ML-EvalPro: Machine Learning Evaluation Profiler for Supervised Tasks

Giza Systems

ML-EvalPro: Machine Learning Evaluation Profiler for Supervised Tasks

Mohamed Maher[1,2], Reem Ayman[1], Sondos Akram[1], Omar Marie[1], and Radwa ElShawi[2]

[1] Innovation Hub, Giza Systems
{reem.aymen,sondos.akram,omar.marie}@gizasystems.com
[2] Institute of Computer Science, University of Tartu, Tartu, Estonia
{mohamed.abdelrahman,radwa.elshawi}@ut.ee

Abstract. In today's rapidly evolving ML landscape, ensuring robust, fair, and ethical model usage is paramount. Achieving these objectives requires rigorous evaluation of performance, bias, variance, inference time, ethical considerations, and regulatory compliance, yet existing approaches often lack accessibility for non-technical stakeholders. Additionally, many models function as black boxes, complicating assessment when training data is unavailable—especially in medical contexts where privacy is a concern. ML-EvalPro addresses these challenges through a user-friendly platform that thoroughly evaluates black-box models without requiring direct access to training data. By demystifying complex metrics, it enables informed decision-making and promotes transparency, accountability, and ethical alignment. Ultimately, ML-EvalPro aims to foster safer, fairer, and more effective integration of machine learning in high-stakes domains such as healthcare.

Keywords: AutoML · Evaluation · Supervised Learning · LCNC.

1 Introduction

Rapid advances in machine learning (ML) have led to widespread adoption in critical domains, including healthcare [1]. However, many modern ML models function as "black boxes," leading to unintended biases, performance gaps, and ethical concerns. Traditional evaluation often centers on metrics like accuracy, precision, and recall, overlooking broader aspects such as fairness, interpretability, and regulatory compliance [2]. As models grow more complex, stakeholders like clinicians face increasing challenges in trusting and interpreting these outputs.

We introduce ML-EvalPro, a novel open-source platform that automates evaluation of supervised ML pipelines across performance, fairness, interpretability, and ethical considerations. Unlike existing fairness libraries (e.g., IBM Fairness 360[3] and Google's What-If Tool[4]), ML-EvalPro operates on black-box models

[3] https://aif360.res.ibm.com/
[4] https://pair-code.github.io/what-if-tool/

without requiring direct access to training data, thereby preserving privacy and confidentiality. Additionally, it provides a user-friendly interface geared toward domain experts, enhancing transparency and accountability in high-stakes contexts such as healthcare. This demonstration details `ML-EvalPro`'s architecture, key components, and functionalities, illustrating how it streamlines ethical, robust, and interpretable ML adoption across diverse supervised learning scenarios.

2 `ML-EvalPro` Architecture Modules

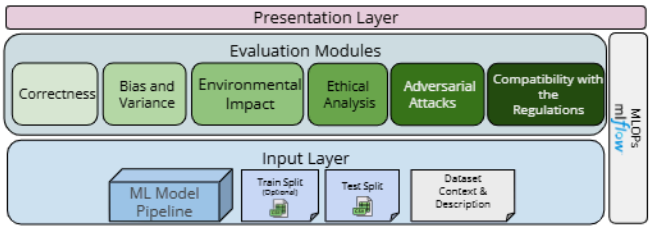


Fig. 1. The architecture of the `ML-EvalPro` package for evaluating supervised models.

The architecture of the `ML-EvalPro` system comprises six modules, including a) the performance evaluation module, which focuses on assessing correctness of the model predictions; b) the bias and variance examination module, which aims to detect bias and variance in the model; c) the environmental impact assessment module, which evaluates the ecological footprint of the model during inference, considering factors such as energy consumption; d) the ethical analysis module, which scrutinizes the ethical usage of input data features to ensure fairness in model predictions; e) the adversarial attack resilience module, which tests the model's robustness against potential adversarial attacks; and f) the regulatory compliance module, which ensures that the model operates under relevant legal standards.

A typical user flow starts by uploading a trained model (following the `mlflow`[5] interface) along with an unseen dataset split. `ML-EvalPro` currently supports models generated using Scikit-learn, H2O, PyTorch, ONNX, TensorFlow, Keras, XGBoost, LightGBM, CatBoost, Statsmodels, and Spark ML. Users indicate whether the model is for classification or regression and specify the target column. Optionally, they can also upload the original training split and the descriptions of input features. The system then processes these inputs through the six modules, ultimately presenting the results via a user-friendly dashboard designed to be interpretable by both technical and non-technical stakeholders.

[5] https://mlflow.org/

2.1 Performance Evaluation

ML-EvalPro provides a wide range of metrics for both classification and regression tasks. Users can conveniently select and apply these metrics to evaluate model performance. In addition, ML-EvalPro offers a reliability assessment feature wherein calibration curves (reliability diagrams) are plotted, and the Expected Calibration Error ([3] is computed for probabilistic classification models. This functionality helps quantify the divergence between model accuracy and confidence levels.

