Cloud-based XAI Services for Assessing Open Repository AI Models Under Adversarial Attacks

Zerui Wang, Yan Liu
Concordia University

Preprint: https://arxiv.org/abs/2401.12261
Avaliable: https://github.com/ZeruiW/XAIport
My Portfolio: https://deep-learning.ca/

Introduction to Explainable AI (XAI)

Explainable AI (XAI): The methods and techniques that provide insights into the decision-making processes of AI models, allowing users to comprehend and trust the results and actions of AI systems.

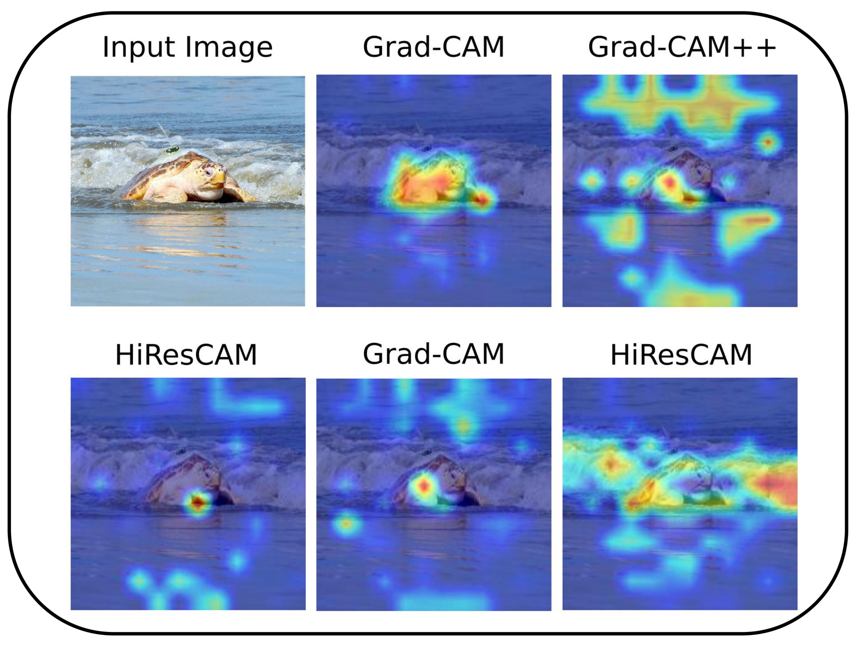

Figure 1. Saliency Map Visual Explanations for Vision Transformer Model with Image Example.

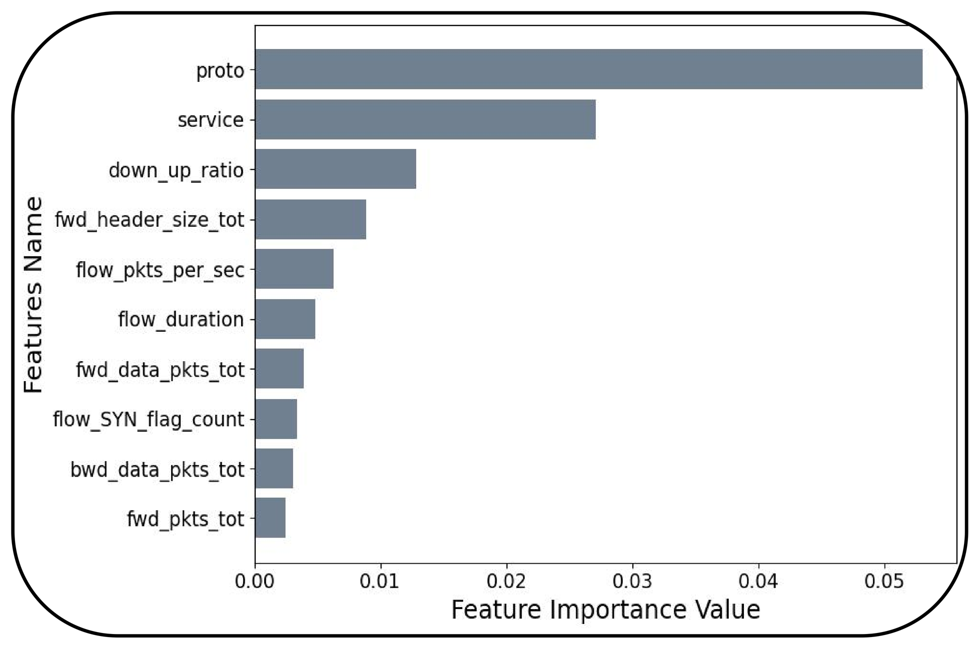

Figure 2. Top 10 out of 83 SHAP Feature Importance Explanations from FT Transformer on RT-IoT Cybersecurity Threats Dataset.

