Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Multi-Dimensional Model Integrity and Responsibility Assessment Index and Scoring Framework

Phuc Truong Loc Nguyen[†*], Thanh Hung Do[†],
Truong Thanh Hung Nguyen[‡], Hung Cao[‡]
[†] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[‡] University of New Brunswick, Canada

Abstract

Artificial intelligence in high-stakes tabular domains cannot be evaluated by predictive performance alone, yet current practice still assesses explainability, fairness, robustness, privacy, and sustainability mostly in isolation. We propose the Model Integrity and Responsibility Assessment Index (MIRAI), a unified evaluation framework that measures tabular models across these five dimensions under a controlled comparison setting and aggregates them into a single score. MIRAI combines established metrics through normalized and direction-aligned dimension scores, which enables direct comparison across models with different architectural and computational profiles. Experiments on healthcare, financial, and socioeconomic datasets show that higher predictive performance does not necessarily imply better overall integrity and responsibility. In several cases, simpler models achieve a stronger cross-dimensional balance than more complex deep tabular architectures. MIRAI provides a compact and practical basis for responsible model selection in regulated settings.

Keywords: Responsible AI, Model Integrity and Responsibility Metrics.

1. Introduction

Artificial Intelligence (AI) is increasingly used in high-stakes domains such as healthcare, finance, and public decision support, where model outputs can directly affect individuals and institutions [1–6]. In these settings, tabular data remains a dominant format for predictive modeling [7, 8]. Predictive performance alone is not sufficient for deployment, because a highly accurate model may still be difficult to interpret, unfair across demographic groups, vulnerable to perturbations, prone to privacy leakage, or expensive to operate. These properties often conflict, which makes responsible model selection inherently multi-dimensional [9–11]. Yet current evaluation practice remains fragmented. Explainability, fairness, robustness, privacy, and sustainability are still assessed largely in isolation, often with separate metrics and toolchains [12–14]. Recent integrated efforts move toward broader responsible-AI assessment [9, 15, 16], but model selection still lacks a compact framework that can translate heterogeneous evidence into a clear comparative judgment [10, 17, 18].

We address this gap with the *Model Integrity and Responsibility Assessment Index (MIRAI)*, a unified evaluation framework for tabular models. MIRAI quantifies explainability, fairness, sustainability, robustness, and privacy, aligns heterogeneous metrics to a common scoring direction, and aggregates them into a single comparative index. This provides a compact and decision-oriented basis for responsible model assessment in regulated settings [9, 17]. Experiments on healthcare, financial, and socioeconomic datasets show that stronger predictive performance does not necessarily imply better overall integrity and responsibility. In several cases, simpler models achieve a better cross-dimensional balance than more complex deep tabular architectures.

[*]Corresponding author: loc.pt.nguyen@fau.de

2. Background and Related Work

Recent work has established a rich set of metrics and toolkits for evaluating responsible AI properties at the dimension level. In explainability, quantitative evaluation has moved beyond visual inspection toward protocol-based assessment of faithfulness, robustness, randomization, and complexity, with Quantus becoming a widely used benchmark framework [12, 19]. Fairness evaluation is similarly supported by mature libraries such as Fairlearn and AIF360, while robustness and privacy are commonly examined through adversarial testing and leakage analysis frameworks such as ART and related toolkits [13, 14, 20]. Sustainability is mostly assessed through carbon and computational cost accounting, which extends evaluation beyond predictive utility to environmental and resource efficiency [21, 22].

Building on these foundations, several studies have moved toward integrated responsible-AI assessment. RAISE proposes a unified scoring pipeline for tabular models across multiple dimensions [9], while SAFE introduces an internally consistent integrated metric formulation [16]. Complementary compliance-oriented frameworks further stress the need for reproducible and auditable assessment pipelines in regulated settings such as those targeted by the EU AI Act [15, 18]. Yet an important gap remains. Existing approaches still vary in dimension coverage, metric aggregation, and comparison protocols, while trade-offs across fairness, explainability, privacy, robustness, and efficiency can substantially alter model rankings [9, 10]. This makes deployment judgments difficult when raw metrics point in different directions. MIRAI is designed to address this gap through a compact, controlled, and directly comparative evaluation framework for tabular models.

