TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization

TAILOR

Grant Agreement Number 952215

D7.7 AutoAI Benchmarks v.3 Report

Document type (nature)	Report
Deliverable No	7.7
Work package number(s)	7
Date	Due 1 July 2024
Responsible Beneficiary	ULEI, ID #7
Author(s)	Annelot Bosman, Holger Hoos
Publicity level	Public
Short description	Version 3 of AutoAI Benchmarks: Curated, regular evaluations of AutoAI techniques and their contribution to trustworthiness, to measure and monitor progress in the field.

History			
Revision	**Date**	**Modification**	**Author**
v.1 0	2024-11-01	-	Annelot Bosman, Holger Hoos, Joaquin Vanschoren, Jan N. van Rijn, Mitra Barachi, Dragi Kocev

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
André Meyer-Vitali	DFKI	29-10-2024
Carla Pacheco	IST-UL	21-10-2024

Table of Contents

Summary of the report

This report gives an overview of existing benchmarks in the area of automated AI (AutoAI), and more specifically, in the areas of the five AutoAI topics covered in T7.1-T7.5 of the TAILOR project.

Organisation

Partner ID / Acronym	Name	Role
#7, ULEI	Holger Hoos Koen van der Blom Mitra Baratchi Jan N. van Rijn Laurens Arp	WP7, T7.2, T7.3 Leader
#17, ALU-FR	Frank Hutter Eddie Bergman	Task Leader
#14, JSI	Dragi Kocev	Participant
#12, TUE	Joaquin Vanschoren Pieter Gijsbers	T7.5 Task Leader

1. Introduction

The development of AutoAI techniques and assessment of their quality in terms of performance and trustworthiness hinges on high-quality benchmarks. Many AI systems are manually constructed or configured to perform a specific task, AutoAI aims to automate (parts of) this construction process. Take, for instance, the widely studied AutoML task of automated selection of classification algorithms and the optimisation of their hyperparameters (e.g., AutoWEKA), or the automated optimisation of machine learning pipelines which consider the same type of approaches (e.g., Auto-sklearn).

When it comes to AI benchmarks, they must only enable the comparison of AI systems, whereas AutoAI benchmarks need to enable the operation of the AutoAI techniques in addition to enabling the comparison of the AI systems resulting from using AutoAI. For the example above this is the difference between needing just training data for the ML algorithms and needing to also specify the AutoML scenario. This scenario includes things such as which ML algorithms are considered and for instance the budget available to tune the hyperparameters. So, although AutoAI benchmarks may share a common part with regular AI benchmarks, they also require additional components.

As part of WP7, one of the goals is to create awareness about the available AutoAI benchmarks, and another goal is to identify gaps for the development of new benchmarks to

complement existing work. To this end we aim to capture the most important benchmarks here, but this is not necessarily an exhaustive list.

In the following, existing AutoAI benchmarks are given per research task in WP7; this was done to ensure broad coverage. In addition, non-AutoAI benchmarks with the potential to be extended to AutoAI are listed, with particular attention to areas where no AutoAI benchmarks are available yet. Work from TAILOR network members is highlighted in boldface, both for the AutoAI and non-AutoAI benchmarks.

2. AutoML in the wild [T7.2, ALU-FR]

This task aims to facilitate the usability of machine learning by non-experts. Our efforts in TAILOR concentrate on two research lines:

1. Making Neural Architecture Search (NAS) usable in the wild

2. Design methods to automatically handle messy real-world data in AutoML.

2.1 Neural Architecture search

In Neural Architecture Search (NAS) benchmarking is particularly heavily used. Task leader Frank Hutter, together with collaborators from Google, introduced the first tabular NAS benchmark, NAS-Bench-101. The research community is using this benchmark heavily and created almost a dozen new tabular NAS benchmarks since. In the following, we typeset NAS benchmarks by TAILOR participants in boldface.

- **[YingEtAl19]** Chris Ying, Aaron Klein, Esteban Real, Eric Christiansen, Kevin Murphy, Frank Hutter, "Nas-bench-101: Towards reproducible neural architecture search," Proceedings of the International Conference on Machine Learning, 2019.

- [DongYang19] Xuanyi Dong and Yi Yang, “NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search,” Proceedings of the International Conference on Learning Representations, 2019.

- [ZiEtAl21] Chaojian Li, Zhongzhi Yu, Yonggan Fu, Yongan Zhang, Yang Zhao, Haoran You, Qixuan Yu, Yue Wang, Yingyan Lin. HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark. Proceedings of the International Conference on Learning Representations, 2021.

- [KlyuchnikovEtAl20] Nikita Klyuchnikov, Ilya Trofimov, Ekaterina Artemova, Mikhail Salnikov, Maxim Fedorov, Evgeny Burnaev. “NAS-Bench-NLP: Neural Architecture Search Benchmark for Natural Language Processing”, arXiv 2020

- [MehrotraEtAl21] Abhinav Mehrotra, Alberto Gil C. P. Ramos, Sourav Bhattacharya, Łukasz Dudziak, Ravichander Vipperla, Thomas Chau, Mohamed S Abdelfattah, Samin Ishtiaq, Nicholas Donald Lane. “NAS-Bench-ASR: Reproducible Neural Architecture Search for Speech Recognition”, Proceedings of the International Conference on Learning Representations, 2021.

- **[ZelaEtAl20]** Arber Zela, Julien Siems, Frank Hutter. “NAS-Bench-1Shot1: Benchmarking and Dissecting One-shot Neural Architecture Search”. Proceedings of the International Conference on Learning Representations, 2020.
- **[SiemsEtAl20]** Julien Siems, Lucas Zimmer, Arber Zela, Jovita Lukasik, Margret Keuper, Frank Hutter. “NAS-Bench-301 and the Case for Surrogate Benchmarks for Neural Architecture Search”, arXiv 2020
- **[YanEtAl21]** Shen Yan, Colin White, Yash Savani, Frank Hutter. “NAS-Bench-x11 and the Power of Learning Curves”, CVPR workshop on NAS 2021.

Currently, work is underway at the University of Freiburg to provide all these NAS benchmarks through a unified interface.

Related to work on NAS is hyperparameter optimisation, and there are several benchmarks available:

- **HPOlib: https://github.com/automl/HPOlib**
- **HPOBench: https://github.com/automl/HPOBench**
- Bayesmark: https://github.com/uber/bayesmark

There is also a related benchmark on dynamic algorithm configuration (DAC), which covers the dynamic optimization of hyperparameters:

https://github.com/automl/DACBench

- **[EimerEtAl21]** Theresa Eimer, Andre Biedenkapp, Maximilian Reimer, Steven Adriaensen, Frank Hutter, Marius Lindauer, DACBench: A Benchmark Library for Dynamic Algorithm Configuration, IJCAI 2021

Regarding the handling of messy data, there are also several efforts:

- The Data wrangling dataset repository: http://dmip.webs.upv.es/datawrangling/catalog.html
- Spreadsheets: Benchmarks and quality assurance techniques for spreadsheets: https://spreadsheets.ist.tugraz.at/

Furthermore, a paper was published on handling messy data, especially string categorical features, at the ECMLPKDD Workshop on Automated Data Science:

- **[LithVanschoren21]** John W. van Lith and Joaquin Vanschoren. From strings to data science: practical framework for automated string handling. ECMLPKDD workshop on Automated Data Science, 2021.

There has been a large increase in Neural Architecture Search (NAS) benchmarks over the last two years, mostly aimed at moving beyond tabular setups and using surrogate models to predict the performance of architectures, allowing us to extend NAS benchmarks to larger spaces.

- **[ZelaEtAl22]** Zela, Arber; Siems, Julien; Zimmer, Lucas; Lukasik, Jovita; Keuper, Margret; Hutter, Frank, “Surrogate NAS Benchmarks: Going Beyond the Limited Search Spaces of Tabular NAS Benchmarks”, In: International Conference on Learning Representations (ICLR), 2022.

We have also made substantial progress in developing NASLib to provide a uniform interface to all available NAS benchmarks and allow researchers and practitioners to directly validate their methods over a variety of search spaces and conditions.

- **NASLib**: https://github.com/automl/NASLib

In the last report, we introduced HPOBench; since then, the benchmark has been extended to integrate additional new benchmarks. More importantly, we are actively working towards multi-objective benchmarks, a fundamental test bed for research on new multi-objective optimizers, a trending requirement of real AutoML systems in the wild.

- **[EggenspergerEtAl21]** Eggensperger, Katharina; Müller, Philipp; Mallik, Neeratyoy; Feurer, Matthias; Sass, René; Klein, Aaron; Awad, Noor; Lindauer, Marius; Hutter, Frank, “HPOBench: A Collection of Reproducible Multi-Fidelity Benchmark Problems for HPO”, In: Vanschoren, J.; Yeung, S. (Ed.): Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021. https://github.com/automl/HPOBench

2.2 Physics applications

In high energy physics (HEP), jets are collections of correlated particles produced ubiquitously in particle collisions such as those at the CERN Large Hadron Collider (LHC). Machine learning (ML)-based generative models, such as generative adversarial networks (GANs), have the potential to significantly accelerate LHC jet simulations. However, despite jets having a natural representation as a set of particles in momentum-space, a.k.a. a particle cloud, there exist no generative models applied to such a dataset.

In recent work**,** we have introduced and released a new particle cloud dataset to serve as a benchmark in applications of ML in high energy physics (HEP) and more specifically in accelerating Large Hardon Collider jet simulations **[KansalEtAl21].**

Existing GANs are found to be inadequate for physics applications, hence we develop a new message-passing GAN (MPGAN), which outperforms existing point cloud GANs on virtually every metric and shows promise for use in HEP. We propose JetNet as a novel point-cloud-style dataset for the ML community to experiment with and set MPGAN as a benchmark to improve upon for future generative models. Additionally, to facilitate research and improve accessibility and reproducibility in this area, we release the open-source JetNet Python package with interfaces for particle cloud datasets, implementations for evaluation and loss metrics, and more tools for ML in HEP development.

- JetNet: https://zenodo.org/record/6302454

Another benchmark dataset has been created for the removal of clouds in optical remote sensing (satellite) imagery. Satellite data is a prominent and impactful example of data that can be messy in the wild, due to its susceptibility to noise (e.g., sensor faults, solar glint), but particularly due to the problems caused by cloud cover. Although not an explicit AutoAI benchmark on its own, the dataset allows for the systematic evaluation of cloud removal methods, which often form a key part of the data processing pipeline in remote sensing applications, with implications for AutoAI pipelines in particular. Moreover, satellite data is inherently diverse, with highly variable types of environments affecting which method (in this case for cloud removal) performs best, usually rendering single, general models infeasible making it an interesting area to apply AutoAI.