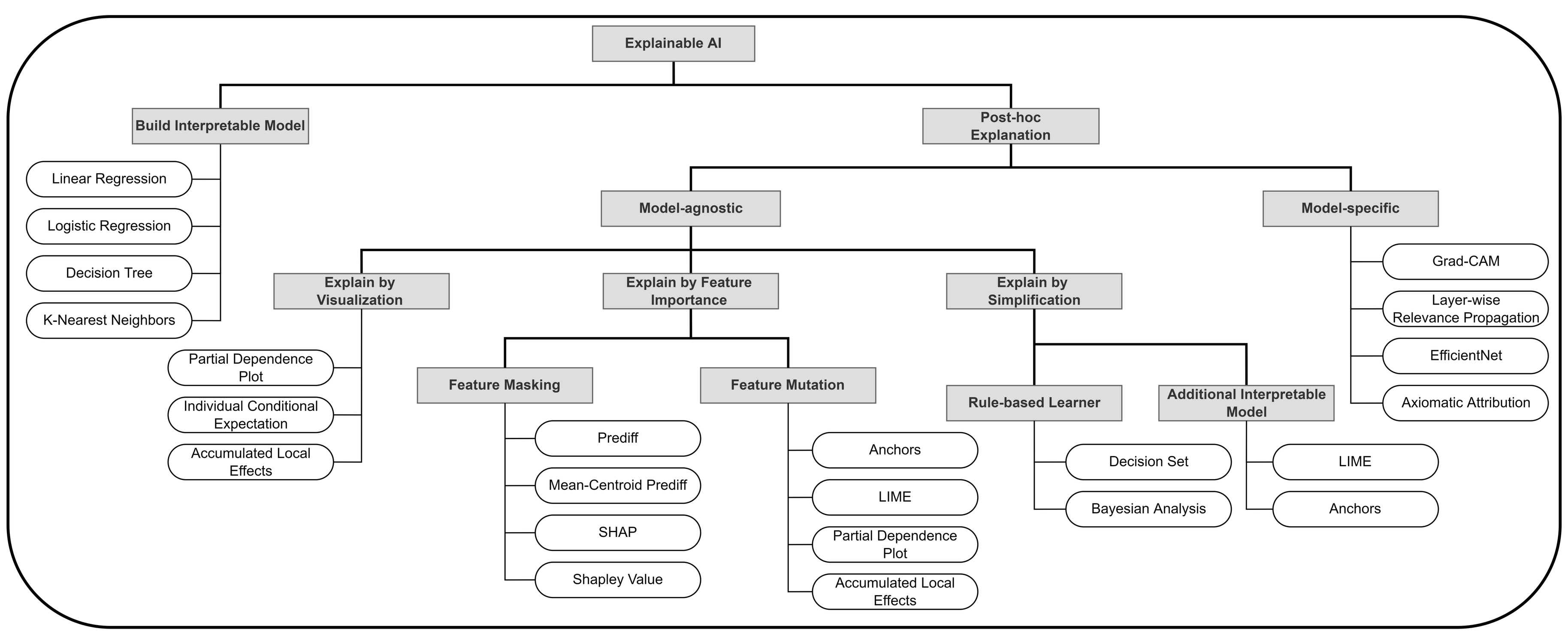

Figure 3. The Taxonomy of XAI Methods.

Research Questions

- *RQ1: Are the explanation deviation generated by XAI methods variable across models with different structures?*
- *RQ2: What is the relationship between computational cost and explanation deviation in model-XAI combinations?*
- *RQ3: Considering the known impacts of adversarial perturbations on model performance metrics, how do these perturbations influence the explanation deviation?*