3. MIRAI Scoring Framework

The MIRAI framework performs controlled comparative evaluation for tabular binary classification. It requires at least two candidate models from major tabular learning families, including Decision Tree (DT), XGBoost (XGB), Support Vector Machine (SVM), Multilayer Perceptron (MLP), TabResNet (TRN), and Feature Tokenizer Transformer (FTT) [23]. For each candidate, it supports model-specific architectural settings, such as tree depth and hidden dimensions, together with training hyperparameters. It also records key dataset properties, including sample size, feature dimensionality, domain, and task type. In addition, it designates a Target Model as the intended deployment candidate and uses it as the reference for ranking and comparison. This makes the evaluation decision-oriented, since model quality, computational overhead, and cross-dimensional trade-offs can be judged relative to a concrete baseline and used to support model and hyperparameter refinement.

Each model is evaluated across five dimensions: explainability, fairness, sustainability, robustness, and privacy. Explainability is computed in two stages. SHAP [24] first generates feature attributions, and Quantus [12] then evaluates them with eight metrics: Local Lipschitz Estimate [25] and Consistency [26] for robustness, Faithfulness Correlation [27] and Faithfulness Estimate [28] for faithfulness, Model Parameter Randomization Test [29] and Random Logit Test [30] for randomization, and Sparseness [31] and Complexity [27] for complexity. Fairness is assessed through subgroup performance in Accuracy, Precision, Recall, True Positive Rate, and False Positive Rate, followed by absolute inter-group disparities, Demographic Parity [32], and Equalized Odds [33], using Fairlearn [13] and AIF360 [14]. Sustainability combines carbon emissions, parameter count, FLOPs, and MACs. Carbon impact is estimated with the Lacoste score [21], where p_c is average CPU power, p_g is average GPU power, and $p_t = (p_c + p_g)/1000$ is total power in kilowatts. This value is multiplied by the 2023 Canadian electricity emission rate [34] and then normalized by the 2023 daily per-capita emission reference [35]. Robustness is measured through the HopSkipJump Attack [36] accuracy gap with ART [20] and prediction-space Maximum Mean Discrepancy

drift detection with Alibi-Detect [37]. Privacy is quantified through Membership Inference Privacy [38] and SHAPr Privacy [39], both implemented with ART.

All raw metrics are normalized to $[0, 1]$, where 1 denotes the most desirable outcome. Metrics for which lower raw values are preferable are direction-aligned through $1 - \text{raw}$. The normalized metrics within each dimension are averaged to obtain a Dimension Score, and the final MIRAI score is computed as $\text{MIRAI} = \sum_{d=1}^{5} w_d \text{DS}_d$. Equal weights are used by default, with $w_d = 0.2$, but user-defined weights are also supported to reflect application-specific priorities, such as fairness in regulated settings or robustness in safety-critical deployment. Predictive Accuracy and F1-score are reported separately, so MIRAI complements rather than replaces standard performance measures. This formulation preserves dimension-specific evidence while enabling compact, context-aware, and directly comparative model selection.

4. Experiment and Results

We evaluate MIRAI on six classifiers that span major tabular learning paradigms: DT, XGB, SVM, MLP, TRN, and FTT. The evaluation uses three public high-stakes tabular datasets from healthcare, finance, and socioeconomics: Diabetes Hospitals [40], German Credit [41], and Census Income [42]. All models are trained under controlled conditions to ensure fair comparison. For fairness evaluation, "gender" is used as the sensitive attribute and "male" is treated as the privileged group, following the protocol in [14, 43]. MIRAI then compares the models jointly across explainability, fairness, sustainability, robustness, and privacy. Results are presented in Table 1, Table 2, and Table 3.

Across datasets, MIRAI reveals a consistent trade-off between predictive strength and cross-dimensional model quality. Deep tabular models, especially TRN and FTT, remain competitive in predictive performance and can achieve strong explainability, with TRN reaching top-tier explainability on Census Income and German Credit. However, these gains are often offset by weaker sustainability and privacy, especially for FTT, whose high computational cost leads to severe efficiency penalties. By contrast, MLP and XGB maintain very strong sustainability, while SVM shows the strongest privacy behavior on Diabetes and German Credit. Fairness and robustness also depend on the data regime. On Diabetes, most models achieve both high robustness and high fairness, with XGB reaching near-perfect fairness. On smaller or more imbalanced settings, such as German Credit, deeper models lose ground in fairness and privacy.