- **SEN2-MSI-T (working dataset title): Arp, Laurens; Baratchi, Mitra; Hoos, Holger; Van Bodegom, Peter; Francis, Alistair; Wheeler, James, "Model-free cloud removal for ground-level Sentinel-2 imagery using value propagation interpolation"**

We have also created a new benchmark dataset for validating super-resolution (SR) methods (methods for increasing the resolution of images) for satellite image datasets. We introduce a new real-world single-image SR dataset, SENT-NICFI **[WasalaEtAl]**. Many SR methods are currently trained and evaluated on synthetic datasets, which require matching images obtained from different sensors. Synthetic datasets are created under a scale invariance assumption using downsampling procedures, and therefore, this approach can misrepresent the information captured in the image. The performance of a model trained on synthetic data might overestimate the model's performance on real-world data. Therefore, we also introduce a new dataset consisting of matching lower-resolution Sentinel-2 and higher-resolution Planet images, called SENT-NICFI. SENT-NICFI is a real-world dataset that is expected to address these problems. This dataset can be used widely for evaluating general algorithms and AutoAI methods for SR. In our research, we validate our own proposed AutoML Approach to SR for Earth Observation Images.

Furthermore, we have prepared six datasets containing telemetry data of the Mars Express Spacecraft (MEX), a spacecraft orbiting Mars operated by the European Space Agency. The data, consisting of context data and thermal power consumption measurements, capture the status of the spacecraft over three Martian years, sampled at six different time resolutions that range from 1 min to 60 min. Given the heterogeneity, complexity, and magnitude of the data, they can be employed in a variety of scenarios and analysed through the prism of different machine learning tasks, such as multi-target regression, learning from data streams, anomaly detection, clustering, etc. Analysing MEX's telemetry data is critical for aiding very important decisions regarding the spacecraft's status and operation, extracting novel knowledge, and monitoring the spacecraft's health, but the data can also be used to benchmark artificial intelligence methods designed for a variety of tasks.

- Data: https://doi.org/10.6084/m9.figshare.c.5360420.v1

- **[PetkovićEtAl22]** Petković, M., Lucas, L., Levatić, J., Breskvar, M., Stepišnik, T., Kostovska, A., ... & Kocev, D. (2022). Machine-learning ready data on the thermal power consumption of the Mars Express Spacecraft. *Scientific Data*, *9*(1), 229.

Finally, we have developed a FAIR (Findable, Accessible, Interoperable, and Reusable) catalogue with semantic annotations of multilabel datasets **[KostovskaEtAl22]**. Multilabel classification (MLC) is a machine learning task where the goal is to learn to label an example with multiple labels simultaneously. Considering the increased interest in this task from the research community, ensuring proper, correct, robust, and trustworthy benchmarking is of utmost importance for the further development of the field. We believe that this can be achieved by adhering to the recently emerged data management standards, such as the FAIR and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles. We introduce an ontology-based online catalogue of MLC datasets originating from various application domains following these principles. The catalogue extensively describes many MLC datasets with comprehensible meta-features, MLC-specific semantic descriptions, and different data provenance information. The MLC data catalogue is available at: http://semantichub.ijs.si/MLCdatasets.

3. Beyond standard supervised learning [T7.2, ULEI]

So far, work on AutoML (an important special case of AutoAI) has largely been limited to standard supervised learning scenarios on tabular data. This “Beyond standard supervised learning” task aims to expand the scope of AutoML to more diverse and richer learning settings. To achieve this, two lines of research are being pursued:

1. Bring together AutoML- and domain experts to design flexible AutoML frameworks - including algorithm configuration and meta-learning - for domains such as multi-target regression, unsupervised learning, semi-supervised learning and learning on spatio-temporal data.

2. Collect and make publicly available (whenever possible) benchmark data and scenarios for these new settings, building on existing repositories and libraries.

Work on multi-target regression is limited, but a benchmark framework for multi-label classification was proposed by:

- [WeverEtAl21] Wever M, Tornede A, Mohr F, Hullermeier E. AutoML for Multi-Label Classification: Overview and Empirical Evaluation. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2021: https://ieeexplore.ieee.org/abstract/document/9321731

Some initial work on semi-supervised AutoML exists, but it is not very clear which datasets were used or how they were adapted to fit the semi-supervised+AutoML scenario, nor was anything made publicly available for future benchmarking use. For instance:

- [LiEtAl19] Li, Y.-F., Wang, H., Wei, T., & Tu, W.-W. (2019). Towards Automated Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 4237-4244. https://doi.org/10.1609/aaai.v33i01.33014237

TUE also proposed benchmarks for unsupervised AutoML, specifically clustering and outlier detection, and invented a way to do unsupervised AutoML via meta-learning (more on this in Section 6). The benchmarks can be found in:

- **[SinghVanschoren23a]** Singh P, and Vanschoren J. AutoML for outlier detection with optimal transport distances. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23), 2023.

- **[SinghVanschoren23b]** Singh P, and Vanschoren J. Applications of Optimal Transport Distances in Unsupervised AutoML NeurIPS 2023 Workshop Optimal Transport and Machine Learning. 2023

TUE also presented a benchmark for AutoML on time series, and performed a benchmark of all current AutoML systems that support time series:

- **Thirthapura Sreedhara A. Can Time Series Forecasting be Automated? A Benchmark Analysis. MSc Thesis, Eindhoven University of Technology.**

Spatio-temporal learning does not yet seem to have AutoML benchmarks in place and is thus one direction to focus on in further work in WP7 of TAILOR. Fortunately, a variety of existing (non-AutoML) work exists that can be built upon.

For human trajectory prediction, there are currently no benchmarks available that are suited for AutoAI research. However, there have been efforts to identify publicly available datasets that can be used for this purpose. For instance:

- [AmirianEtAl20] Amirian J, Zhang B, Castro FV, Baldelomar JJ, Hayet JB, Pettré J. Opentraj: Assessing prediction complexity in human trajectories datasets. InProceedings of the Asian Conference on Computer Vision 2020. https://github.com/crowdbotp/OpenTraj

For other spatio-temporal AutoML applications, spatial datasets could be used to develop AutoML benchmarks. Such datasets include:

- AIREO for Earth observation data: https://eo4society.esa.int/projects/aireo/

In addition, as part of the following publication, we have identified 7 datasets (and collected them in a single repository) that can be used for AutoML research for remote sensing image classification tasks. This dataset should make it easier for others to perform research in this area:

- **[SalinasEtAl21]** Palacios Salinas, N. Rosaura and Baratchi, M. and van Rijn, J. N. and Vollrath, A, "Automated Machine Learning for Satellite Data: Integrating Remote Sensing Pre-trained Models into AutoML Systems", in Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML-PKDD 2021, 2021. https://github.com/palaciosnrps/automl-rs-project

Other domains with existing (non-AutoML) benchmarks that could be extended include:

- High-dimensional particle and astrophysics optimisation benchmark: https://arxiv.org/abs/2101.04525

- OpenML: https://www.openml.org/

- OpenAI gym: https://gym.openai.com/

- OpenAI baselines: https://github.com/openai/baselines
- DeepMind's RL benchmark problem library b-suite: https://deepmind.com/research/open-source/bsuite

In addition, in the future, novel real-world datasets for AutoML benchmarking may be introduced through further work with partners from the industry.

For these new datasets, it should be investigated how they might be integrated with existing AutoML benchmarking libraries covering automated algorithm configuration:

- AClib: Algorithm Configuration Library: https://aclib.net/
- AClib: Algorithm Configuration Library 2.0: https://bitbucket.org/mlindauer/aclib2/src/master/
- **[EimerEtAl21]** Theresa Eimer, Andre Biedenkapp, Maximilian Reimer, Steven Adriaensen, Frank Hutter, Marius Lindauer, DACBench: A Benchmark Library for Dynamic Algorithm Configuration, IJCAI 2021. https://github.com/automl/DACBench

Automated algorithm selection:

- [BischlEtAl16] Bischl B, Kerschke P, Kotthoff L, Lindauer M, Malitsky Y, Fréchette A, Hoos H, Hutter F, Leyton-Brown K, Tierney K, Vanschoren J. Aslib: A benchmark library for algorithm selection. Artificial Intelligence. 2016. http://www.aslib.net

Automated hyperparameter optimization:

- **HPOlib: https://github.com/automl/HPOlib**
- **HPOBench: https://github.com/automl/HPOBench**
- Bayesmark: https://github.com/uber/bayesmark

Code smells are bad design patterns that come up in code. More precisely, code smells are the inverse of code quality. Recent work, a large benchmark has been provided by Madeyski et al, in which several experienced developers were asked to annotate certain code snippets with whether they experience it as a code smell. This benchmark has been made publicly available. Later, Soomlet et al. extended this dataset with all sorts of measurable metrics, so that machine learning can be used to mimic this developer perception. Both datasets are made publicly available online. Code smell detection could provide a further useful basis for interesting, new AutoAI benchmarks:

- [MadeyskiLewowski20] Madeyski, L., Lewowski, T.: Mlcq: Industry-relevant code smell data set. In: Proceedings of the Evaluation and Assessment in Software Engineering, pp. 342–347(2020)
- **[SoomlekEtAl21]** Soomlek, C., van Rijn, J.N., Bonsangue, M.M.: Automatic human-like detection of code smells, Discovery Science 2021.

Reinforcement learning (RL) is a general and widely used approach for addressing a wide variety of problems, due to its flexibility to describe and optimise problems. However, the training costs of these agents are often excessively high and robust general agents that can

generalise to multiple tasks are preferred. To this end, there have been two specific works to introduce a benchmark that evaluates RL agents with respect to an ever-changing world using context:

- **[BenjaminsEtAl21]** Benjamins, Carolin; Eimer, Theresa; Schubert, Frederik; Biedenkapp, André; Rosenhan, Bodo; Hutter, Frank; Lindauer, Marius CARL: A Benchmark for Contextual and Adaptive Reinforcement Learning, In: Workshop on Ecological Theory of Reinforcement Learning (EcoRL@NeurIPS'21), 2021.

- **[BenjaminsEtAl22]** Benjamins, Carolin; Eimer, Theresa; Schubert, Frederik; Mohan, Aditya; Biedenkapp, André; Rosenhan, Bodo; Hutter, Frank; Lindauer, Marius, Contextualize Me – The Case for Context in Reinforcement Learning, In: arXiv:2202.04500, 2022.

Another work towards Reinforcement Learning (RL) based benchmarks is MDPPlayground which provides dynamic generation of fast and extensible environments in which to test reinforcement learning agents.

- **[RajanEtAl21]** Rajan, Raghu; Diaz, Jessica Lizeth Borja; Guttikonda, Suresh; Ferreira, Fabio; Biedenkapp, André; von Hartz, Jan Ole; Hutter, Frank, MDP Playground: A Design and Debug Testbed for Reinforcement Learning, In: arXiv:1909.07750, 2021.

There has been a serious push towards establishing a set of benchmarks for Dynamic Algorithm Configuration (DAC), a very general framework in which we can adjust the parameters of an algorithm while it is operating to improve performance compared to keeping the parameters fixed during execution. To efficiently advance this topic forward, improvements were made over DACBench [EimEtAl21], leading to the *GECH track* **best paper award at GECCO'22!**

- **[BiedenKappEtAL22]** Biedenkapp, André; Dang, Nguyen; Krejca, Martin S.; Hutter, Frank; Doerr, Carola, "Theory-inspired Parameter Control Benchmarks for Dynamic Algorithm Configuration", In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'22), 2022

- **[EimerEtAl21]** Eimer, Theresa; Biedenkapp, André; Reimer, Maximilian; Adriaensen, Steven; Hutter, Frank; Lindauer, Marius, "DACBench: A Benchmark Library for Dynamic Algorithm Configuration", In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI'21), ijcai.org, 2021.