A dataset representing patient information for suggesting a treatment strategy. **Columns:** 1) **Age (Numerical):** ... 2) **Blood Pressure (Numerical):** Systolic blood pressure measured at admission. 3) **Ethnicity (Categorical):** Self-reported patient ethnicity.... N)**Treatment (Target):** Binary label indicating treatment strategy. **Sample rows:** (`65, 140, Hispanic, ..., A`), (`45, 125, Caucasian, ..., B`). Suppose **Ethnicity** is identified as the most influential predictors for a machine-learning model used to decide on treatment strategies. In this context, is it ethically sound to include them for decision-making?

Fig. 2. Example prompt for identifying potential ethical issues from related features.

2.2 Model Bias and Variance

To detect bias and variance in black-box ML models, ML-EvalPro supports two variance assessment approaches based on data availability. If the training set is provided, it compares performance on training and test splits to flag overfitting (high variance). If not, small perturbations are introduced to the unseen data, and large fluctuations in predictions suggest high variance.

For bias detection, the framework examines disparities across categorical features and computes the equalized odds [8] to ensure similar true/false positive rates across groups. For numerical features, K-means clustering (guided by an automated elbow rule with silhouette scoring) bins feature values. Significant performance differences among bins indicate bias toward specific feature ranges.

2.3 Environmental Impact

ML-EvalPro estimates the carbon footprint of model inference by measuring the average inference time per instance (t in seconds) and converting it into equivalent CO_2 emissions [4]. Specifically, the system computes how many predictions ($N = \frac{3600}{EF \cdot P \cdot t}$) produce 1 kg of CO_2, where P (in kW) is the power consumption of the hosting machine's CPU/GPU core[6] and EF (in kg CO_2/kW·hr) is the emission factor of energy generation [4].

[6] https://www.cpubenchmark.net/

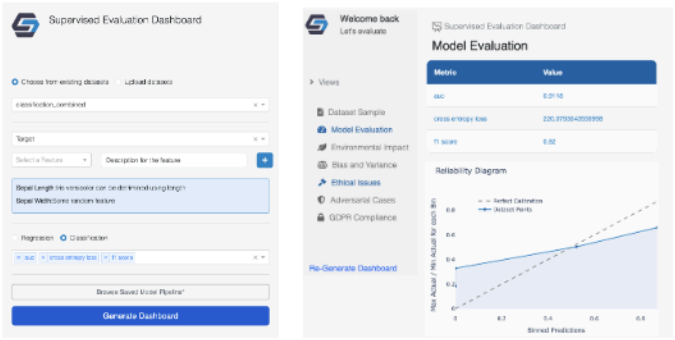


Fig. 3. The interface of `ML-EvalPro`. Framework inputs are on the Left. The model performance evaluation (module 1) is on the Right.

2.4 Ethical Use of Input Features

To evaluate ethical considerations around input features, `ML-EvalPro` employs SHAP [6] to identify globally influential features. Building on methods from [7], relevant context is appended to these features (e.g., domain-specific details, sample rows), and an LLM (Llama3.2 (3B) [7]) as a judge is prompted for ethical guidance (Figure 2). This process combines quantitative feature importance with qualitative user input, enabling transparency and accountability in model development, consistent with principles of responsible AI [9].

2.5 Robustness against Adversarial Attacks

`ML-EvalPro` evaluates susceptibility to adversarial attacks by first training a random forest surrogate model using predictions from the original black-box model. It then applies ZOO-based black-box attacks [5] to probe the surrogate's robustness. If these attacks transfer to the original model, the latter is deemed similarly vulnerable [10]. For regression tasks, the target variable is discretized via K-means clustering (similar to Section 2.3) so that classification-based adversarial methods can be applied, thereby providing a unified framework for both classification and regression models.

2.6 Regulations Compliance

`ML-EvalPro` generates a succinct report on `GDPR` compliance, highlighting interpretability, feature importance, and pipeline transparency. It leverages an LLM to detect potentially unethical features requiring user consent and to assess model reliability. While these results offer initial guidance on transparency, privacy, and reliability, the platform underscores that its compliance summary is advisory, not definitive, and should be supplemented with domain-specific legal expertise.

[7] https://huggingface.co/meta-llama/Llama-3.2-3B

3 Demo Scenario

ML-EvalPro is available both as a web application and a Python package.[8] In this demonstration, we showcase how users can upload a trained model, a testing dataset split, feature descriptions, and chosen evaluation metrics (Figure 3, left). The system then walks non-experts through various evaluation modules, offering a multifaceted assessment of model performance and reliability (Figure 3, right).