The key result is that predictive metrics and MIRAI rankings do not coincide. Models with the best F1 scores do not necessarily provide the strongest overall integrity and responsibility profile when all five dimensions are considered jointly. On German Credit, TRN attains the top F1 score, yet MLP achieves the stronger MIRAI ranking. On Diabetes, MLP records the highest MIRAI score, while FTT ranks much lower despite competitive predictive performance. Across the benchmarks, MLP and SVM provide the most favorable cross-dimensional balance. This indicates that higher model complexity does not guarantee better all-round behavior and that simpler models can be stronger candidates for deployment in regulated tabular settings.

5. Conclusion

We presented MIRAI, a unified evaluation framework for tabular models that integrates explainability, fairness, sustainability, robustness, and privacy beyond predictive performance alone. Our results show that higher model complexity does not necessarily translate

Table 1. **Diabetes Hospitals**: 101763 samples, 22 features. The best results are in **bold**. The second-best results are underlined.

Model	DT	XGB	SVM	MLP	TRN	FTT
MIRAI	0.7635	0.7763	0.7724	**0.7776**	0.7607	0.5636
Accuracy	0.7870	0.8880	0.8880	0.8840	0.8850	0.8880
F1 Score	0.7960	0.8360	0.8360	0.8380	0.8370	0.8360
Explainability	0.4456	**0.5126**	0.4312	0.4850	0.5101	0.4635
Complexity	0.5619	0.6363	0.6237	0.6454	0.6988	0.6639
Faithfulness	0.6105	0.7211	0.5285	0.5318	0.5571	0.6910
Robustness (Expl.)	0.1101	0.1931	0.0725	0.2016	0.1730	0.1260
Randomization	0.5000	0.5000	0.5000	0.5611	0.6117	0.3732
Fairness	0.9980	**0.9993**	0.9645	0.9947	0.9887	0.9155
Accuracy Diff*	0.0020	0.0040	0.0050	0.0040	0.0030	0.0040
Precision Diff*	0.0040	0.0000	0.2000	0.0240	0.0410	0.5000
TPR Diff*	0.0020	0.0000	0.0030	0.0020	0.0110	0.0010
FPR Diff*	0.0010	0.0000	0.0010	0.0000	0.0000	0.0010
Demographic Parity Diff*	0.0010	0.0000	0.0010	0.0000	0.0020	0.0000
Equalized Odds Diff*	0.0020	0.0000	0.0030	0.0020	0.0110	0.0010
Sustainability	0.9899	**0.9992**	0.9639	0.9987	0.8913	0.0000
Parameter Count*	0.9698	0.9976	0.8988	0.9964	0.7022	0.0000
FLOPs per Sample*	1.0000	1.0000	0.9976	0.9998	0.9861	0.0000
MACs per Sample*	1.0000	1.0000	0.9952	0.9998	0.9857	0.0000
Normalized $kgCO_2e$*	0.9925	0.9989	0.9720	0.9987	0.8850	0.0000
Robustness	0.8676	0.8619	0.8665	0.8506	0.8553	**0.8730**
HSJA Robustness	0.9558	0.9257	0.9304	0.8889	0.9058	0.9237
Drift Robustness	0.7794	0.7980	0.8027	0.8123	0.8048	0.8224
Privacy	0.5164	0.5144	**0.6361**	0.5590	0.5582	0.5635
MI Privacy	0.4762	0.4813	0.7158	0.5666	0.5639	0.5724
SHAPr Privacy	0.5566	0.5475	0.5565	0.5514	0.5526	0.5545

Metrics marked with an asterisk (*) are lower-is-better by definition. Their values have been inverted using 1 − raw.

Table 2. **German Credit**: 1000 samples, 22 features. The best results are in **bold**. The second-best results are underlined.

