SustainaML: Enhancing Transparency, Control, and Green Sustainability in AutoML

Mehak Mushtaq Malik[1] and Radwa El Shawi(✉)[1]

University of Tartu, Estonia
{mehak.mushtaq.malik, radwa.elshawi}@ut.ee

Abstract. Automated machine learning (AutoML) enhances accessibility but often suffers from a lack of transparency and user control due to its complex and opaque processes. We introduce SustainaML, a lightweight visualization interface built atop FLAML, H2O, and MLJAR, enabling interactive refinement of AutoML search spaces and evaluation based on both performance and sustainability metrics. SustainaML offers flexible configurations and actionable visual feedback. A user study comparing SustainaML with ATMSeer demonstrates superior usability and effectiveness in promoting transparent, resource-efficient AutoML workflows.

1 Introduction

Automated Machine Learning (AutoML) speeds up model development by reducing the need for manual intervention and expert knowledge [2,4,9,1]. Open-source tools like H2O AutoML [7], FLAML [11], and MLJAR [5] enable rapid experimentation and model deployment across diverse domains. Despite these advances, most AutoML systems remain opaque, offering limited insight into decision processes [12]. This lack of transparency hinders trust, and the incorporation of domain knowledge—especially in complex workflows requiring informed or context-sensitive decisions. As a result, the collaborative potential between human experts and AutoML systems remains underexploited.

Recent tools aim to improve transparency and user control [3]. ATMSeer [12] supports visualization and refinement of hyperparameter tuning within the ATM framework. AutoAIViz [13] visualizes full pipeline evolution and comparisons, yet favors observability over steerability [6]. SigOpt [10], a commercial platform for Bayesian optimization, offers multi-metric tracking and convergence visualization but lacks access to internal pipeline processes or interactive configuration. Importantly, none of these tools consider environmental impact, such as energy use or CO_2 emissions [8]. To address this gap, we introduce `SustainaML`[1], an interactive online tool designed to improve transparency, steerability, and sustainability in AutoML workflows. It integrates sustainability metrics (energy consumption and CO_2 emissions) with performance evaluation, enabling users to explore trade-offs between performance and environmental impact. The tool allows real-time refinement of search spaces, comparison across frameworks (FLAML, H2O, MLJAR), and greater control over pipeline configurations.

[1] Source code is available on: https://github.com/DataSystemsGroupUT/SustainaML

2 SustainaML Overview

`SustainaML` is an interactive tool for refining AutoML search spaces and analyzing results. It supports pipeline comparison across metrics and frameworks with configurable search options. In addition to performance metrics, it integrates sustainability indicators, such as energy consumption (µWh) and CO_2 emissions (µg), measured via CodeCarbon's energy usage and carbon emission, enabling environmentally conscious model selection. We demonstrate its functionality on a classification task using the Heart Disease dataset [2,3].

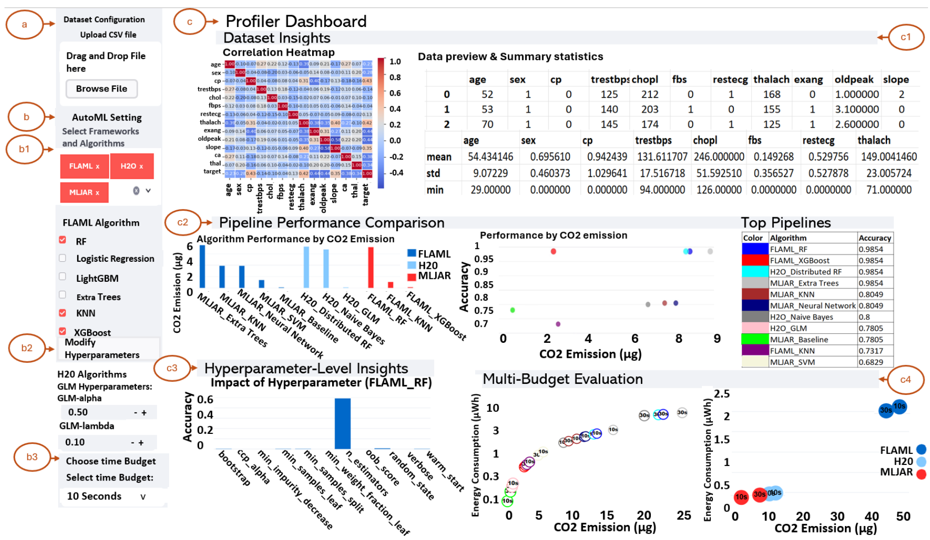

Fig. 1: Interface of SustainaML.