Services Architecture for the Pipeline

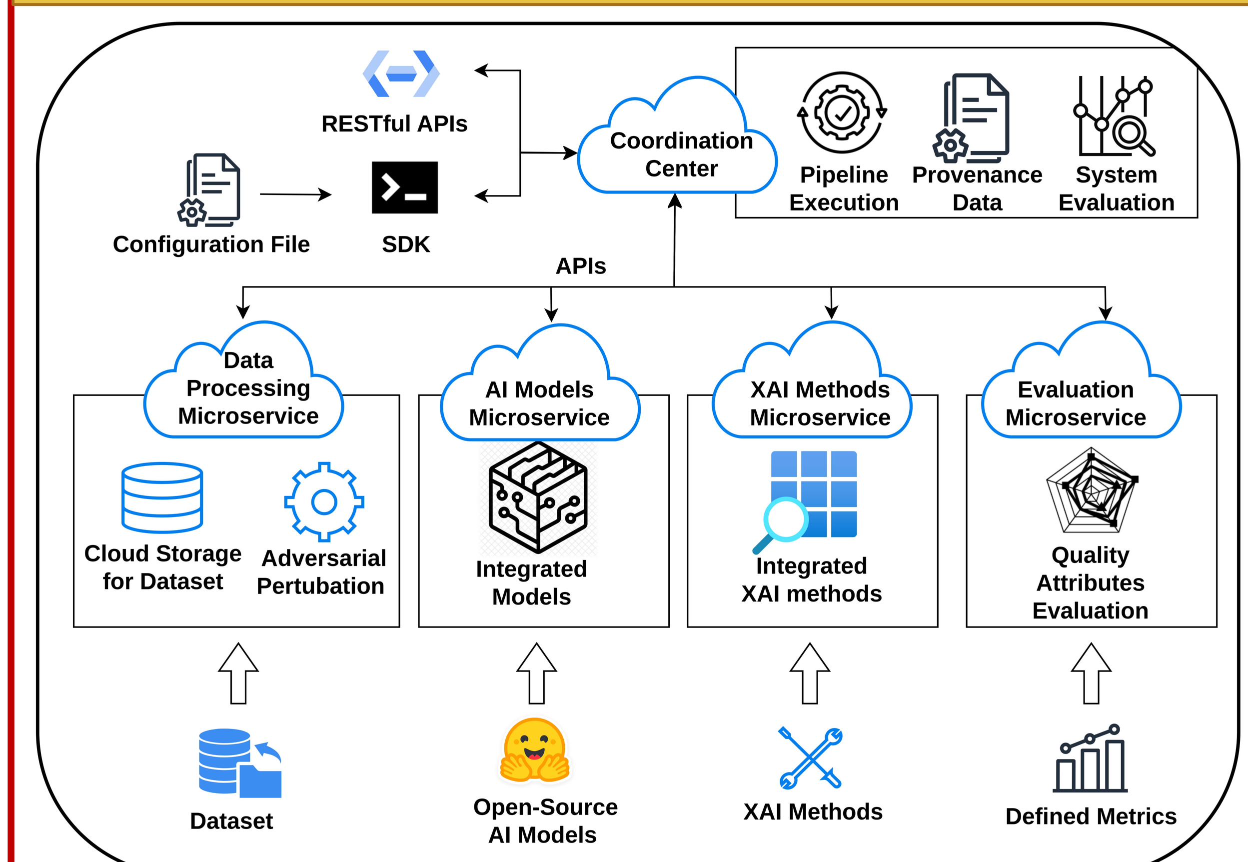

Figure 5. Cloud-based XAI Service Architecture.

Features:

1. **Adaptive Integration:** The architecture supports integration and testing of different AI models and XAI methods with flexibility.
2. **Comprehensive Evaluation:** Enables investigation of the combaination between AI models, datasets, and XAI methods, in defined metrics.
3. **Reusable Components:** Each microservice is designed for reusability, facilitating consistent and efficient reuse in multiple scenarios.

Components of the Architecture:

- **Coordination Center:** Manages operations, communication, and records data for transparency.
- **Data Processing:** Formats data and applies adversarial attack conditions.
- **Model Microservice:** Deploys pre-trained AI models, including community contributions.
- **XAI Method Microservice:** Provides explainable AI tools and algorithms for generating explanations.
- **Evaluation Microservice:** Aggregates results and evaluates quality attributes.

Introduction to Adversarial Attacks

Adversarial Attacks: The techniques by which an attacker creates inputs to a machine learning model that cause the model to make mistakes. These inputs are specially crafted by making small, often imperceptible, changes to the data that force the model to misclassify, mispredict, or otherwise fail to perform as intended.