4. Self-monitoring AI systems [T7.3, ULEI]

Self-monitoring AI systems concern systems that are able to monitor the robustness of their performance. Here robustness is considered in a broad sense and can include natural drift, changes in the system, and adversarial attacks. For AI systems to be reliable they should be able to detect, report on, and ultimately automatically correct themselves. This task aims to produce:

1. General-purpose methods to self-calibrate AI systems, both for machine learning and other areas of AI

2. Metrics to assess those self-calibration approaches.
3. Benchmarks for these problems

Initial work exists to support the detection of incorrect predictions by machine learning systems. Indeed, where machine learning models in some cases can give an uncertainty quantification (e.g. a Gaussian process does this natively, and a random forest can assess the number of agreeing trees in the ensemble) there is often improvement to be made, by estimating the demographics of a certain observation. We explored this in the following work:

- **[KönigEtAl20]** Matthias König, Holger H Hoos and Jan N van Rijn. Towards Algorithm-Agnostic Uncertainty Estimation: Predicting Classification Error in an Automated Machine Learning Setting. In ICML Workshop on Automated Machine Learning. 2020.

Another direction considering predictions of machine learning systems is Neural Network Verification. This field does not necessarily research wrong predictions but aims to verify whether under certain conditions a Network retains its ability to provide correct predictions. Complete verification is expensive in terms of computing resources. This limits authors in benchmarking efforts. A yearly competition series (VNN comp), where developers can submit their verification tools, has alleviated part of this problem.

This workpackage has taken this one step further and has created a benchmark, that allowed for assessing the state of the art and the complementarity of different verification tools. This effort was awarded with a best paper award at the SafeAI@AAAI workshop and an extended version was recently published at JMLR. This extensive investigation can help practitioners.

- Github Repo: https://github.com/AWbosman/nn-verification-assessment
- **[KönigEtAl23]** König, M., Bosman, A. W., Hoos, H. H., & van Rijn, J. N. (2023). Critically Assessing the State of the Art in CPU-based Local Robustness Verification**.** Proceedings of the Workshop on Artificial Intelligence Safety 2023 (SafeAI 2023) co-located with the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI2023)**, best paper award.**
- **[KönigEtAl24]** König, M., Bosman, A. W., Hoos, H. H., & van Rijn, J. N. (2024). Critically assessing the state of the art in neural network verification. *Journal of Machine Learning Research*, *25*(12), 1-53.

For natural language inference (NLI) in a supervised setting, an adversarial benchmark exists; it works by testing machine learning systems and asking adversarial human annotators to break it:

- [NieEtAl19] Nie Y, Williams A, Dinan E, Bansal M, Weston J, Kiela D. Adversarial NLI: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599. 2019. https://arxiv.org/pdf/1910.14599.pdf

Other adversarial settings, such as neural network verification, are starting to adopt AutoML techniques, but specific benchmarks are still missing:

- **[KönigEtAl22]** Matthias König, Holger H Hoos and Jan N van Rijn. Speeding up neural network robustness verification via algorithm configuration and an optimised

mixed integer linear programming solver portfolio. In Machine Learning Journal (MLJ), 2022.

Even so, for neural network verification, there is work that could lead to such benchmarks. For instance, from the following competition:

- Verification of Neural Networks Competition: https://sites.google.com/view/vnn20/

In addition, while they do not define a benchmark, the following work provides rather extensive evaluations along with tooling to reproduce these:

- [TjengEtAl19] Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating Robustness of Neural Networks with Mixed Integer Programming. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), 2019.

The WILDS benchmarking environment considers distribution shifts for realistic situations between the training and test data. Datasets are included that cover both domain generalisation and subpopulation shifts, as well as the combination of the two.

- [KohEtAl21] Koh PW, Sagawa S, Xie SM, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, Lee T, David E. Wilds: A benchmark of in-the-wild distribution shifts. In International Conference on Machine Learning 2021. https://wilds.stanford.edu/

Similar to the task of moving beyond standard supervised learning, here, too, the extension and integration of benchmarks into the same existing platforms for automated algorithm configuration, automated algorithm selection, and automated hyperparameter optimisation should be considered.

The sat4j platform now supports Dynamic Algorithm Configuration to select some of its heuristics (Bumping strategies for Pseudo Boolean Solving right now). As such, it can now be used as a platform to experiment with Self-monitoring AI systems for reasoning engines.

- https://gitlab.ow2.org/sat4j/sat4j/-/tree/DAC

5. Multi-objective AutoAI [T7.4, INRIA]

Most works in AutoAI are searching for the best algorithm, algorithm configuration, or full pipeline design and configuration that optimises the performance of the resulting algorithm/pipeline, i.e., are driven by a single objective. However, and this is especially true in the context of TAILOR, though performance will always be of interest, there are other objectives pertaining to trustworthiness that might also be important (e.g., explainability, frugality, robustness, etc.). There is unfortunately little doubt that improving these trustworthiness properties will result in decreasing performance: some trade-off needs to be adopted.

Finding trade-offs between antagonist objectives is the goal of multi-objective optimization (MOO), also known as multi-criteria optimisation. There are several approaches for MOO, from Evolutionary Algorithms (e.g. [DebEtAl02, CoeLec02]) to Iterated Local Search (e.g.

[BloEtAl15, DubEtAl15, DerEtAl16]) to Bayesian Optimisation (e.g. [GalEtAl20, EmmEtAl16]), that aim at finding the Pareto set, i.e. the set of best trade-offs for the competing objectives, points such that no objective can be improved without degrading another objective.

A first important remark is that there exist several works aiming at algorithm selection, algorithm configuration, and even algorithm design of multi-objective algorithms (MOAs) [BloEtAl17]. However, this work does not truly consider multi-objective AutoAI: these works look for the design/selection/configuration of MOAs **from a single-objective point of view**, optimizing the performance of the resulting algorithm, usually in terms of hypervolume, or some other multi-objective indicator, and do not consider multiple objectives for AutoAI. This is the case with most, if not all, continuous optimisation platforms proposing datasets of problems.

COCO: A platform their state-of-the-art solutions, from COCO to Nevergrad, and has been present in the corresponding competitions (BBOB, BBComp). For comparing continuous optimizers in a black-box setting. Optimisation Methods and Software. 2021

- BBOB: http://numbbo.github.io/workshops/index.html
- [BennetEtAl21] Bennet P, Doerr C, Moreau A, Rapin J, Teytaud F, Teytaud O. Nevergrad: black-box optimisation platform. ACM SIGEVOlution. 2021: https://github.com/facebookresearch/nevergrad
- BBcomp - Black-box Optimisation Competition: https://www.ini.rub.de/PEOPLE/glasmtbl/projects/bbcomp/

Similar work could be done in other domains where benchmarks exist for AS/AC of single-objective combinatorial functions (ASLib, ACLib, Souza). But we will not consider these any more in this section, for the reasons above.

- AClib: Algorithm Configuration Library: https://aclib.net/
- AClib: Algorithm Configuration Library 2.0: https://bitbucket.org/mlindauer/aclib2/src/master/
- [BischlEtAl16] Bischl B, Kerschke P, Kotthoff L, Lindauer M, Malitsky Y, Fréchette A, Hoos H, Hutter F, Leyton-Brown K, Tierney K, Vanschoren J. Aslib: A benchmark library for algorithm selection. Artificial Intelligence. 2016. http://www.aslib.net
- **[SouzaEtAl]** Marcelo de Souza, Marcus Ritt and Manuel López-Ibáñez. Capping Methods for the Automatic Configuration of Optimization Algorithms. Computers & Operations Research 139: 105615. **https://doi.org/10.1016/j.cor.2021.105615 https://github.com/souzamarcelo/supp-cor-capopt**

On the other hand, it seems straightforward to extend the simplest existing (single-objective) AutoAI algorithms that do some direct meta-optimisation of algorithms hyperparameters, to handle multiple objectives, replacing the optimisation algorithm at work in the AutoAI at hand with the corresponding multi-objective optimiser. Indeed, this has been proposed for instance for ParamILS [HutEtAl09], with the MO-ParamILS platform, where the MO-ParamILS

algorithm was applied to find the Pareto-set for several bi-objective algorithm configuration problems, balancing performance and complexity/CPU cost.

- [BlotEtAl16] Blot A, Hoos HH, Jourdan L, Kessaci-Marmion MÉ, Trautmann H. MO-ParamILS: A multi-objective automatic algorithm configuration framework. In International Conference on Learning and Intelligent Optimisation 2016 May 29 (pp. 32-47). Springer, Cham.

Another interesting approach is NSGA-Net, in which NSGA-II is used for Neural Architecture Search (NAS), using a specific representation of DNN architectures, to discover the best trade-off between accuracy (on popular datasets like MNIST and Cifar-10) and size of the network (as a proxy for complexity of the learning phase).

- [LuEtAl19] Lu Z, Whalen I, Boddeti V, Dhebar Y, Deb K, Goodman E, Banzhaf W. Nsga-net: neural architecture search using multi-objective genetic algorithm. InProceedings of the Genetic and Evolutionary Computation Conference 2019 Jul 13 (pp. 419-427).

Two works done in the context of TAILOR use NSGA-II to tune hyperparameters of the Federated Learning approach where the objectives are to maximize the accuracy of the global AI model while at the same time the communication overhead is reduced.

- **[MartinezEtAl22]** José Ángel Morell Martínez, Zakaria Abdelmoiz Dahi, Francisco Chicano, Gabriel Luque, Enrique Alba: Optimising Communication Overhead in Federated Learning Using NSGA-II. EvoApplications 2022: 317-333

- **[MorellEtAl24]** José Á. Morell, Zakaria Abdelmoiz Dahi, Francisco Chicano, Gabriel Luque, Enrique Alba: A multi-objective approach for communication reduction in federated learning under devices heterogeneity constraints. Future Gener. Comput. Syst. 155: 367-383 (2024)

However, these works have remained isolated and, in particular, no recognised benchmarks in the area of Multi-objective AutoAI have been proposed yet, nor do existing works even compare their results to those of other Multi-objective AutoAI approaches. There are several reasons for that.

- Comparing the results of two MOAs (i.e., comparing two Pareto fronts) is not straightforward. Several measures have been proposed, and hence several rankings among Multi-objective AutoAI algorithms will need to be established

- There is no clear way to measure most of the trustworthiness objectives: whereas existing works use as second objective some measure of complexity (or CPU cost of learning or optimisation), and this could, with some twist, be considered as a proxy for explainability (though a DNN with 300,000 weights can be considered more explainable than another one with 3M weights, it is in fact not reasonable to talk about explainability in such context).