2.1 Interface and Demo Scenario

The interface consists of three panels: (a) dataset configuration, (b) AutoML settings, and (c) profiler dashboard (Figure 1). Panel (a) enables dataset upload. In (b1), users select frameworks (e.g., FLAML, H2O, MLJAR) and define their search spaces by specifying included algorithms (e.g., XGBoost, KNN, Random Forest for FLAML). Hyperparameters can be configured in (b2); otherwise, default search spaces are used. A time budget is set in (b3). After execution, panel (c) visualizes outcomes, providing dataset statistics (c1), pipeline comparisons (c2), hyperparameter analysis (c3), and multi-budget evaluations (c4).

[2] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
[3] Demo video: https://www.youtube.com/watch?v=I0nxUgzN4dI

2.2 Profiler Dashboard

The AutoML Profiler includes dataset insights (c1) and provides multi-level analysis of the AutoML process at the algorithm (c2), hyperparameter (c3), and time-budget (c4) levels. These visualizations help users interpret performance and sustainability trade-offs. `Dataset Insights (c1)` presents dataset-level statistics including minimum, mean, and standard deviation values for each feature, along with a feature correlation heatmap (see Figure 1(c1)). Users can identify strongly correlated features (e.g., slope and oldpeak), which informs feature relevance prior to model fitting. `Pipeline Performance Comparison (c2)` visualizes the performance of top pipelines across selected AutoML frameworks (e.g., FLAML, H2O, MLJAR) based on user-selected metrics. In this demo scenario, accuracy, energy consumption, and CO_2 emissions are selected to evaluate the pipelines of the selected frameworks. As shown in the middle plot of Figure 1(c2), `MLJAR_Extra Trees` and `H2O_distributed RF` pipelines achieve similar accuracy, but `H2O_distributed RF` incurs lower CO_2 emissions. These insights enable users to identify efficient pipelines that balance between predictive performance and sustainability. `Hyperparameter-Level Analysis (c3)` provides an on-demand breakdown of how hyperparameters affect performance. In this demo, we analyze the `FLAML_RF` pipeline, and as shown in the left plot of Figure 1(c3), the number of estimators has the greatest impact on accuracy. `Multi-Budget Evaluation (c4)` provides an on-demand breakdown that enables users to assess selected pipelines across varying time budgets based on user-defined metrics. In this demo, 11 pipelines are selected from the table in panel (c2) for comparative visualization of performance and resource consumption. The left plot in Figure 1(c4) displays CO_2 emissions versus energy consumption across selected time budgets, highlighting MLJAR_baseline as the most efficient pipeline at 10s and 30s. The right plot in panel (c4) compares the top-performing pipeline (by accuracy) from each framework—FLAML, H2O, and MLJAR—across time budgets highlighting trade-offs between performance and environmental impact.

3 User Study

User Study Setup: We conducted a user study to evaluate `SustainaML`, focusing on its support for algorithmic analysis, environmental assessment, and decision support. To assess usability, we benchmarked `SustainaML` against `ATMSeer`, comparing shared features and identifying unique capabilities. The study involved 15 participants with prior machine learning or data science experience, two of whom had used AutoML tools. The Heart Disease dataset[4] was used for evaluation. A 16-item survey captured user experience, including nine Likert-scale questions (Q1–Q9) on usability and seven open-ended questions (Q10–Q16) evaluating interpretability and decision support using `SustainaML`.

[4] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

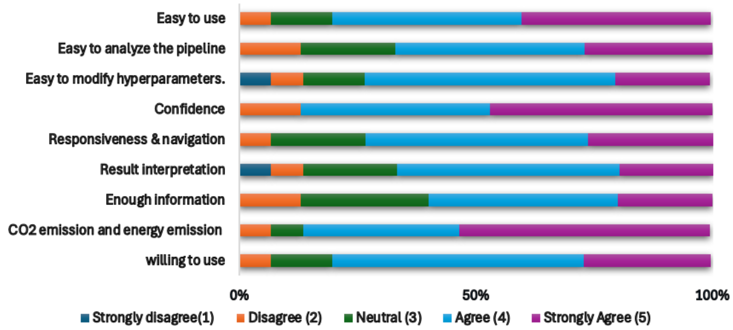

Fig. 2: Usability Results (Q1–Q9)

Table 1: Interpretability and Decision Support Results (Q10-Q16)

	Algorithm	**Environmental Impact**	**Feature Analysis**
QUESTIONS	Q10. Find best performing algorithm in terms of accuracy. **(15/15)**	Q13. Find the algorithm that consumes less energy and emits less CO_2. **(15/15)**	Q15. Find the feature that influences the most. **(14/15)**
	Q11. Find the stable algorithm. **(15/15)**	Q14. Find time budget's impact on algorithms' performance. **(15/15)**	Q16. Which feature is more important for the framework? **(12/15)**
	Q12. Which framework is more suitable for the specific dataset? **(15/15)**		

Procedure: Each session began with a brief demonstration of `SustainaML` and `ATMSeer`, followed by approximately 30 minutes of hands-on interaction, task completion, and feedback. For `SustainaML`, participants used three AutoML frameworks—FLAML, H2O, and MLJAR—across 20 algorithms.