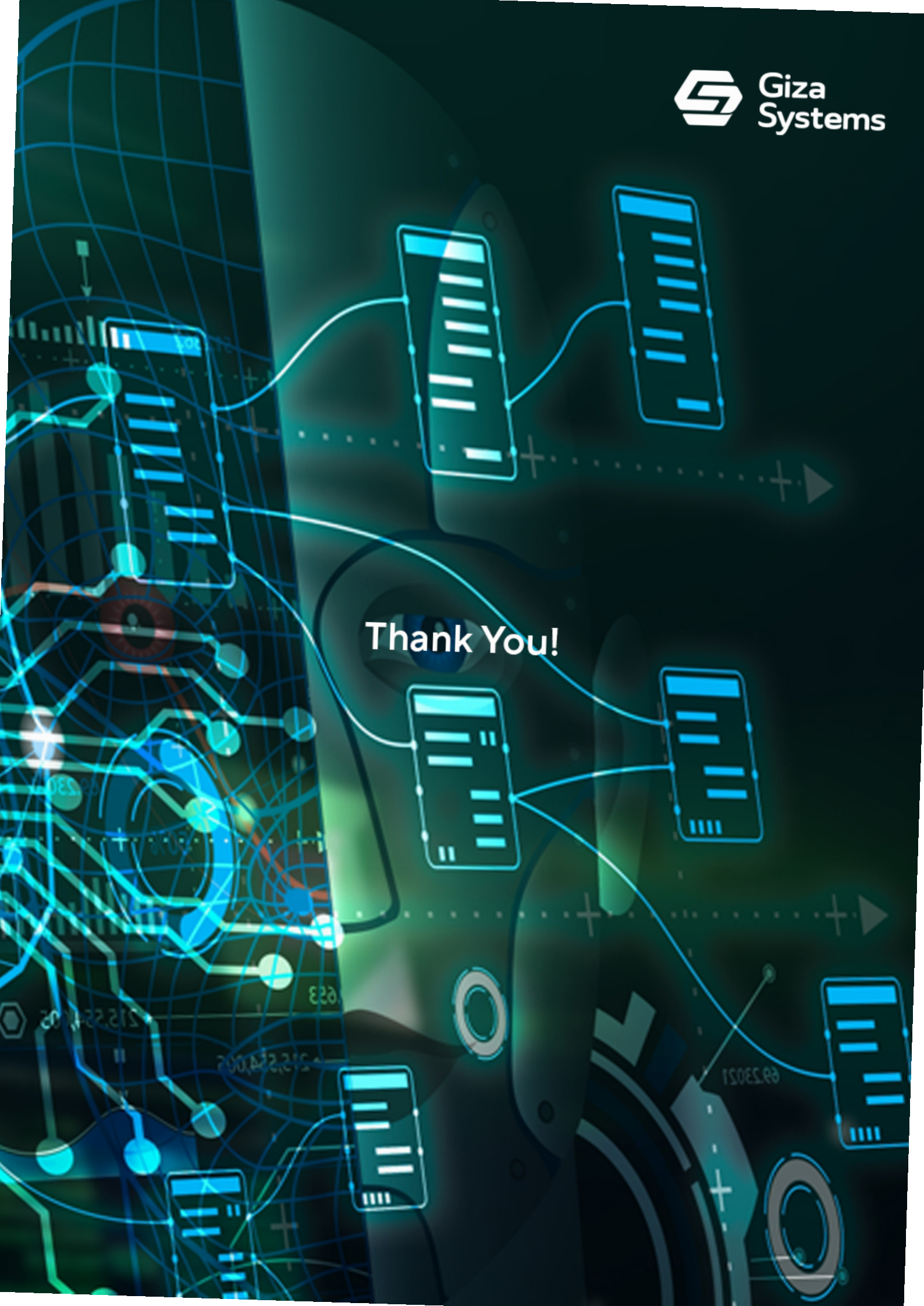
Throughout the interface, tooltips clarify assumptions—such as the proxy nature of bias metrics—and reduce the risk of misuse. We also provide an opportunity for real-time exploration across diverse models and datasets, demonstrating the platform's versatility.

Acknowledgments. This work was supported by the project "Increasing the knowledge intensity of Ida-Viru entrepreneurship" co-funded by the European Union and the innovation hub at Giza Systems.

References

1. Pugliese, R., Regondi, S., Marini, R.: Machine learning-based approach: Global trends, research directions, and regulatory standpoints. Data Science and Management **4**, 19–29 (2021). Elsevier
2. Kaul, S.: Speed and accuracy are not enough! Trustworthy machine learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 372–373 (2018)
3. Levi, D., Gispan, L., Giladi, N., Fetaya, E.: Evaluating and Calibrating Uncertainty Prediction in Regression Tasks. Sensors **22**(15), 5540 (2022). https://doi.org/10.3390/s22155540
4. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the Carbon Emissions of Machine Learning. arXiv preprint arXiv:1910.09700 (2019)
5. Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.-J.: 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, pp. 15–26 (2017). https://doi.org/10.1145/3128572.3140448
6. Lundberg, S.M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30, pp. 4765–4774 (2017)
7. Hollmann, N., Müller, S., Hutter, F.: Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. Advances in Neural Information Processing Systems **36** (2024)
8. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems **29** (2016)
9. Zheng, L., Chiang, W.-L., Sheng, Y., et al.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023)
10. Sadeghi, K., Banerjee, A., Gupta, S.: A System-Driven Taxonomy of Attacks and Defenses in Adversarial Machine Learning. IEEE Transactions on Emerging Topics in Computational Intelligence **4**, 450–467 (2020). https://doi.org/10.1109/TETCI.2020.2968932

[8] https://github.com/sondosakramm/ML-EvalPro


Giza
Systems
Thank You!