Algorithm selection on a meta level

Alexander Tornede[1] **Lukas Gehring**[1] **Tanja Tornede**[1] **Marcel Wever**[2] **Eyke Hüllermeier**[2]

[1]Heinz Nixdorf Institut, Paderborn University, 33098, Paderborn, Germany
[2]University of Munich (LMU), 80538, Munich, Germany

Abstract

1 Reproducibility 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] We described the methodological framework mentioned in the abstract in Sections 3-5 and present the superior performance of the methods presented inside this framework in Section 6, in particular Section 6.6.1.

   (b) Did you describe the limitations of your work? [Yes] We describe the limitations of each presented approach within the corresponding subsection of Section 5. Moreover, we discuss the scope of our results and in particular the limitation of conclusions drawn from it in Section 6.7.1.

   (c) Did you discuss any potential negative societal impacts of your work? [No] This is not done within the workshop paper. However, we do this as part of the broader impact statement for this submission.

   (d) Have you read the ethics author's and review guidelines and ensured that your paper conforms to them? `https://automl.cc/ethics-accessibility/` [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

   (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 version), an instructive `README` with installation, and execution commands (either in the supplemental material or as a URL)? [Yes] We have a detailed README with instructions on how to run the code and reproduce the results as part of our GitHub repository linked in the paper: `https://github.com/alexandertornede/as_on_a_meta_level`

   (b) Did you include the raw results of running the given instructions on the given code and data? [No] The results are stored in a database. However, we did not include a raw dump of the results.

   (c) Did you include scripts and commands that can be used to generate the figures and tables in your paper based on the raw results of the code, data, and instructions given? [Yes] They can be found in the README.

(d) Did you ensure sufficient code quality such that your code can be safely executed and the code is properly documented? [Yes] The code can be safely executed. While documentation in the form of the README is extensive, inline code documentation could certainly be improved.

(e) Did you specify all the training details (e.g., data splits, pre-processing, search spaces, fixed hyperparameter settings, and how they were chosen)? [Yes] All training details can be found within the code. Moreover, if we changed any default parameters, we noted this.

(f) Did you ensure that you compared different methods (including your own) exactly on the same benchmarks, including the same datasets, search space, code for training and hyperparameters for that code? [Yes] All methods were compared on the exact same splits provided by ASlib such that the results are truly comparable.

(g) Did you run ablation studies to assess the impact of different components of your approach? [Yes] For each approach presented, we exchanged some of it components with different options in order to see how this influences performance. Details can be found in Section 6.

(h) Did you use the same evaluation protocol for the methods being compared? [Yes] All methods were compared on the exact same splits provided by ASlib using the same underlying evaluation code such that the results are truly comparable.

(i) Did you compare performance over time? [N/A]

(j) Did you perform multiple runs of your experiments and report random seeds? [Yes] We performed a ten-fold cross-validation as suggested by ASlib. The used random seed corresponds to the fold number. This information be found in the code.

(k) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]

(l) Did you use tabular or surrogate benchmarks for in-depth evaluations? [Yes] We used ASlib, which can be considered a tabular benchmark.

(m) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

(n) Did you report how you tuned hyperparameters, and what time and resources this required (if they were not automatically tuned by your AutoML method, e.g. in a NAS approach; and also hyperparameters of your own method)? [N/A] We did not tune hyperparameters.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We use ASlib and cite the authors.

(b) Did you mention the license of the assets? [No]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, we provide some meta AS datasets as part of the GitHub repository.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The ASlib benchmark is freely available.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. 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]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]