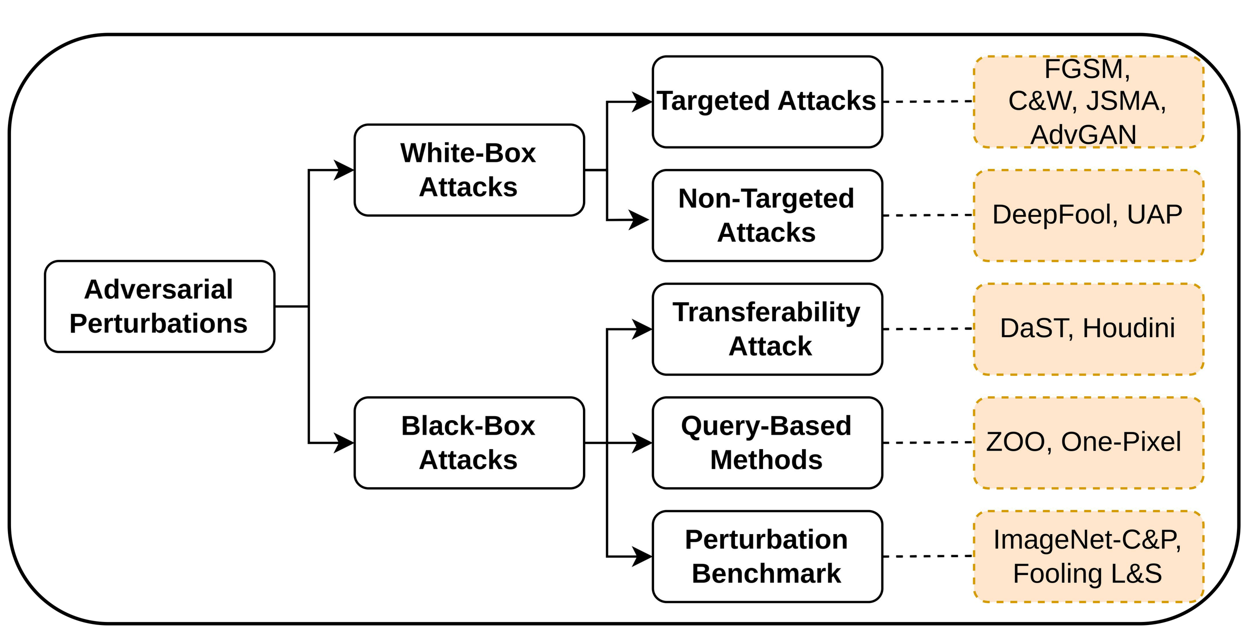

Figure 4. The Taxonomy of Adversarial Attack Methods.

White-box attacks,the attacker needs certain knowledge of the model, such as architecture and parameters.

Black-box attacks, where the attacker has no information about the model's internals, the focus shifts to transferability attacks, query-based methods, and perturbation benchmarks.