- Many (single-objective) AutoAI algorithms are not simply made of one meta-optimisation algorithm, and hence cannot easily be turned into multi-objective AutoAI. Approaches exist that are based on recommendation processes [MısSeb17], or on multi-armed bandits [RakEtAl19] (even though multi-objective MAB algorithms

exist); or algorithms that optimise the whole pipeline, like Auto-sklearn [FeuEtAl19], TPOT [OlsEtAl16], MOSAIC [RakEtAl19], or EvoFlow [BarEtAl24].

Nevertheless, some low-hanging fruit exists, and future work should address the following avenues:

- Continue with the performance/complexity bi-objective AutoAI: in that area at least, some benchmarks could be set up, building on the existing datasets and platforms.
    - A particular issue in the framework of algorithm selection for continuous optimisation is that of the performance measure. Two points of view exist: measure the time (or number of calls to the objective function) needed to reach a given performance, as done in COCO/BBOB, allowing easy aggregation of performances for functions with different orders of magnitude of values; or measure the best objective value reached in a given time, as done in many papers, as well as in Nevergrad and in all BBComp competitions. Running multi-objective algorithm selection/configuration with the best value/number of iterations consumed could reconcile both approaches.
    - In the combinatorial optimisation domain, there exist many specialised libraries, like tsplib for the Travelling Salesman Problem, or the past instances of competitions for problems like SAT or Planning, and all of them could easily be turned into benchmarks for multi-objective algorithm selection and configuration:
        - TSPLIB: http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/
        - SAT competition: http://www.satcompetition.org/
        - Planning competitions: https://www.icaps-conference.org/competitions/
    - In AutoML (for supervised learning), many works deal with optimizing the whole learning pipeline, and a popular benchmark has emerged from the OpenML dataset, the OpenML100. It has since been by OpenML-CC18, which could easily be transformed into a bi-objective AutoAI problem by also minimising the learning time, provided all runtimes are scaled fairly (e.g., from the training time of some simple dummy algorithm).
        - [BischlEtAl19] Bischl, B. and Casalicchio, G. and Feurer, M. and Hutter, F. and Lang, M. and Mantovani, R. G. and van Rijn, J. N. and Vanschoren, J. Openml benchmarking suites, arXiv preprint arXiv:1708.03731. 2019.
    - NAS (aka AutoDL) also offers a nice benchmark with the NASBench initiatives, offering either tabular datasets:
        - **[YingEtAl19]** Chris Ying, Aaron Klein, Esteban Real, Eric Christiansen, Kevin Murphy, Frank Hutter, "Nas-bench-101: Towards

reproducible neural architecture search," Proceedings of the International Conference on Machine Learning, 2019.

or surrogate datasets of pre-computed performances of many DNN architectures:

- **[SiemsEtAl20]** Julien Siems, Lucas Zimmer, Arber Zela, Jovita Lukasik, Margret Keuper, Frank Hutter. “NAS-Bench-301 and the Case for Surrogate Benchmarks for Neural Architecture Search”, arXiv 2020

- Propose indicators beyond complexity, for other trustworthiness objectives, and extend all the above to other multi-objective settings:
  - Robustness seems a first easy and agnostic target. In optimisation, one can simply average the performances over some sampling around the point of interest - though many options remain open regarding such sampling. In Machine Learning, some noise can be added to the training samples, or to the test sample, or both.

    In the specific domain of DNNs, random noise will not demonstrate anything, and robustness against adversarial examples is another challenge that still requires some research - and it seems we are still far from benchmarks.
  - Explainability on the other hand will probably require defining some proxies: complexity is one, though not satisfying on its own, and one difficulty will be to come up with some agnostic indicator (even complexity is model-dependent). Interaction with users might be mandatory, as explainability depends on the expertise of the target human.
  - Fairness in the context of supervised learning, requires probably even more work, as some biased datasets need to be designed (but many models of biases can be used, from unbalanced training sets to purposely biased outcomes), and fairness of the resulting models measured - all fields still subject to active research.
- Using the above indicators, extend existing AutoAI algorithms that operate on complex search spaces using Evolutionary Computation to multi-objective settings, and discover new learning paradigms following the path opened by AutoM-Zero [ReaEtAl20].

Presented at the first AutoML conference in July 2022 (though available on Arxiv since Sept. 2021), YAHPO is a new benchmark based on surrogate models (and hence in which function evaluations are very fast to compute) that allows for multi-objective hyperparameter optimisation.

- **YAHPO Gym – An Efficient Multi-Objective Multi-Fidelity Benchmark for Hyperparameter Optimization** Florian Pfisterer, Lennart Schneider, Julia Moosbauer, Martin Binder, Bernd Bischl OpenReview AutoML 2022

6. Ever-learning AutoAI [T7.5, TU/e]

“Ever-learning AutoAI” aims to ensure that AutoAI gets better over time, producing better models with less data, and avoids the computational overhead of starting from scratch for any new use case, or change in scenario.

The science of learning how to learn better or faster through experience is called meta-learning (or learning to learn). This can be done in several ways:

- Keeping the model architecture and design decisions (hyperparameters) fixed, and then learning good initial model parameters and/or how to update them for a certain set of tasks. This is useful when many similar tasks exist, and the model itself is large and flexible (e.g., a large neural network), so that re-tuning the weights, without changing the architecture, is sufficient to adapt to new tasks.

- Meta-learning the model architecture itself. This is the combination of meta-learning and AutoAI, which is more generally applicable since it can also work when new tasks are quite different from previous tasks. This also works with smaller models as they can completely adapt to new tasks. Ideally, it also leads to models which are smaller and more tuned to the given task. We will focus mainly on this approach in this WP.

- Doing any of the above, while also choosing or generating the next task to solve. The idea here is to learn simple variations of a task first and use that experience to learn increasingly hard variations of the task.

A notable related field is continual learning, where there is a single task, but it itself evolves over time (see, e.g., [DelEtAl21]). New data points may come in that are slightly different from those before (concept drift), which may require retraining or even a change in architecture. For instance, if outliers suddenly appear, the learning pipeline may need to be adapted to deal with these outliers.

While this is an active and fast-evolving field, it is also quite young, and there exist almost no commonly used benchmarks.

For one particular subfield, few-shot learning, there exists the meta-dataset

- [TriantafillouEtAll19] Triantafillou E, Zhu T, Dumoulin V, Lamblin P, Evci U, Xu K, Goroshin R, Gelada C, Swersky K, Manzagol PA, Larochelle H. Meta-dataset: A dataset of datasets for learning to learn from few examples. arXiv preprint arXiv:1903.03096. 2019. https://github.com/google-research/meta-dataset

This is a curated collection of 10 object recognition (vision) tasks. While it is well designed and has some adoption, few researchers test their algorithms on all 10 tasks: since these datasets are large and each requires pretraining large models, this is often prohibitively expensive. Often, only a few tasks are selected (e.g., omniglot and/or mini-imagenet). This limits its impact as a benchmark. Some examples of its use by TAILOR partners are:

- Elsken et al. 2017, Meta-NAS. By ALU-FR. Uses meta-learning to speed up NAS and evaluates 2 tasks also appearing in the meta-dataset.

- Gonzalez and Vanschoren, 2019, Meta-Reinforcement learning for NAS. By TUE. Trains an RL agent to do NAS and evaluates on 4 tasks from the meta-dataset.

For the combination of meta-learning and AutoAI, OpenML is often used as a source of meta-data to either learn which hyperparameters are most important to tune [VanHut18], to warm-start the search for good model configurations based on what worked on similar tasks [**FeuEtAl19**], or to train meta-models to predict which configurations will work well [BraEtAl21]. Especially the OpenML100 or OpenML-CC18 mentioned earlier are used often, although some authors also use datasets other than those from OpenML.

To improve the state of benchmarking in this field, our efforts in TAILOR concentrate on three research lines:

- Set up central infrastructure, building on OpenML, to collect large amounts of meta-data. This will represent a 'global memory' of AI approaches and their performance that will speed up and imbue new rigour to AI research.

- Create meta-datasets, curated sets of many related tasks, to stimulate research into techniques that can learn effectively across tasks, and benchmark them to measure progress in meta-learning.

- Nurture research in meta-learning and transfer learning to leverage this meta-data to build better models, as well as research that combines these techniques with AutoAI approaches to yield ever-learning AutoAI techniques.

We discuss progress on these lines below.

6.1. AutoML Benchmark

TUE continued their work on the AutoML Benchmark, an open-source benchmarking tool for AutoML frameworks. Since the last update, a large-scale evaluation of 9 different AutoML frameworks on a mix of over 100 classification and regression tasks has been completed. Further, they wrote a paper about the design of the benchmark and the experimental results, which is now published in JMLR:

- **[GijsbersEtAl24]** Gijsbers P., Bueno M.L.P, Coors S., LeDell E., Poirier S., Thomas J., Bischl B., Vanschoren, J. AMLB: an AutoML Benchmark. Journal of Machine Learning Research (JMLR), 25 (101), pp. 1-65.

- Project website: **https://openml.github.io/automlbenchmark/**

The project website contains several additional tools for the community to analyse the obtained results, including an interactive app and multiple Python notebooks.

This benchmark has already known widespread adoption:

- Over 15 AutoML frameworks were included by their original authors, many of which are from industry. This includes AutoGluon (Amazon), AutoSKLearn (U Freiburg),

GAMA (U Eindhoven), H2O-AutoML (H2O), ML.NET AutoML (Microsoft), Auto-XGBoost and MLR3AutoML (U Munich), FLAML (Microsoft), MLJAR-AutoML (MLJAR), OBOE (Cornell), LightAutoML (Sberbank AI), hyperopt-sklean (U Waterloo), and MLPlan (U Paderborn).

- The AutoML benchmark is now a key part of the development and testing pipeline on many top AutoML systems, especially in AutoGluon.
- A first run was completed in 2019, with results published at the ICML AutoML Workshop. A second, much larger, run has evaluated a much wider range of AutoML systems on more than 100 classification and regression datasets and has been published in the JMLR journal.
- To ensure fair evaluation, all systems are run with equal resources as docker images on identical hardware on AWS.
- It is dynamic: the collection of tasks used changes over time, to avoid people (or AutoML systems) overfitting a certain set of datasets.

We are currently extending this benchmark to cover a much wider range of datasets, especially datasets which are 'hard' for state-of-the-art AutoML systems, to encourage further development, and aim to release an update later this year. Moreover, we also aim to extend the benchmark to explicitly allow meta-learning under certain controlled conditions, to encourage the wider adoption of Meta-Learning in these AutoML frameworks.

Finally, via a program funded by the Dutch Science Foundation (NWO), we will use the infrastructure created in the AutoML benchmark to create a new open science service in which researchers can submit their datasets, and we will run multiple AutoML systems on them (depending on their requirements), to automatically generate machine learning models. From this, we aim to learn how useful AutoML is for accelerating scientific research, and how we can improve it further to enable robust machine learning-driven research.

