1 Broader Impact Statement

We identify two main ways in which our work may have a potential negative impact to society. They are both related to the scale of the benchmark and the resulting computational costs.

First, there is a significant computational cost, and thus carbon emissions, to evaluating an AutoML framework on the entire benchmarking suite. If both the 1 hour and 4 hour constraints are considered, this will be (4h+1h) * (101 datasets) * (10-fold CV) = 5000 wall-clock hours of compute (in our case, with 8 vCPU cores).

Second, the monetary cost to executing this evaluation may also be prohibitive. This naturally affects some people more than others, and can thus potentially reinforce the status quo if a full evaluation of this benchmark becomes the norm.

However, our intention is that people will be able to use experimental results across papers. With the experimental setup standardized, results reported in our benchmark, or those provided by more recent publications that use the benchmark, should be able to be re-used. We hope that the overall effect is a net positive (a reduction in costs), while allowing a more rigorous evaluation.

We are also looking into the effect of reducing the time constraints (and/or the use of early-stopping), which may further reduce the computational costs in the future.

Submission Checklist

1. For all authors…

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [Yes] See Section 5.3.

   (c) Did you discuss any potential negative societal impacts of your work? [No] The paper does not discuss this directly. Instead, the statement is provided in this document (previous page).

   (d) Did you read the ethics review guidelines and ensure that your paper conforms to them? `https://2022.automl.cc/ethics-accessibility/` [Yes] We have read the guidelines and conform to most of them. Since we cannot edit the paper from its journal publications, we indicate where we do not comply:

   - Not all figures have text larger than their caption size, but all figures and text are high definition and text remains sharp when zooming in.
   - We did do our best to select contrasting colors in our plots, unfortunately some plots do rely on consider exclusively to be interpretable.
   - We use the PDF from JMLR, so we cannot alter its metadata.

2. If you ran experiments…

   (a) Did you use the same evaluation protocol for all methods being compared (e.g., same benchmarks, data (sub)sets, available resources)? [Yes] As much as possible.

   All frameworks were evaluated on all datasets of the proposed suites, with the same cross-validation procedure and splits. Some frameworks were evaluated with multiple presets. Additional presets besides the default presets were chosen based on correspondence with the framework authors. All frameworks ran in a docker container on the same type of EC instances (m5.2xlarge).

   We tried to measure inference time for all generated models. For TPOT, we did not report inference times as it does not operate on raw data, this is mentioned in Section 6.3.

   In some cases, models exhibited unstable behaviour during the inference measurements phase. If we concluded that the regular evaluation (training + test) completed successfully but failure was related to the additional inference time measurements for Section 6.3, we would run the experiment again. This was a technical necessity: we could not communicate the inference time constraints to the framework, and at the same time had to limit the total time per compute job to avoid ballooning experimental costs caused by (potentially) halting processes. If it failed a second time, we would run it one last time without the inference time measurements routine. This affected all frameworks, but in total only a small number of experiments.

   Additionally, AutoGluon's 'high quality' presets explicitly ignore a portion of the time constraint, so the time constraint provided to that framework was adjusted accordingly, as outlined in Appendix C.2.1.

   Finally, we tried to evaluate as much as possible given our computational constraints. For the 4 hour time constraint, this meant that in some cases we report results obtained in previous experiments instead (part of the 2021 evaluation for the original submission). In those cases, changes between releases for the frameworks are small and unlikely to affect results in a meaningful manner.

(b) Did you specify all the necessary details of your evaluation (e.g., data splits, pre-processing, search spaces, hyperparameter tuning)? [Yes]

(c) Did you repeat your experiments (e.g., across multiple random seeds or splits) to account for the impact of randomness in your methods or data? [Yes] We performed 10-fold cross-validation for each of the datasets and time constraints. The AutoML framework's seed was set differently for each fold. We did not repeat the experiment with different splits because of the computational costs involved.

(d) Did you report the uncertainty of your results (e.g., the variance across random seeds or splits)? [Yes] In Appendix B we report the mean and variance across the 10 splits for each (dataset x framework x runtime constraint).

(e) Did you report the statistical significance of your results? [Yes] For the mean rank only (Figure 2).