Assessment Scenarios

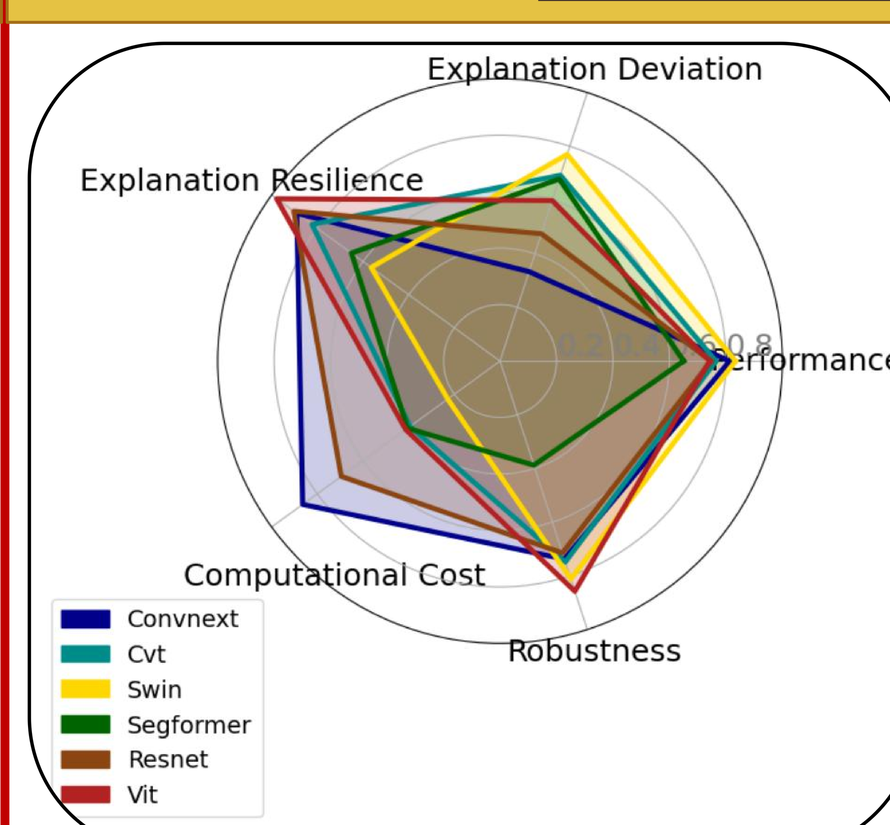

Figure 6. Comprehensive Overview of Quality Attributes Assessment for the Vision Models. **(RQ1)**

Table I. Explanation Deviation and Energy Consumption of the Selected Models. **(RQ2)**

Models(Vision/Tabular)	Explanation Deviation ↓	Energy (Wh) ↓
Swin	0.770	10.07
CVT	0.692	5.86
SegFormer	0.678	5.73
ViT	0.598	5.58
ResNet	0.474	3.32
ConvNeXt	0.333	2.67
TabNet	0.983	0.56
TabTransformer	0.977	0.49
FT Transformers	0.974	0.47

Table II. Explanation Deviation and Energy Consumption of the Selected Models. **(RQ3)**

Attribute	Significant ($p > 0.05$)	Non significant ($p \leq 0.05$)
Performance	88.89%	11.11%
Deviation	69.44%	30.56%