6.2. OpenML upgrades and Croissant

We have also upgraded OpenML in several ways to improve benchmarking on a wider range of tasks. OpenML already has extensive support for creating new benchmarks and making them easy to use, as demonstrated in a paper accepted in the NeurIPS Datasets and Benchmarks track, acknowledging the support from TAILOR:

- **[BischlEtAl21]** Bischl B, Casalicchio G., Feurer M., Gijsbers P., Hutter F., Lang M., Gomes Mantovani R., van Rijn, J,N., Vanschoren. J. OpenML Benchmarking suites. Advances in Neural Processing Systems, Datasets and Benchmarks (NeurIPS 2021)

We have been working on extending the range of datasets it could efficiently serve:

- The OpenML API has been updated to allow binary data formats (in particular Parquet). This allows the inclusion of more types of datasets into OpenML, including large image and text datasets. We successfully ran and stored experiments on datasets of 8GB and larger.

- The OpenML Python library [FeuEtAl21] has been updated to transparently handle binary formats. End users can submit and receive datasets natively from many Python data structures (e.g. Pandas, Dask, Numpy,...), facilitating experimentation (e.g. evaluations of new pipelines). They can also easily run experiments and stream results back to the platform.
- We have close to 250k yearly users and close to 1k daily visitors to the website, in addition to high-frequency bursts of calls to our APIs. Our servers are regularly experiencing heavy loads. We set up a Kubernetes environment and S3 storage to ensure scalability, high availability, and conformance to modern standards.

In future work, we aim to leverage OpenML to create new benchmarks that can serve as a reference for work leveraging meta-learning in AutoAI. For this, we also aim to offer new services that compute and return rich meta-data that can be directly used by any AutoML system. This will drastically reduce the cost of leveraging meta-learning, lower the threshold to use meta-learning more intensively in AutoML systems, while at the same time also making it more systematic, reproducible, and comparable.

More recently, TUE has partnered with other leading machine learning platforms, and have created a meta-data format, **Croissant**, specifically for machine learning datasets while being compatible with existing standards for describing datasets in general. This new data format allows datasets to be more easily exchanged and more easily imported into many existing machine learning libraries. It enables us to extend OpenML to support a wider range of datasets, import them from existing platforms, and use them to create new benchmarks. Currently, this format is supported by OpenML, HuggingFace, Kaggle, Google Dataset Search, and TensorFlow Datasets. Is it also being adopted by Harvard Dataverse, and the NeurIPS conference is recommending it to be used for all new datasets submitted to the conference. This work also won the **best paper award** at the DEEM Workshop at SIGKDD 2024:

- **[AkhtarEtAl24]** Akhtar M., Benjelloun O., Conforti C., Gijsbers P., Giner-Miguelez J., Jain N., Kuchnik M., Lhoest Q., Marcenac P., Maskey M., Mattson P., Oala L., Ruyssen P., Shinde R., Simperl E., Thomas G., Tykhonov S., Vanschoren J., van der Velde J., Vogler S., Wu C.J. Croissant: A Metadata Format for ML-Ready Datasets. SIGKDD Workshop on Data Management for End-to-End Machine Learning, 2024.

In future work, we aim to extend this format to also support the reporting and exchange of benchmarking results, which would also have a tremendous impact on the way we perform machine learning research.

6.3. Meta-learning challenges

We co-organised two meta-learning challenges, one at AAAI 2021 and one at NeurIPS 2021. In these challenges, AutoAI systems for NAS are allowed run-up time for Meta-Learning: instead of starting the search from scratch, they are allowed to spend some time to collect metadata about the challenge tasks and use that to search for the best architectures more efficiently. While these challenges only offer resources for a certain

amount of time (until the challenge is over), they can serve as inspiration for future benchmarks in this area.

The most recent work in this area can be found in the NeurIPS workshop on meta-learning (https://meta-learn.github.io/2020/) and the ICML 2021 workshop on AutoML (https://sites.google.com/view/automl2021).

Works published since September 2020 (highlighted if part of TAILOR):

1. **[BrazdilEtAl21]** Brazdil, P., van Rijn, J.N., Soares, C., Vanschoren, J. Metalearning. Applications to Automated Machine Learning and Data Mining. Springer 2021
2. **[CelikEtAl21]** Celik, Bilge, and Joaquin Vanschoren. "Adaptation strategies for automated machine learning on evolving data." IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).
3. **[ElskenEtAl]** Elsken, T., Staffler, B., Metzen, J. H., & Hutter, F. (2020). Meta-learning of neural architectures for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12365-12375).
4. **[FeurerEtAl21]** Feurer, Matthias, et al. "Openml-python: an extensible python api for openml." Journal of Machine Learning Research 22.100 (2021): 1-5.
5. **[GijsbersEtAl21]** Gijsbers, Pieter, et al. "Meta-Learning for Symbolic Hyperparameter Defaults." GECCO 2021 (2021).
6. **[GijsbersEtAl20]** Gijsbers, Pieter, and Joaquin Vanschoren. "GAMA: a General Automated Machine learning Assistant." ECMLPKDD 2020 (2020).
7. **[WeertsEtAl20]** Weerts, Hilde JP, Andreas C. Mueller, and Joaquin Vanschoren. "Importance of tuning hyperparameters of machine learning algorithms." arXiv preprint arXiv:2007.07588 (2020).

There are plans to extend this benchmark to allow for (explicit) multi-objective optimisation to support the trade-off between obtaining high model accuracy and other aspects, such as fairness or inference time. An additional set of experiments will also be carried out later this year, where this trade-off for AutoML frameworks will be investigated.

In collaboration with Laurens Bliek (Eindhoven University of Technology) the possibility of curating AutoML benchmark tasks centered around climate change data will be explored.

6.4. MetaDL and Meta-Album

Meta-Learning, whose goal is to learn across datasets, can be seen as a form of AutoAI. Challenges (open competitions during which competitors submit their best algorithms for blind comparisons with the other submissions) usually remain open after the competition has ended, and hence de facto become benchmarks. A survey of benchmarks in AutoAI must include the Cross Domain MetaDL challenge, a TAILOR challenge within Task 2.4 that is currently running, and will become a benchmark after Oct. 1, the date of the publication of the results. More details are given in Deliverable 2.3.

- **[BazEtAl21]** El Baz, A., Guyon, I., Liu, Z., Treguer, S., van Rijn, J.N. and Vanschoren, J.: Advances in MetaDL: AAAI 2021 challenge and workshop. AAAI Workshop on Meta-Learning and MetaDL Challenge, 1-16, 2021.

- **NeurIPS'22 Cross-Domain MetaDL competition: [CarrionEtAl22]** Design and baseline results. Dustin Carrión-Ojeda, Hong Chen, Adrian El Baz, Sergio Escalera, Chaoyu Guan, Isabelle Guyon, Ihsan Ullah, Xin Wang, and Wenwu Zhu. https://drive.google.com/file/d/145t-KVmHNIFCweiIjbPwimmAXMvHHf7e/view

The underlying dataset used in this challenge is Meta-Album, an extensible multidomain meta-dataset, including (so far) 40 image classification datasets from 10 different domains. Meta-Album was specifically designed to facilitate meta-learning research in the cross-domain few-shot setting, which is more realistic than commonly used evaluation protocols. All datasets and Open Source code is available at https://meta-album.github.io/. A paper giving all details about how this dataset of datasets has been built is available on OpenReview. Further details about Meta-Album are presented in Deliverable 2.3 (foundational benchmarks and challenges).

- **Meta-Album: [UllahEtAl22]** Multi-domain Meta-Dataset for Few-Shot Image Classification. Ihsan Ullah, Dustin Carrion, Sergio Escalera, Isabelle M Guyon, Mike Huisman, Felix Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, Phan Anh Vu. *NeurIPS 2022 Datasets and Benchmarks Track.*

6.5. Meta-learning for unsupervised AutoML

TUE proposed a way to do unsupervised AutoML by leveraging meta-learning. AutoML usually needs a ground truth (a golden standard) to get a signal of whether one model is better than another. This is usually not available in unsupervised problems (if it is, then all standard AutoML techniques could be used). In the absence of this feedback, it is still possible to recommend unsupervised techniques based on prior knowledge. By recording the performance of unsupervised techniques on many prior problems, we can recommend techniques based on how similar a new problem is to prior ones. This similarity can be accurately measured using 'optimal transport', captured by the Wasserstein distance, which measures the (dis-)similarity between two data distributions. Indeed, if two data distributions are very similar, then the same unsupervised techniques will likely work well, which we demonstrated empirically on large benchmarks of unsupervised tasks:

- **[SinghVanschoren23a] Singh P, and Vanschoren J. AutoML for outlier detection with optimal transport distances. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23), 2023.**

- **[SinghVanschoren23b] Singh P, and Vanschoren J. Applications of Optimal Transport Distances in Unsupervised AutoML NeurIPS 2023 Workshop Optimal Transport and Machine Learning. 2023**

6.6. Meta-learning for Bayesian Optimisation

Bayesian optimisation (BO) is a very efficient method to build AutoML systems, and meta-learning is an effective way to leverage knowledge from related tasks to optimize new tasks faster. However, existing meta-learning methods for BO rely on surrogate models that are not scalable or are sensitive to varying input scales and noise types across tasks. Moreover, they often overlook the uncertainty associated with task similarity, leading to unreliable task adaptation when a new task differs significantly or has not been sufficiently explored yet. TUE, in collaboration with Bosch, proposed a novel meta-learning BO approach that bypasses the surrogate model and directly learns the utility of queries across tasks via a deep neural network with a meta-learning module. It explicitly models task uncertainty and includes an auxiliary model to enable robust adaptation to new tasks. Extensive experiments show that this method achieves strong performance and outperforms multiple meta-learning BO methods across various benchmarks. This work has been accepted at ICML 2024:

- **[PanEtAl24]** Pan J, Falker S., Berkenkamp F., Vanschoren J, MALIBO: Meta-learning for Likelihood-free Bayesian Optimization. Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

6.7. Meta-learning optimisers for continual learning

Continual learning (CL) refers to the ability to continually learn over time by accommodating new knowledge while retaining previously learned experience. While this concept is inherent in human learning, current machine learning methods are highly prone to overwrite previously learned patterns and thus forget past experiences. Instead, model parameters should be updated selectively and carefully, avoiding unnecessary forgetting while optimally leveraging previously learned patterns to accelerate future learning.

Through a Connectivity Fund exchange, researchers from TUE and Halmstad University proposed to use meta-learning to train a transformer-based optimiser to enhance CL. This meta-learned optimiser uses attention to learn the complex relationships between model parameters across a stream of tasks and is designed to generate effective weight updates for the current task while preventing catastrophic forgetting on previously encountered tasks.

In short, it learns which weights in the network should be updated when given a new task, and leaves other weights untouched, to learn new tasks effectively while not forgetting previous tasks. This work was accepted in CoLLAs 2024:

- **[VettoruzzoEtAl24]** Vettoruzzo A., Vanschoren J., Bouguelia M-R., Rognvaldsson, T. Learning to learn without forgetting using attention. Third Conference on Lifelong Learning Agents (CoLLAs), 2024.