(f) Did you use tabular or surrogate benchmarks for in-depth evaluations? [No] There are no tabular/surrogate benchmarks available that cover the search spaces for each AutoML framework.

(g) Did you compare performance over time and describe how you selected the maximum duration? [Yes] Partially. We decided to evaluate the frameworks on two different time constraints as a proxy for anytime performance (see also Figure 4 and Section 5.3.1). The choice of time constraints is somewhat historical (we started in 2018). The choice for a 1 hour constraint was based on the evaluation in Feurer et al. (2015), combined with the fact it was the default time-out period for a number of frameworks. We wanted a second more generous time limit to evaluate the usefulness of extending the time limit. We felt that 4 hours was a good trade-off between keeping the experiments computationally feasible and having enough extra time that we expected a noticeable difference.

(h) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

We mentioned the cloud provider (AWS), instance type (EC2 m5.2xlarge), and number of experiments and safety margins. The experimental results are composed of new evaluations (2023), and some old evaluations (2021). The new evaluations accounted for roughly 35 thousand hours of 'm5.2xlarge' time, and the old evaluations for approximately 44 thousand hours.

(i) Did you run ablation studies to assess the impact of different components of your approach? [No]

We evaluated the frameworks as a whole, as we expect most end-users to use them. As such, we can not attribute characteristics of the framework to any individual design decision. This limitation is explicitly mentioned in Section 5.3.

3. With respect to the code used to obtain your results…

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results, including all requirements (e.g., `requirements.txt` with explicit versions), random seeds, an instructive `README` with installation, and execution commands (either in the supplemental material or as a URL)? [Yes] Yes. See `https://openml.github.io/automlbenchmark/docs/getting_started/` for general usage and `https://openml.github.io/automlbenchmark/docs/using/reproducing/` for reproducibility.

(b) Did you include a minimal example to replicate results on a small subset of the experiments or on toy data? [Yes]

See https://openml.github.io/automlbenchmark/docs/using/reproducing/.

These guidelines are able to reproduce similar results: the exact generated data will be different, but it should lead to the same conclusion. Reproducibility in the sense of generating identical data is outside our influence, as we can do little more than provide the AutoML frameworks our random seed.

(c) Did you ensure sufficient code quality and documentation so that someone else can execute and understand your code? [Yes]

We have had individual contributors write integration for a new containerization framework (#88, i.e., https://github.com/openml/automlbenchmark/pull/88), new problem types (#494), AutoML frameworks (#211, #563), and, for individual use, modalities (#436). However, improving the state of documentation and code is an on-going process, and we identified a few key areas (#566, #279) which will make future use easier.

(d) Did you include the raw results of running your experiments with the given code, data, and instructions? [Yes] https://openml.github.io/automlbenchmark/results.html has links to all artifacts generated by our experiments.

(e) Did you include the code, additional data, and instructions needed to generate the figures and tables in your paper based on the raw results? [Yes] Both as a static repository with notebook and data files at https://github.com/PGijsbers/amlb-results, and as a code repository with interactive tool (does not yet generate all visualizations) at https://automlbenchmark.streamlit.app (https://github.com/PGijsbers/amlb-streamlit).

4. If you used existing assets (e.g., code, data, models)...

(a) Did you cite the creators of used assets? [Yes]

Each AutoML evaluated framework is cited, as is the SCIKIT-LEARN package that we use for our baselines and metric calculations. For the datasets, we provide links to our benchmarking suite on OpenML.

(b) Did you discuss whether and how consent was obtained from people whose data you're using/curating if the license requires it? [N/A]

For the datasets used, there were no known restrictions to their usage.

(c) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No new data was collected for these experiments, only already publicly available datasets were used.

5. If you created/released new assets (e.g., code, data, models)...

(a) Did you mention the license of the new assets (e.g., as part of your code submission)? [Yes] We released our code under the MIT license.

(b) Did you include the new assets either in the supplemental material or as a URL (to, e.g., GitHub or Hugging Face)? [Yes] URLs are included in the paper.

6. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or used human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use crowdsourcing or used human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use crowdsourcing or used human subjects.

7. If you included theoretical results…

   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We did not include theoretical results.

   (b) Did you include complete proofs of all theoretical results? [N/A] We did not include theoretical results.

References

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in neural information processing systems*, 28.