Model	DT	XGB	SVM	MLP	TRN	FTT
MIRAI	0.7282	0.7086	0.7377	**0.7422**	0.6540	0.4815
Accuracy	0.7000	0.7550	0.7100	0.7350	0.7500	0.7350
F1 Score	0.7070	0.7460	0.6910	0.7330	0.7520	0.7340
Explainability	0.4601	0.4371	0.4451	0.4951	**0.4982**	0.4690
Complexity	0.6065	0.5606	0.6072	0.6154	0.6104	0.6343
Faithfulness	0.5265	0.4981	0.5395	0.5585	0.4722	0.5045
Robustness (Expl.)	0.2074	0.1898	0.1339	0.1797	0.1651	0.1656
Randomization	0.5000	0.5000	0.5000	0.6270	0.7451	0.5716
Fairness	0.8907	**0.9465**	0.9017	0.8862	0.8170	0.9202
Accuracy Diff*	0.1210	0.0550	0.0450	0.0190	0.1520	0.0620
Precision Diff*	0.0790	0.0420	0.0190	0.0160	0.0580	0.0370
TPR Diff*	0.1650	0.0650	0.1000	0.1000	0.2650	0.1000
FPR Diff*	0.0000	0.0250	0.1500	0.2000	0.1250	0.0750
Demographic Parity Diff*	0.1260	0.0690	0.1260	0.1480	0.2330	0.1050
Equalized Odds Diff*	0.1650	0.0650	0.1500	0.2000	0.2650	0.1000
Sustainability	**0.9999**	0.9993	0.9951	0.9987	0.8913	0.0000
Parameter Count*	0.9996	0.9978	0.9863	0.9964	0.7022	0.0000
FLOPs per Sample*	1.0000	1.0000	0.9997	0.9998	0.9861	0.0000
MACs per Sample*	1.0000	1.0000	0.9993	0.9998	0.9857	0.0000
Normalized $kgCO_2e$*	0.9992	0.9985	0.9920	0.9978	0.8900	0.0000
Robustness	0.6715	0.5308	0.6830	**0.6930**	0.4927	0.4853
HSJA Robustness	0.5900	0.2700	0.5750	0.5650	0.2300	0.2000
Drift Robustness	0.7530	0.7916	0.7911	0.8210	0.7555	0.7706
Privacy	0.6188	0.6295	**0.6635**	0.6382	0.5706	0.5329
MI Privacy	0.5333	0.7095	0.7000	0.6476	0.5809	0.5476
SHAPr Privacy	0.7042	0.5495	0.6270	0.6289	0.5602	0.5181

Metrics marked with an asterisk (*) are lower-is-better by definition. Their values have been inverted using 1 − raw.

Table 3. **Census Income**: 32561 samples, 14 features. The best results are in **bold**. The second-best results are underlined.

Model	DT	XGB	SVM	MLP	TRN	FTT
MIRAI	0.6925	0.6890	**0.7209**	0.7189	0.6881	0.5698
Accuracy	0.8130	0.8690	0.8530	0.8520	0.8500	0.8490
F1 Score	0.8140	0.8630	0.8440	0.8500	0.8460	0.8480
Explainability	0.4491	0.4271	0.4547	0.5078	**0.5250**	0.4809
Complexity	0.6287	0.6259	0.6334	0.6014	0.6704	0.7005
Faithfulness	0.5339	0.4810	0.5856	0.5864	0.5963	0.6026
Robustness (Expl.)	0.1339	0.1016	0.0999	0.1171	0.0975	0.1003
Randomization	0.5000	0.5000	0.5000	0.7262	0.7358	0.5200
Fairness	0.9035	0.9012	0.9117	0.9168	0.9210	**0.9387**
Accuracy Diff*	0.1210	0.1010	0.1130	0.1150	0.1060	0.1100
Precision Diff*	0.0860	0.0590	0.0520	0.0090	0.0640	0.0050
TPR Diff*	0.0000	0.0990	0.0730	0.0490	0.0120	0.0170
FPR Diff*	0.0970	0.0630	0.0620	0.0770	0.0660	0.0480
Demographic Parity Diff*	0.1780	0.1720	0.1570	0.1720	0.1600	0.1400
Equalized Odds Diff*	0.0970	0.0990	0.0730	0.0770	0.0660	0.0480
Sustainability	0.9971	**0.9992**	0.9779	0.9991	0.8864	0.0000
Parameter Count*	0.9914	0.9976	0.9403	0.9976	0.7023	0.0000
FLOPs per Sample*	1.0000	1.0000	0.9978	0.9998	0.9787	0.0000
MACs per Sample*	1.0000	1.0000	0.9957	0.9998	0.9781	0.0000
Normalized $kgCO_2e$*	0.9968	0.9990	0.9735	0.9988	0.8840	0.0000
Robustness	0.5948	0.5092	0.6406	0.5967	0.5160	**0.8607**
HSJA Robustness	0.3740	0.3840	0.4100	0.3560	0.4300	0.9280
Drift Robustness	0.8156	0.6345	0.8712	0.8373	0.6020	0.7934
Privacy	0.5171	0.6098	**0.6166**	0.5710	0.5922	0.5700
MI Privacy	0.4767	0.6541	0.6729	0.5813	0.6037	0.5841
SHAPr Privacy	0.5574	0.5654	0.5602	0.5607	0.5807	0.5559

