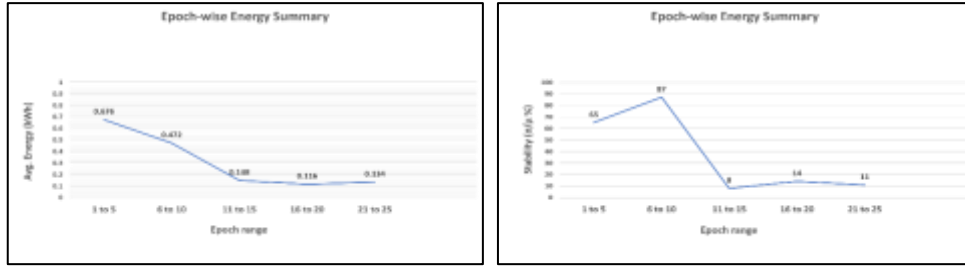
The system evaluates models across four dimensions (Figure. 3.1).

Energy Tracking Module (CodeCarbon) tracks energy (CPU + RAM + GPU), emissions, and stability (σ/μ) for early-stopping recommendations. Fairness Module (Fairlearn) computes disparity metrics (demographic parity, equalized odds, FPR/FNR) with ExponentiatedGradient mitigation. Explainability Module (SHAP/LIME) – SHAP gives global attributions using Shapley values while LIME gives local attributions whose agreement is determined using the Jaccard index. Data Governance Module (Presidio + spaCy) applies Presidio in combination with NER from spaCy to detect PII/PHI (30+ types) with 0.7 confidence level and compliance with GDPR, HIPAA, ISO 27001 and PCI-DSS.

4. EXPERIMENTS AND RESULTS

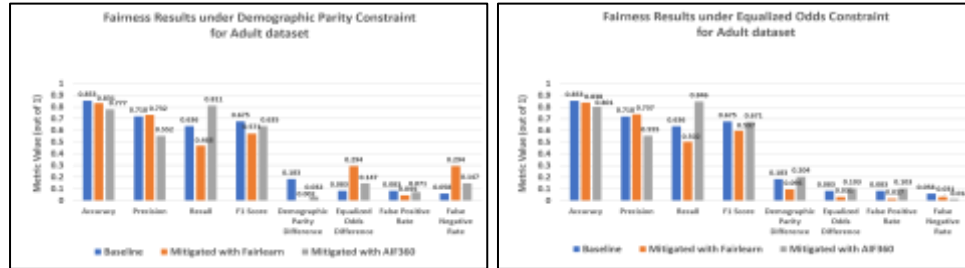
4.1. Energy Efficiency

Tested on a CatBoostClassifier (400 records, 25 epochs): 0.0078 Wh, 0.0055 g CO2 (CPU 84%, RAM 16%). Early stopping at epoch 15 saved 72% energy while retaining accuracy (Figure. 4.1).

Figure 4.1. *Epoch-wise Energy Summary*

4.2. Fairness Analysis

Evaluated on the UCI Adult Income dataset [9] using 'sex': Equalized Odds fell from 0.083 to 0.031 (F1 0.675 $\rightarrow$ 0.597); Demographic Parity fell from 0.183 to 0.002 at higher cost (F1 0.675 $\rightarrow$ 0.571) (Figure. 4.2).

Figure 4.2. *Fairness Results for Adult Income dataset*

4.3. Explainability Analysis

A RandomForestClassifier trained on the Adult Income dataset [9] (16,281 samples, 14 attributes): SHAP and LIME both ranked relationship, education_num, marital_status, age, and capital_gain as top features (~90% Jaccard agreement).

4.4. Data Governance and Compliance

Tested on 1,000 synthetic transactions (589 unmasked PII, 411 anonymized), the module correctly flagged all actual PII while leaving properly tokenized records unflagged, validating Presidio's threshold-based filtering for automated compliance screening.

5. CONCLUSIONS

This research proposes an integrated Responsible AI pipeline for energy efficiency, fairness, interpretability, and auditability, showing multi-dimensional auditing to be feasible and scalable; future work targets training-phase efficiency and unstructured-data risk scoring.

6. REFERENCES

- [1] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of BLOOM, a 176B parameter language model," J. Mach. Learn. Res., vol. 24, no. 1, art. no. 253, pp. 11990–12004, 2023. doi: 10.5555/3648699.3648952

- [2] A.-J. Aberkane, G. Poels, and S. V. Broucke, "Exploring automated GDPR-compliance in requirements engineering," *IEEE Access*, vol. 9, pp. 66542–66559, 2021. doi: 10.1109/ACCESS.2021.3076921
- [3] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, Transparency (FAT*)*, Atlanta, GA, USA, Jan. 2019, pp. 220–229. doi: 10.1145/3287560.3287596
- [4] A. Vaidya, S. Tagalpallewar, A. Thanekar, S. Mone, and M. Khodaskar, "Fairness to Footprint: A Comprehensive Audit Framework for Sustainable AI Systems," *PICT's Int. J. Eng. Technol. (PIJET)*, vol. 3, no. 1, pp. 120–135, Dec. 2025.
- [5] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Redmond, WA, USA, Tech. Rep. MSR-TR-2020-32, May 2020. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>
- [6] S. Nazim, M. M. Alam, S. S. Rizvi, J. C. Mustapha, S. S. Hussain, and M. M. Suud, "Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM," *PLoS ONE*, vol. 20, no. 5, art. no. e0318542, May 2025. doi: 10.1371/journal.pone.0318542
- [7] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *J. Am. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 596–606, May 2017. doi: 10.1093/jamia/ocw156
- [8] D. Asimopoulos, I. Siniosoglou, V. Argyriou, T. Karamitsou, E. Fountoukidis, S. K. Goudos, et al., "Benchmarking advanced text anonymisation methods: A comparative study on novel and traditional approaches," in *Proc. 13th Int. Conf. Modern Circuits Syst. Technol. (MOCASST)*, 2024, pp. 1–6. doi: 10.1109/MOCASST61810.2024.10615642
- [9] R. Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," in *Proc. 2nd Int. Conf. Knowledge Discovery Data Mining (KDD-96)*, E. Simoudis, J. Han, and U. M. Fayyad, Eds. Menlo Park, CA, USA: AAAI Press, 1996, pp. 202–207.

Biographies



Arnav Vaidya holds a B.E. in IT from PICT (2026) and works as a cybersecurity consultant, researching web security, AI, and cloud.



Arya Thanekar holds a B.E. in IT from PICT (2026) and works as an Associate Engineer in Industrial Automation, researching machine learning and software engineering.



Shruti Mone holds a B.E. in IT from PICT (2026) and works in data analytics and consulting, researching AI and data analytics.



Sampada Tagalpallewar holds a B.E. in IT from PICT (2026) and works as a Product Manager in Fintech, researching AI and deep learning.



Mr. Manish R. Khodaskar, M.E. in Computer Engineering, is an Assistant Professor in the Department of Information Technology at Pune Institute of Computer Technology (PICT), Pune. He has over 20 years of academic experience along with one year of industry experience.