Pipelines of XAI Centric Assessment of Open Models Quality Attributes

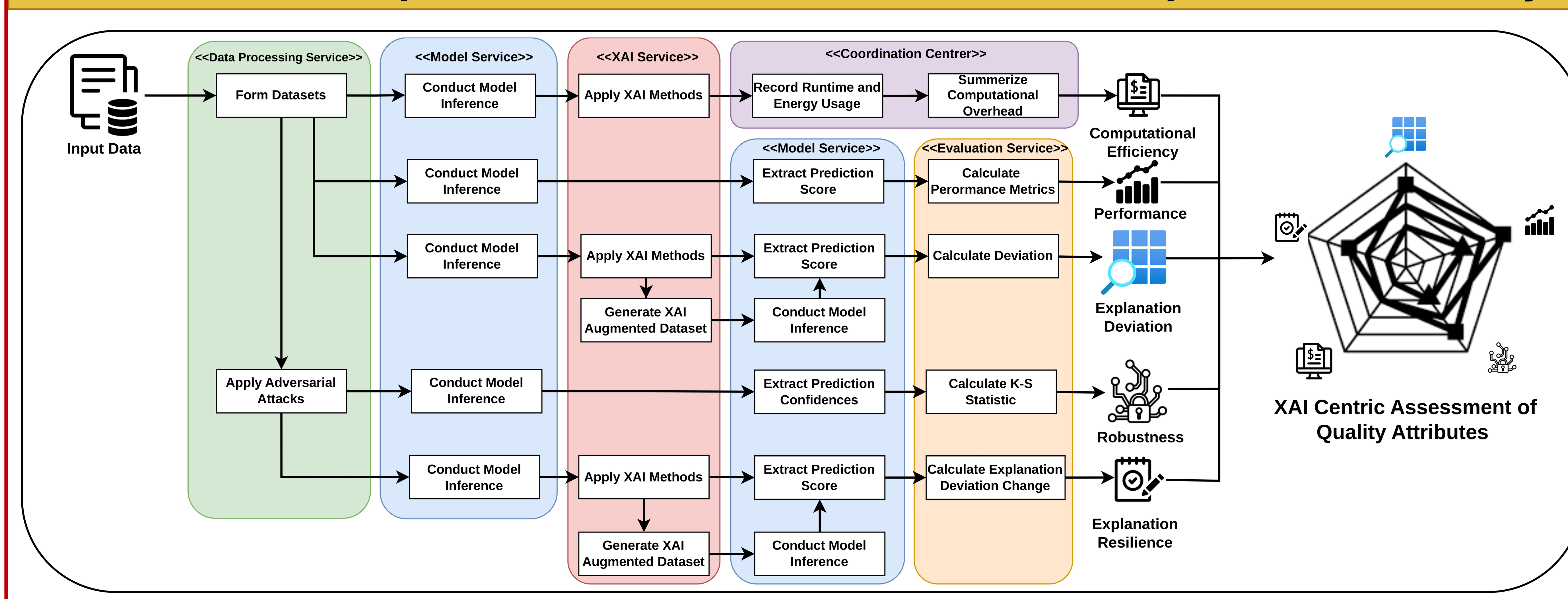

Figure 7. Assessment Pipelines for Open-source AI Model Quality Attributes.

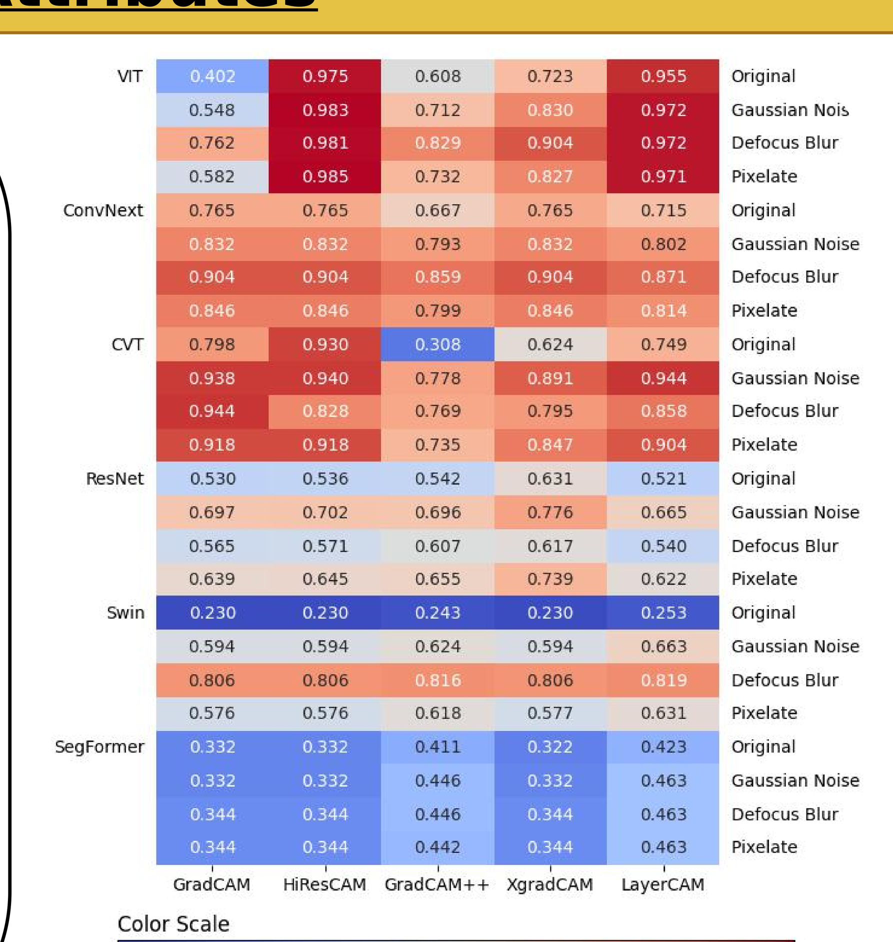

Figure 8. Heatmaps Illustrating Median Prediction Change Percentage for Original and Adversarial Perturbed Images. Lower Values Indicate Better Explanation Deviation.

REFERENCE

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
[2] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 ieee symposium on security and privacy (sp). Ieee, 2017, pp. 39–57.
[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016, pp. 372–387.
[4] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," arXiv preprint arXiv:1801.02610, 2018.
[5] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582.
[6] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1765–1773.
[7] M. Zhou, J. Wu, Y. Liu, S. Liu, and C. Zhu, "Dast: Data-free substitute training for adversarial attacks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 234–243.
[8] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," arXiv preprint arXiv:1707.05373, 2017.
[9] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in Proceedings of the 10th ACM workshop on artificial intelligence and security, 2017, pp. 15–26.
[10] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 828–841, 2019.
[11] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," arXiv:1903.12261, 2019.
[12] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.