7. Hardware Dimensioning of AI algorithms

The problem of determining the right hardware (HW) architecture and its configuration to run an AI algorithm under required performance and budget limits is called hardware dimensioning. This is a challenging and complex task for many companies that need

assistance from AI experts in handling it. Algorithm developers need to answer questions such as: (1) if an AI algorithm has to run under real-time constraints, respecting some budget constraints and guaranteeing a certain solution quality, which HW architecture should be used? (2) given a certain set of (possibly heterogeneous) HW resources, what is the best algorithm to solve a problem respecting user-defined constraints? These are not trivial questions and require domain and AI expert knowledge to be solved, owing to the complexity of knowing beforehand the behaviour of an algorithm on different HW architectures and evaluating the effect of all possible choices (HW and configurations). Moreover, even AI experts might not know exactly how to dimension the HW resources, typically resorting to over-provisioning (requesting more resources than those strictly needed) and/or using heuristics, with the risk of sub-optimal solutions.

To address this issue, HADA **[DeFEtAl22]** has been proposed, an automated approach for HW dimensioning where ML models are embedded within an optimisation problem to enable decision-making over complex real-world systems. To create such ML models, empirical data is needed to learn the relationship between algorithm and hardware configuration and performance. To this end, we collected a dataset which we then made publicly available. The dataset consists of benchmarks of online/offline optimisation algorithms applied to the energy system domain, executed on a variety of heterogeneous resources. The collected data (e.g., performance metrics) enable the construction of ML models to estimate the behaviour of an AI application on different hardware architecture, thus enabling the creation of the matchmaking engine (hardware dimensioning). The data set might be of interest to other AI experts for the creation of ML and optimisation models for similar tasks. The data set contains both numerical and categorical datal. Data is mostly quantitative. The data is raw, with minimal pre-processing. The data is stored using a standard format, such as CSV files and/or a binarised format typically used in Python environments, namely pickle.

The data set is available on Zenodo: https://zenodo.org/record/5838437/

Further details can be found in the accompanying paper **[DeFEtAl22]**

- **[DeFEtAl22]** De Filippo A, Borghesi A, Boscarino A, Milano M. HADA: An automated tool for hardware dimensioning of AI applications. Knowledge-Based Systems. 2022 Sep 5;251:109199.

8. Machine Learning and Language Processing

UPV introduced Europarl-ASR, a large speech and text corpus of parliamentary debates including 1300 hours of transcribed speeches and 70 million tokens of text in English extracted from European Parliament sessions. The training set is labelled with the Parliament's non-fully verbatim official transcripts, time-aligned. As verbatimness is critical for acoustic model training, we also provide automatically noise-filtered and automatically verbatimised transcripts of all speeches based on speech data filtering and verbatimisation techniques. Additionally, 18 hours of transcribed speeches were manually verbatimised to build reliable speaker-dependent and speaker-independent development/test sets for streaming ASR benchmarking. The availability of manual non-verbatim and verbatim transcripts for dev/test speeches makes this corpus useful for the assessment of automatic filtering and verbatimisation techniques. A recent publication (see below) describes the

corpus and its creation and provides off-line and streaming ASR baselines for both the speaker-dependent and speaker-independent tasks using the three training transcription sets. The corpus is publicly released under an open licence.

- **[GarcésEtAl21]** G. Garcés, J. A. Silvestre, J. Jorge, A. Giménez, J. Iranzo, P. Baquero, N. Roselló, A. Pérez, J. Civera, A. Sanchis, A. Juan. Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization. In Proc. Interspeech 2021, pp. 3695–3699, Brno (Czech Republic), 2021. **doi:10.21437/Interspeech.2021-1905**
- https://www.mllp.upv.es/wp-content/uploads/2021/09/europarl-asr-presentation-ext[...]
- https://www.youtube.com/watch?v=Tc0gNSDdnQg&list=PLlePn-Yanvnc_LRhgmmaNmH12B[...]

9. Gap analysis

Benchmarks are essential for assessing and advancing the performance of models and systems. However, current benchmarks often fail to capture the complexity and diversity of real-world challenges. Looking ahead, it is clear that future benchmarks should evolve to address tasks that include multiple, potentially conflicting objectives, such as fairness or the balancing of interests among various stakeholders. For instance, differing values between citizen groups, organisations, and environmental concerns like fossil fuel use are notoriously difficult for humans to navigate.

Though some initial efforts, such as YAPHO gym, have made progress in this direction, these attempts are still in their infancy. The need for further development is significant. It would be particularly interesting to explore how advanced AI techniques can aid in tackling these multifaceted benchmarks. A promising starting point is the widely used neural architecture search (NAS) benchmarks, such as the NAS-bench series. Expanding these benchmarks to consider multi-objective tasks could offer a richer and more comprehensive evaluation framework.

Another set of interesting benchmarks concern more complex neural network verification tasks. These could help push the community to develop approaches and systems beyond local robustness verification and focus on diverse properties neural networks should adhere to, depending on the task at hand. Verification can ensure that large and complex black-box machine learning models trained with non-deterministic techniques and subject to unforeseen perturbations of their inputs behave as desired even under worst-case assumptions. This will make such models increasingly safe and robust to use in practice.

In the realm of datastreams, the challenge of data drift is another issue that current benchmarks do not fully address. Much research has focused on static datasets, even though real-world data is often dynamic and continually updated. Weather data, customer demographics, and trends are all in flux, and these changes can affect the accuracy and reliability of AI systems. Effective data drift detection is crucial to flagging when an AI system's performance begins to degrade, ensuring the system remains trustworthy over time. Although some companies are developing their own solutions, there is a shortage of

accessible real-world datasets for broader research. This lack of data hinders the development of more effective data drift detection methods, which are essential for practical, self-monitoring AI systems.

Finally, many existing benchmarks focus on single modalities, such as image processing, even though real-world AI applications increasingly rely on multi-modal data. Systems that integrate text, images, video, and sensor data require more comprehensive benchmarks to evaluate their performance effectively. Expanding current benchmarks or developing entirely new ones to assess multi-modal systems will be crucial for fostering innovation and ensuring that AI systems are adequately evaluated in diverse, real-world scenarios.

In summary, the future of AI benchmarking lies in addressing more complex, multi-objective tasks, ensuring safety through comprehensive verification, managing the challenges of data drift in dynamic environments, and developing benchmarks for multi-modal systems. By pushing the boundaries of current benchmarking practices, the AI community can ensure that future systems are not only more powerful but also more robust, fair, and aligned with real-world needs.

10. Possible drawbacks of AutoAI

As the goal of creating AutoAI methods and systems is to make AI techniques more accessible even for non-experts in AI, it is crucial that the AI systems or components produced by AutoAI approaches are of high quality. If an end-user is not capable of checking and validating, e.g., the machine learning pipeline produced by an AutoML system, it is necessary that the system takes this into account, i.e., it should be self-monitoring and have automated evaluation mechanisms.

Getting to a point where a system can be fully used independently and where it does not need interfering from a human is tedious. Ensuring that a system is fully capable and what are the limitations is a hard task, this is also described in the following work:

- Karmaker, S. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2021). Automl to date and beyond: Challenges and opportunities. *ACM Computing Surveys (CSUR)*, *54*(8), 1-36.

Integrating all possible methods for a specific task is a tedious and complicated task for the designer of AutoAI methods. It is important that AutoAI systems are carefully designed and documented, such that designers of specific algorithms can easily incorporate their work into existing AutoAI systems. At the same time, this means that the AutoAI systems should be well-maintained, which can be difficult to achieve for any software developed in academic environments.

Even though AutoAI methods open up the possibility of effectively creating AI solutions for different tasks, this often comes at high computational cost and substantial energy consumption. Often, running the AutoAI methods require large-scale computational infrastructure with specialised hardware, such as GPUs, which can be difficult to afford for some stakeholders. This is one of the drawbacks described in this best-practices survey:

- Serban, A., van der Blom, K., Hoos, H., & Visser, J. (2024). Software engineering practices for machine learning—Adoption, effects, and team assessment. *Journal of Systems and Software*, *209*, 111907.

They also highlight that developing AutoML methods, specifically, require more in-depth expertise. This slows down the adoption of the promising techniques.

It has been shown that AutoAI techniques can achieve higher performance than human experts; however, they are sensitive to changes in the data. This means that it should be extremely clear how AutoAI techniques have been calibrated for their task (training/selection/configuration), such that an informed decision can be made whether to use an AutoAI technique in a given context (e.g., for training an ML model based on the dataset at hand). This is especially important as AutoAI methods can be very inflexible as highlighted in this survey:

- Azevedo, K., Quaranta, L., Calefato, F., & Kalinowski, M. (2024). A Multivocal Literature Review on the Benefits and Limitations of Automated Machine Learning Tools. *arXiv preprint arXiv:2401.11366*.

A counterintuitive result of a recent study showed that employing Neural Architecture Search for finding and training a suitable Neural Network can lead to AI models that are more vulnerable to adversarial attacks:

- Pang, R., Xi, Z., Ji, S., Luo, X., & Wang, T. (2022). On the security risks of {AutoML}. In *31st USENIX Security Symposium (USENIX Security 22)* (pp. 3953-3970).

This preliminary study shows the importance of the creation of benchmarks that are multi-objective and AutoAI methods that make the need for fairness and robustness is inherent to the selection and training of models.

Furthermore, the use of AutoAI will accelerate change in the job market. In some cases, jobs will be automated entirely, in others, they may change significantly, for example through heavy reliance on AI tools. This may well give rise to the urgent need to create additional opportunities for displaced workers to learn and carry out new jobs.

Finally, although various initiatives exist to combine AutoML and algorithmic fairness, there are both opportunities and challenges which make this non-trivial without having humans in the loop. We published a paper in the Journal of AI Research that highlights this:

H. Weerts, F. Pfisterer, M. Feurer, K. Eggensperger, E. Bergman, N. Awad, J. Vanschoren, M. Pechenizkiy, B. Bischl, F. Hutter. Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML. Journal of Artificial Intelligence Research (79), 2024.

In conclusion, while AutoAI holds great promise for making AI technology more accessible, it also presents challenges in terms of system design, maintenance, computational cost, job market impact, and fairness. Addressing these issues will be essential for ensuring that AutoAI fulfils its potential to democratise AI while remaining robust, ethical, and efficient.

11. Conclusion

This report provides an overview of the progress in AutoAI benchmarking with a focus on the main topics researched within work package 7. Besides this, we included a section on hardware as well as language processing, which has become of major interest in society.

In the context of TAILOR, a book chapter **[EgeEtAl24]** has been delivered on best practices for dataset development (which essentially can be used for benchmarking machine learning research). A similar best practices document for creating AutoML benchmarks does not yet exists and would be very useful to steer the many contributions that are being developed.

During the TAILOR project, various efforts have been made that have an emphasis on tabular surrogate benchmarks (such as NAS-Bench-101, 201 and 301). One advantage of such surrogate benchmark is that many experiments can be ran without requiring significant compute time: Running HPO experiments takes a significant amount of time, but when executing it on a surrogate benchmark can be done at less time. While these efforts all come with their specific benefit, there is still room for filling specific gaps, for example a large-scale surrogate benchmark centred around learning curves (indeed YAPHO Gym, LCBench or LCDB1.0 all do provide them to a certain degree, but come with their specific limitations), large language models, etc.