Metrics marked with an asterisk (*) are lower-is-better by definition. Their values have been inverted using 1 − raw.

into better overall integrity and responsibility. In several high-stakes settings, simpler models such as MLP and SVM achieve a stronger cross-dimensional balance than transformer-based architectures. These findings suggest that compact, well-balanced models can be better suited to deployment in regulated domains.

Acknowledgment

This work was supported by NSERC Discovery Grant No RGPIN-2025-04478 and NSERC Discovery Supplement Award No DGECR-2025-00129.

References

[1] L. P. T. Nguyen et al. "Motion2Meaning: A Clinician-Centered Framework for Contestable LLM in Parkinson's Disease Gait Interpretation". In: *Proceedings of 9th International Symposium on Chatbots and Human-centred AI (CONVERSATIONS) 2025*. 2025.

[2] H. Nguyen et al. "Heart2Mind: Human-Centered Contestable Psychiatric Disorder Prediction System Using Wearable ECG Monitors". In: *ACM Trans. Comput. Healthcare* (2026).

[3] H. T. T. Nguyen et al. "XEdgeAI: A human-centered industrial inspection framework with data-centric Explainable Edge AI approach". In: *Information Fusion* 116 (2025), p. 102782. ISSN: 1566-2535.

[4] L. P. T. Nguyen et al. "ODExAI: A Comprehensive Object Detection Explainable AI Evaluation". In: *KI 2025: Advances in Artificial Intelligence*. 2026, pp. 118–133.

[5] H. Nguyen et al. "LangXAI: Integrating Large Vision Models for Generating Textual Explanations to Enhance Explainability in Visual Perception Tasks". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. Aug. 2024.

[6] T. T. H. Nguyen, P. T. L. Nguyen, M. Wachowicz, and H. Cao. "MACeIP: A Multimodal Ambient Context-Enriched Intelligence Platform in Smart Cities". In: *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. 2024, pp. 1–4.

[7] R. Shwartz-Ziv and A. Armon. "Tabular data: Deep learning is not all you need". In: *Information Fusion* 81 (2022), pp. 84–90. ISSN: 1566-2535.

[8] D. McElfresh, S. Khandagale, J. Valverde, V. Prasad C, G. Ramakrishnan, M. Goldblum, and C. White. "When do neural nets outperform boosted trees on tabular data?" In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 76336–76369.

[9] L. P. T. Nguyen and H. T. Do. "RAISE: A Unified Framework for Responsible AI Scoring and Evaluation". In: *PRIMA 2025: Principles and Practice of Multi-Agent Systems*. Cham: Springer Nature Switzerland, 2026, pp. 453–460. ISBN: 978-3-032-13562-9.

[10] N. Kemmerzell and A. Schreiner. "Quantifying the Trade-Offs Between Dimensions of Trustworthy AI - An Empirical Study on Fairness, Explainability, Privacy, and Robustness". In: *KI 2024: Advances in Artificial Intelligence*. 2024, pp. 128–146.

[11] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri. "On Adversarial Bias and the Robustness of Fair Machine Learning". In: *arXiv preprint arXiv:2006.08669* (2020).

[12] A. Hedstrom et al. "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond". In: *Journal of Machine Learning Research* 24.34 (2023).

[13] H. Weerts et al. "Fairlearn: Assessing and Improving Fairness of AI Systems". In: *Journal of Machine Learning Research* 24.257 (2023), pp. 1–8.