Usability Results (Q1–Q9; see Figure 2): Participants rated nine aspects of usability on a five-point Likert scale. Users found `SustainaML` highly usable and informative: 87% agreed it was easy to use, 79% found it easy to analyze pipelines, and 73% agreed it was easy to modify hyperparameters. Confidence in decision-making was high (87%), and 84% rated the tool as responsive. Clarity in interpreting results was noted by 77%, while 80% felt the tool provided sufficient information and clearly conveyed energy consumption and CO_2 emissions. In comparison, usability ratings for `ATMSeer` were slightly lower: `SustainaML` outperformed `ATMSeer` in both confidence (87% vs 84.6%) and willingness to use (93% vs 92.3%).

Interpretability and Decision Support Results (Q10–Q16; see Table 1): Participants consistently identified the best-performing algorithm in terms of accuracy (15/15), selected the most stable algorithm (15/15), and chose the most suitable framework for the dataset (15/15). They accurately identified the algorithm with the lowest energy and CO_2 emissions (15/15), understood the impact of time budget on performance (15/15), and recognized the most influential feature (14/15). Additionally, 12 out of 15 correctly identified the key feature influencing framework selection.

Acknowledgments. This work was supported by the project "Increasing the knowledge intensity of Ida-Viru entrepreneurship" co-funded by the European Union.

References

1. El Shawi, R., Rozgonjuk, D.: Onlineautoclust: A framework for online automated clustering. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 3870–3874 (2023)
2. Eldeeb, H., Maher, M., Elshawi, R., Sakr, S.: Automlbench: a comprehensive experimental evaluation of automated machine learning frameworks. Expert Systems with Applications **243**, 122877 (2024)
3. ElShawi, R., Sakr, S.: csmartml-glassbox: Increasing transparency and controllability in automated clustering. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 47–54. IEEE (2022)
4. Maher, M., Oun, O.F., Mesmeh, M.S., Shawi, R.E.: Fedforecaster: An automated federated learning approach for time-series forecasting. In: Simitsis, A., Kemme, B., Queralt, A., Romero, O., Jovanovic, P. (eds.) Proceedings 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025. pp. 867–873. OpenProceedings.org (2025). https://doi.org/10.48786/EDBT.2025.70, https://doi.org/10.48786/edbt.2025.70
5. Mota, B., Faria, P., Ramos, C.: Automated machine learning and explainable artificial intelligence in predictive maintenance: An mljar framework review. In: Int. Symposium on Distributed Computing and Artificial Intelligence. Springer (2024)
6. Olson, R.S., Bartley, N., Urbanowicz, R.J., Moore, J.H.: Evaluation of a tree-based pipeline optimization tool for automating data science. In: Proceedings of the genetic and evolutionary computation conference 2016. pp. 485–492 (2016)
7. Prusty, S., Patnaik, S., Dash, S.K., Prusty, S.G.P., Rautaray, J., Sahoo, G.: Predicting cervical cancer risk probabilities using advanced h20 automl and local interpretable model-agnostic explanation techniques. PeerJ Computer Science **10** (2024)
8. Radersma, R.: Green coding: Reduce your carbon footprint. Ethical Software Engineering **19** (2022)
9. Sayed, E., Maher, M., Sedeek, O., Eldamaty, A., Kamel, A., El Shawi, R.: Gizaml: A collaborative meta-learning based framework using llm for automated time-series forecasting. In: EDBT. pp. 830–833 (2024)
10. Sorokin, A., Zhu, X., Lee, E.H., Cheng, B.: Sigopt mulch: An intelligent system for automl of gradient boosted trees. Knowledge-Based Systems **273**, 110604 (2023)
11. Wang, C., Wu, Q., Weimer, M., Zhu, E.: Flaml: A fast and lightweight automl library. Proceedings of Machine Learning and Systems **3**, 434–447 (2021)
12. Wang, Q., Ming, Y., Jin, Z., Shen, Q., Liu, D., Smith, M.J., Veeramachaneni, K., Qu, H.: Atmseer: Increasing transparency and controllability in automated machine learning. In: Proceedings of the 2019 CHI. pp. 1–12 (2019)
13. Weidele, D.K.I., Weisz, J.D., Oduor, E., Muller, M., Andres, J., Gray, A., Wang, D.: Autoaiviz: opening the blackbox of automated artificial intelligence with conditional parallel coordinates. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. pp. 308–312 (2020)