TAILOR partners have also worked on the infrastructure required for benchmarking. Major contributions to OpenML have been made, including the ability to define reproducible benchmarks through benchmarking suites, supporting modern binary data formats, and migrating to a scalable software stack based on kubernetes. In a collaboration with other leading ML platforms, we have developed the new “croissant” metadata format for describing ML datasets. This format is now supported by many popular ML platforms and is used at the NeurIPS Datasets and Benchmarks track.

New AI benchmarks have been developed, and existing ones improved. For cross-domain meta-learning, *Meta-Album* is a new Meta-Dataset that contains 40 image datasets on 10 different domains. Meta-learning challenges for NAS have been run and can serve as inspiration for future benchmarks. The OpenML100, OpenML-CC18, and the AutoML benchmark are sets of benchmarking suites with tabular data each focusing on different data-readiness levels. The AutoML benchmark is not only the de facto standard for tabular AutoML research, the benchmarking suite and the accompanying software tool is also adopted by the industry for development and testing.

Future benchmarks should consider multi-objective optimisation and multiple modalities. Considering objectives such as fairness or energy consumption in addition to predictive accuracy is important, especially with ML models becoming increasingly pervasive in society. We need dedicated benchmarks to aid research which focuses on multi-objective optimization and to evaluate deployed ML systems. Additionally, most current benchmarks only use one modality. Modern models frequently make use of multiple modalities together, such as text, images, and structured data, so existing benchmarks should be updated or new ones developed.

In the context of Multi-objective AutoML, more research is required to combine several trustworthy objectives. In particular, robustness measures must be clearly defined and used together with performance metrics (like accuracy) in AutoML tasks. Robustness can be

defined in many different ways (robustness against (i) noise in the input data, (ii) input perturbations, or (iii) stochasticity during the training process, etc.), and has been deeply analysed in the context of optimisation. When it comes to robustness against input perturbations, the work by Konig et al (2024) provides a first benchmark study that compares various robustness verification techniques against each other. A logical next step would be to define a community benchmark, that is easily used and extended by others.

The analysis of robustness in the optimisation domain could serve as a basis to define robustness in the context of machine learning. Other trustworthy objectives, like explainability and fairness, also require more attention. One easy step forward is to use the existing measures for these objectives and include them into a multiobjective approach. Another important trustworthy objective to consider is ethics. The research done in value-alignment in the context of TAILOR [MonSie2021] is a good starting point to build a multi-objective approach.

References

Work with TAILOR partners is prefaced with a boldfaced indicator.

[AmirianEtAl20] Amirian J, Zhang B, Castro FV, Baldelomar JJ, Hayet JB, Pettré J. Opentraj: Assessing prediction complexity in human trajectories datasets. InProceedings of the Asian Conference on Computer Vision 2020. https://github.com/crowdbotp/OpenTraj

[AkhtarEtAl24] Akhtar M., Benjelloun O., Conforti C., Gijsbers P., Giner-Miguelez J., Jain N., Kuchnik M., Lhoest Q., Marcenac P., Maskey M., Mattson P., Oala L., Ruyssen P., Shinde R., Simperl E., Thomas G., Tykhonov S., Vanschoren J., van der Velde J., Vogler S., Wu C.J. Croissant: A Metadata Format for ML-Ready Datasets. SIGKDD Workshop on Data Management for End-to-End Machine Learning, 2024.

[BarEtAl24] Rafael Barbudo, Aurora Ramírez, José Raúl Romero: Grammar-based evolutionary approach for automated workflow composition with domain-specific operators and ensemble diversity. Appl. Soft Comput. 153: 111292 (2024).

[BazEtAl21] El Baz, A., Guyon, I., Liu, Z., Treguer, S., van Rijn, J.N. and Vanschoren, J.: Advances in MetaDL: AAAI 2021 challenge and workshop. AAAI Workshop on Meta-Learning and MetaDL Challenge, 1-16, 2021.

[BenjaminsEtAl21] Benjamins, Carolin; Eimer, Theresa; Schubert, Frederik; Biedenkapp, André; Rosenhan, Bodo; Hutter, Frank; Lindauer, Marius CARL: A Benchmark for Contextual and Adaptive Reinforcement Learning, In: Workshop on Ecological Theory of Reinforcement Learning (EcoRL@NeurIPS'21), 2021.

[BenjaminsEtAl22] Benjamins, Carolin; Eimer, Theresa; Schubert, Frederik; Mohan, Aditya; Biedenkapp, André; Rosenhan, Bodo; Hutter, Frank; Lindauer, Marius, Contextualize Me – The Case for Context in Reinforcement Learning, In: arXiv:2202.04500, 2022.

[BennetEtAl21] Bennet P, Doerr C, Moreau A, Rapin J, Teytaud F, Teytaud O. Nevergrad: black-box optimisation platform. ACM SIGEVOlution. 2021,

[BiedenKappEtAL22] Biedenkapp, André; Dang, Nguyen; Krejca, Martin S.; Hutter, Frank; Doerr, Carola, “Theory-inspired Parameter Control Benchmarks for Dynamic Algorithm Configuration”, In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'22), 2022.

[BischlEtAl16] Bischl B, Kerschke P, Kotthoff L, Lindauer M, Malitsky Y, Fréchette A, Hoos H, Hutter F, Leyton-Brown K, Tierney K, Vanschoren J. Aslib: A benchmark library for algorithm selection. Artificial Intelligence. 2016.

[BischlEtAl19] Bischl, B. and Casalicchio, G. and Feurer, M. and Hutter, F. and Lang, M. and Mantovani, R. G. and van Rijn, J. N. and Vanschoren, J. OpenML benchmarking suites, arXiv preprint arXiv:1708.03731. 2019.

[BloEtAl15] Blot A, Aguirre H, Dhaenens C, Jourdan L, Marmion ME, Tanaka K. Neutral but a winner! How neutrality helps multiobjective local search algorithms. InInternational Conference on Evolutionary Multi-Criterion Optimization 2015 Mar 29 (pp. 34-47). Springer, Cham.

[BlotEtAl16] Blot A, Hoos HH, Jourdan L, Kessaci-Marmion MÉ, Trautmann H. MO-ParamILS: A multi-objective automatic algorithm configuration framework. In International Conference on Learning and Intelligent Optimisation 2016 May 29 (pp. 32-47). Springer, Cham.

[BloEtAl17] Blot A, Jourdan L, Kessaci MÉ. Automatic design of multi-objective local search algorithms: case study on a bi-objective permutation flowshop scheduling problem. InProceedings of the Genetic and Evolutionary Computation Conference 2017 Jul 1 (pp. 227-234).

[BraEtAl21] Brazdil, P., van Rijn, J.N., Soares, C., Vanschoren, J. Metalearning. Applications to Automated Machine Learning and Data Mining. Springer 2021.

[CarrionEtAl22] Design and baseline results. Dustin Carrión-Ojeda, Hong Chen, Adrian El Baz, Sergio Escalera, Chaoyu Guan, Isabelle Guyon, Ihsan Ullah, Xin Wang, and Wenwu Zhu.

[CelikEtAl21] Celik, Bilge, and Joaquin Vanschoren. "Adaptation strategies for automated machine learning on evolving data." IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[CoeLec02] Coello CC, Lechuga MS. MOPSO: A proposal for multiple objective particle swarm optimization. InProceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600) 2002 May 12 (Vol. 2, pp. 1051-1056). IEEE.

[DebEtAl02] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," in IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, April 2002, doi: 10.1109/4235.996017.

[DeFEtAl22] De Filippo A, Borghesi A, Boscarino A, Milano M. HADA: An automated tool for hardware dimensioning of AI applications. Knowledge-Based Systems. 2022 Sep 5;251:109199.

[DelEtAl21] Delange, Matthias, et al. "A continual learning survey: Defying forgetting in classification tasks." IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[DerEtAl16] Bilel Derbel, Arnaud Liefooghe, Qingfu Zhang, Hernan Aguirre, Kiyoshi Tanaka. Multi-objective Local Search Based on Decomposition. International Conference on Parallel Problem Solving from Nature (PPSN 2016), 2016, Edinburgh, United Kingdom. pp.431 - 441

[DongYang19] Xuanyi Dong and Yi Yang, “NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search,” Proceedings of the International Conference on Learning Representations, 2019.

[DubEtAl15] J. Dubois-Lacoste, M. López-Ibáñez and T. Stützle, "Anytime Pareto local search", Eur. J. Oper. Res., vol. 243, no. 2, pp. 369-385, 2015.

[EggenspergerEtAl21] Eggensperger, Katharina; Müller, Philipp; Mallik, Neeratyoy; Feurer, Matthias; Sass, René; Klein, Aaron; Awad, Noor; Lindauer, Marius; Hutter, Frank, “HPOBench: A Collection of Reproducible Multi-Fidelity Benchmark Problems for HPO”, In: Vanschoren, J.; Yeung, S. (Ed.): Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021. https://github.com/automl/HPOBench

[EgeleEtAl24] R. Egele, J. C. S. Jacques Júnior, J. N. van Rijn, I. Guyon, X. Baró, A. Clapés, P. Balaprakash, S. Escalera, T. B. Moeslund, J. Wan, “AI Competitions and Benchmarks: Dataset Development”, CoRR abs/2404.09703 (2024)

[EimerEtAl21] Eimer, Theresa; Biedenkapp, André; Reimer, Maximilian; Adriaensen, Steven; Hutter, Frank; Lindauer, Marius, “DACBench: A Benchmark Library for Dynamic Algorithm Configuration”, In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI'21), ijcai.org, 2021.

[ElskenEtAl] Elsken, T., Staffler, B., Metzen, J. H., & Hutter, F. (2020). Meta-learning of neural architectures for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12365-12375).

[EmmEtAl16] Emmerich M, Yang K, Deutz A, Wang H, Fonseca CM. A multicriteria generalization of bayesian global optimization. InAdvances in Stochastic and Deterministic Global Optimization 2016 (pp. 229-242). Springer, Cham.

[FeuEtAl19] Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Auto-sklearn: efficient and robust automated machine learning. In Automated Machine Learning 2019 (pp. 113-134). Springer, Cham.

[FeuEtAl21] Feurer, Matthias, et al. "Openml-python: an extensible python api for openml." Journal of Machine Learning Research 22.100 (2021): 1-5.

[GalEtAl20] Galuzio PP, de Vasconcelos Segundo EH, dos Santos Coelho L, Mariani VC. MOBOpt—multi-objective Bayesian optimization. SoftwareX. 2020 Jul 1;12:100520.

[GarcésEtAl21] G. Garcés, J. A. Silvestre, J. Jorge, A. Giménez, J. Iranzo, P. Baquero, N. Roselló, A. Pérez, J. Civera, A. Sanchis, A. Juan. Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data

Filtering/Verbatimization. In Proc. Interspeech 2021, pp. 3695–3699, Brno (Czech Republic), 2021.

[GijsbersEtAl21] Gijsbers, Pieter, et al. "Meta-Learning for Symbolic Hyperparameter Defaults." GECCO 2021 (2021).