[14] R. K. Bellamy et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[15] M. Poretschkin et al. "Guideline for Trustworthy Artificial Intelligence–AI Assessment Catalog". In: *arXiv preprint arXiv:2307.03681* (2023).

[16] P. Giudici and V. Kolesnikov. "SAFE AI metrics: An integrated approach". In: *Machine Learning with Applications* 23 (2026), p. 100821. ISSN: 2666-8270.

[17] T. Clement et al. "Towards Quantifying Compliance with the EU AI Act". In: *Proceedings of the 59th Hawaii International Conference on System Sciences (HICSS) 2026*. 2026.

[18] J. Kelly et al. "Navigating the EU AI Act: A methodological approach to compliance for safety-critical products". In: *IEEE Conference on Artificial Intelligence*. IEEE. 2024.

[19] P. Q. Le et al. "Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. Aug. 2023, pp. 6665–6673.

[20] M.-I. Nicolae et al. "Adversarial Robustness Toolbox v1. 0.0". In: *arXiv:1807.01069* (2018).

[21] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. "Quantifying the Carbon Emissions of Machine Learning". In: *arXiv preprint arXiv:1910.09700* (2019).

[22] F. Neutatz et al. “How Green is AutoML for Tabular Data?” In: *EDBT*. 2025.
[23] Y. Gorishniy et al. “Revisiting Deep Learning Models for Tabular Data”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 18932–18943.
[24] S. M. Lundberg and S.-I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.
[25] D. Alvarez-Melis and T. S. Jaakkola. “On the Robustness of Interpretability Methods”. In: *arXiv preprint arXiv:1806.08049* (2018).
[26] S. Dasgupta, N. Frost, and M. Moshkovitz. “Framework for Evaluating Faithfulness of Local Explanations”. In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. PMLR, 2022, pp. 4794–4815.
[27] U. Bhatt, A. Weller, and J. M. F. Moura. “Evaluating and Aggregating Feature-based Model Explanations”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. 2020, pp. 3016–3022.
[28] D. Alvarez Melis and T. Jaakkola. “Towards Robust Interpretability with Self-Explaining Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.
[29] J. Adebayo et al. “Sanity Checks for Saliency Maps”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.
[30] L. Sixt et al. “When Explanations Lie: Why Many Modified BP Attributions Fail”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. PMLR, 2020.
[31] P. Chalasani et al. “Concise Explanations of Neural Networks using Adversarial Training”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. PMLR, 2020, pp. 1383–1391.
[32] R. Berk et al. “Fairness in Criminal Justice Risk Assessments: The State of the Art”. In: *Sociological Methods & Research* 50.1 (2021), pp. 3–44.
[33] M. Hardt, E. Price, and N. Srebro. “Equality of Opportunity in Supervised Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.
[34] Environment and Climate Change Canada. *Annex 13: Electricity in Canada, Summary and Intensity Tables (Electricity Intensity)*. Mar. 2025.
[35] Environment and Climate Change Canada. *Greenhouse Gas Emissions (Canadian Environmental Sustainability Indicators)*. Mar. 2025.
[36] J. Chen, M. I. Jordan, and M. J. Wainwright. “HopSkipJumpAttack: A Query-Efficient Decision-Based Adversarial Attack”. In: *arXiv preprint arXiv:1904.02144* (2019).
[37] A. Van Looveren et al. *Alibi Detect: Algorithms for outlier, adversarial and drift detection*. Version 0.13.0. Dec. 11, 2025.
[38] R. Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017, pp. 3–18.
[39] V. Duddu, S. Szyller, and N Asokan. “SHAPr: An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning”. In: *arXiv preprint arXiv:2112.02230* (2021).
[40] J. Clore, K. Cios, J. DeShazo, and B. Strack. *Diabetes 130-US Hospitals for Years 1999-2008*. UCI Machine Learning Repository. 2014.
[41] H. Hofmann. *Statlog (German Credit Data)*. UCI Machine Learning Repository. 1994.
[42] B. Becker and R. Kohavi. *Adult*. UCI Machine Learning Repository. 1996.
[43] J. M. Brock and R. De Haas. “Discriminatory Lending: Evidence from Bankers in the Lab”. In: *American Economic Journal: Applied Economics* 15.2 (2023), 31–68.