[GijsbersEtAl20] Gijsbers, Pieter, and Joaquin Vanschoren. "GAMA: a General Automated Machine learning Assistant." ECMLPKDD 2020 (2020).

[GijsbersEtAl24] Gijsbers P., Bueno M.L.P, Coors S., LeDell E., Poirier S., Thomas J., Bischl B., Vanschoren, J. AMLB: an AutoML Benchmark. Journal of Machine Learning Research (JMLR), 25 (101), pp. 1-65.

[HutEtAl09] Hutter F, Hoos HH, Leyton-Brown K, Stützle T. ParamILS: an automatic algorithm configuration framework. Journal of Artificial Intelligence Research. 2009 Oct 30;36:267-306.

[KansalEtAl21] Kansal, Raghav, Javier Duarte, Hao Su, Breno Orzari, Thiago Tomei, Maurizio Pierini, Mary Touranakou, and Dimitrios Gunopulos. "Particle cloud generation with message passing generative adversarial networks." Advances in Neural Information Processing Systems 34 (2021): 23858-23871.

[KohEtAl21] Koh PW, Sagawa S, Xie SM, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, Lee T, David E. Wilds: A benchmark of in-the-wild distribution shifts. In International Conference on Machine Learning 2021.

[KönigEtAl20] Matthias König, Holger H Hoos and Jan N van Rijn. Towards Algorithm-Agnostic Uncertainty Estimation: Predicting Classification Error in an Automated Machine Learning Setting. In ICML Workshop on Automated Machine Learning. 2020.

[KönigEtAl21] Matthias König, Holger H Hoos and Jan N van Rijn. Speeding Up Neural Network Verification via Automated Algorithm Configuration. In ICLR Workshop on Security and Safety in Machine Learning Systems. 2021.

[KönigEtAl24] König, M., Bosman, A. W., Hoos, H. H., & van Rijn, J. N. (2024). Critically assessing the state of the art in neural network verification. *Journal of Machine Learning Research*, *25*(12), 1-53.

[KostovskaEtAl22] Kostovska, A., Bogatinovski, J., Džeroski, S., Kocev, D., & Panov, P. (2022). A catalogue with semantic annotations makes multilabel datasets FAIR. *Scientific Reports*, *12*(1), 7267.

[KlyuchnikovEtAl20] Nikita Klyuchnikov, Ilya Trofimov, Ekaterina Artemova, Mikhail Salnikov, Maxim Fedorov, Evgeny Burnaev. “NAS-Bench-NLP: Neural Architecture Search Benchmark for Natural Language Processing”, arXiv 2020

[LakeEtAl15] Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015: https://paperswithcode.com/sota/few-shot-image-classification-on-omniglot-1-1

[LiEtAl19] Li, Y.-F., Wang, H., Wei, T., & Tu, W.-W. (2019). Towards Automated Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 4237-4244.

[LithVanschoren21] John W. van Lith and Joaquin Vanschoren. From strings to data science: practical framework for automated string handling. ECMLPKDD workshop on Automated Data Science, 2021.

[LuEtAl19] Lu Z, Whalen I, Boddeti V, Dhebar Y, Deb K, Goodman E, Banzhaf W. Nsga-net: neural architecture search using multi-objective genetic algorithm. InProceedings of the Genetic and Evolutionary Computation Conference 2019 Jul 13 (pp. 419-427).

[MadeyskiLewowski20] Madeyski, L., Lewowski, T.: Mlcq: Industry-relevant code smell data set. In: Proceedings of the Evaluation and Assessment in Software Engineering, pp. 342–347(2020)

[MartinezEtAl22] José Ángel Morell Martínez, Zakaria Abdelmoiz Dahi, Francisco Chicano, Gabriel Luque, Enrique Alba: Optimising Communication Overhead in Federated Learning Using NSGA-II. EvoApplications 2022: 317-333

[MehrotraEtAl21] Abhinav Mehrotra, Alberto Gil C. P. Ramos, Sourav Bhattacharya, Łukasz Dudziak, Ravichander Vipperla, Thomas Chau, Mohamed S Abdelfattah, Samin Ishtiaq, Nicholas Donald Lane. “NAS-Bench-ASR: Reproducible Neural Architecture Search for Speech Recognition”, Proceedings of the International Conference on Learning Representations, 2021.

[MısSeb17] Mısır M, Sebag M. Alors: An algorithm recommender system. Artificial Intelligence. 2017 Mar 1;244:291-314.

[MonSie20] Nieves Montes, Carles Sierra: Value-Alignment Equilibrium in Multiagent Systems. TAILOR 2020: 189-204

[MorellEtAl24] José Á. Morell, Zakaria Abdelmoiz Dahi, Francisco Chicano, Gabriel Luque, Enrique Alba: A multi-objective approach for communication reduction in federated learning under devices heterogeneity constraints. Future Gener. Comput. Syst. 155: 367-383 (2024)

[NieEtAl19] Nie Y, Williams A, Dinan E, Bansal M, Weston J, Kiela D. Adversarial NLI: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599. 2019. https://arxiv.org/pdf/1910.14599.pdf

[OlsEtAl16] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. In Proc. ACM-GECCO, pages 485–492. ACM Press, 2016

[RakEtAl19] Herilalaina Rakotoarison, Marc Schoenauer, Michèle Sebag. Automated Machine Learning with Monte-Carlo Tree Search. IJCAI-19 - 28th International Joint Conference on Artificial Intelligence, Aug 2019, Macau, China. pp.3296-3303

[ReaEtAl20] Real E, Liang C, So D, Le Q. Automl-zero: Evolving machine learning algorithms from scratch. InInternational Conference on Machine Learning 2020 Nov 21 (pp. 8007-8019).

[PanEtAl24] Pan J, Falker S., Berkenkamp F., Vanschoren J, MALIBO: Meta-learning for Likelihood-free Bayesian Optimization. Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[PetkovićEtAl22] Petković, M., Lucas, L., Levatić, J., Breskvar, M., Stepišnik, T., Kostovska, A., ... & Kocev, D. (2022). Machine-learning ready data on the thermal power consumption of the Mars Express Spacecraft. *Scientific Data*, *9*(1), 229.

[RajanEtAl21] Rajan, Raghu; Diaz, Jessica Lizeth Borja; Guttikonda, Suresh; Ferreira, Fabio; Biedenkapp, André; von Hartz, Jan Ole; Hutter, Frank, MDP Playground: A Design and Debug Testbed for Reinforcement Learning, In: arXiv:1909.07750, 2021.

[SalinasEtAl21] Palacios Salinas, N. Rosaura and Baratchi, M. and van Rijn, J. N. and Vollrath, A, "Automated Machine Learning for Satellite Data: Integrating Remote Sensing Pre-trained Models into AutoML Systems", in Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML-PKDD 2021, 2021. https://github.com/palaciosnrps/automl-rs-project

[SiemsEtAl20] Julien Siems, Lucas Zimmer, Arber Zela, Jovita Lukasik, Margret Keuper, Frank Hutter. "NAS-Bench-301 and the Case for Surrogate Benchmarks for Neural Architecture Search", arXiv 2020

[SinghVanschoren23a] Singh P, and Vanschoren J. AutoML for outlier detection with optimal transport distances. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23), 2023.

[SinghVanschoren23b] Singh P, and Vanschoren J. Applications of Optimal Transport Distances in Unsupervised AutoML NeurIPS 2023 Workshop Optimal Transport and Machine Learning. 2023

[SoomlekEtAl21] Soomlek, C., van Rijn, J.N., Bonsangue, M.M.: Automatic human-like detection of code smells, Discovery Science 2021.

[SouzaEtAl] Marcelo de Souza, Marcus Ritt and Manuel López-Ibáñez. Capping Methods for the Automatic Configuration of Optimization Algorithms. Computers & Operations Research 139: 105615.

[TjengEtAl19] Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating Robustness of Neural Networks with Mixed Integer Programming. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), 2019.

[TriantafillouEtAll19] Triantafillou E, Zhu T, Dumoulin V, Lamblin P, Evci U, Xu K, Goroshin R, Gelada C, Swersky K, Manzagol PA, Larochelle H. Meta-dataset: A dataset of datasets for learning to learn from few examples. arXiv preprint arXiv:1903.03096. 2019. https://github.com/google-research/meta-dataset

[UllahEtAl22] Multi-domain Meta-Dataset for Few-Shot Image Classification. Ihsan Ullah, Dustin Carrion, Sergio Escalera, Isabelle M Guyon, Mike Huisman, Felix Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, Phan Anh Vu. *NeurIPS 2022 Datasets and Benchmarks Track.*

[VanHut18] Van Rijn, Jan N., and Frank Hutter. "Hyperparameter importance across datasets." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

[VettoruzzoEtAl24] Vettoruzzo A., Vanschoren J., Bouguelia M-R., Rognvaldsson, T. Learning to learn without forgetting using attention. Third Conference on Lifelong Learning Agents (CoLLAs), 2024.

[VinyalsEtAl16] Vinyals O, Blundell C, Lillicrap T, Wierstra D. Matching networks for one-shot learning. Advances in neural information processing systems. 2016: https://paperswithcode.com/dataset/miniimagenet-1

[WasalaEtAl] Wasala, Julia; Baratchi, Mitra; Marselis, Suzanne; Arp, Laurens; Longepe, Nicolas; Hoos, Holger, “AutoSR4EO: An AutoML Approach to Super-Resolution for Earth Observation Images”.

[WeertsEtAl20] Weerts, Hilde JP, Andreas C. Mueller, and Joaquin Vanschoren. "Importance of tuning hyperparameters of machine learning algorithms." arXiv preprint arXiv:2007.07588 (2020).

[WeverEtAl21] Wever M, Tornede A, Mohr F, Hullermeier E. AutoML for Multi-Label Classification: Overview and Empirical Evaluation. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2021.

[YanEtAl21] Shen Yan, Colin White, Yash Savani, Frank Hutter. “NAS-Bench-x11 and the Power of Learning Curves”, CVPR workshop on NAS 2021.

[YingEtAl19] Chris Ying, Aaron Klein, Esteban Real, Eric Christiansen, Kevin Murphy, Frank Hutter, "Nas-bench-101: Towards reproducible neural architecture search," Proceedings of the International Conference on Machine Learning, 2019.

[ZelaEtAl20] Arber Zela, Julien Siems, Frank Hutter. “NAS-Bench-1Shot1: Benchmarking and Dissecting One-shot Neural Architecture Search”. Proceedings of the International Conference on Learning Representations, 2020.

[ZelaEtAl22] Zela, Arber; Siems, Julien; Zimmer, Lucas; Lukasik, Jovita; Keuper, Margret; Hutter, Frank, “Surrogate NAS Benchmarks: Going Beyond the Limited Search Spaces of Tabular NAS Benchmarks”, In: International Conference on Learning Representations (ICLR), 2022.

[ZiEtAl21] Chaojian Li, Zhongzhi Yu, Yonggan Fu, Yongan Zhang, Yang Zhao, Haoran You, Qixuan Yu, Yue Wang, Yingyan Lin. HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark. Proceedings of the International Conference on Learning Representations, 2021.