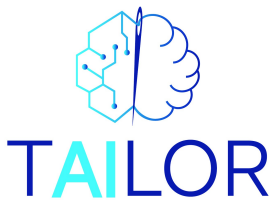

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

AutoAI Benchmarks v.2 Report

Document type (nature)	Report
Deliverable No	7.6
Work package number(s)	7
Date	Due 31 August 2022
Responsible Beneficiary	ULEI, ID #7
Author(s)	Koen van der Blom
Publicity level	Public
Short description (Please insert the text in the Description of Deliverables in the Appendix 1.)	Version 2 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**
1.0	November 2022	first version	Koen van der Blom

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Fredrik Heintz	1 - LiU	2022-11-05
Peter Flach	16 - UNIBRIS	2022-11-14

Table of Contents

Summary of the Report

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

Organisation

The following experts from the TAILOR consortium have been involved in the writing of this report, based on materials collected across a broad range of project partners:

Partner ID / Acronym	**Name**	**Role**
#7, ULEI	Koen van der Blom Laurens Arp Mitra Baratchi	WP7, T7.2, T7.3 Leader
#12, TUE	Pieter Gijsbers Joaquin Vanschoren	T7.5 Leader
#17, ALU-FR	Eddie Bergman	T7.1 Leader
#33, NKUA	Dimitrios Gunopulos	T7.1 Participant
#3, INRIA	Marc Schoenauer	T7.4 Leader
#10, UNIBO	Andrea Borghesi	Participant
#43, UPV	Alfons Juan	Participant
#23, CRIL	Daniel Le Berre	Participant

1. Introduction

This second version of the AutoAI Benchmarks report is limited to the discussion of developments since the first version, i.e., starting from 1 September 2021. As in the previous version, the report is organised by topic, and work originating from TAILOR is highlighted in **bold**. 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. With this in mind, work from outside the TAILOR network but related to the tasks is also covered.

2. AutoML in the wild [T7.1, ALU-FR]

Facilitate the usability of machine learning by non-machine-learning-experts who have data and a clear target to predict, but who are not familiar enough with machine learning to know which neural architecture or machine learning pipeline to use, and how to set its hyperparameters.

There has been a large increase in Neural Architecture Search (NAS) benchmarks in the last 2 years, to move beyond tabular setups and use surrogate models to predict the performance of architectures, allowing us to extend NAS benchmarks to larger spaces.

- **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 gone through great strides to develop NASLib to provide a uniform interface to all of the available NAS benchmarks and allows 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 it has been extended to integrate additional new benchmarks and more importantly, we are actively working towards multi-objective benchmarks, an important test bed for research on new multi-objective optimizers, a trending requirement of real AutoML systems in the wild.

- **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

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 **[KanEtAl21]** 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.
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
- **[KanEtAl21]** 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.

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. The paper for this work is currently being finalised, and the dataset will be hosted publicly along with the paper once it is submitted

- **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, a method for increasing the resolution of images, for satellite image datasets. We introduce a new real-world single-image SR dataset, SENT-NICFI. 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.

- **SENT-NICFI: Wasala, Julia; Baratchi, Mitra; Marselis, Suzanne; Arp, Laurens; Longepe, Nicolas; Hoos, Holger, "AutoSR4EO: An AutoML Approach to Super-Resolution for Earth Observation Images"**

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

Goal: Expand the scope of AutoML, an important special case of AutoAI, to diverse and rich learning settings.

3.1 Reinforcement Learning

Reinforcement learning (RL) is a panacea to address a wide variety of problems due to its flexibility to describe and optimise problems. However, the training costs of these agents are often excessively expensive 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:

- 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.
- 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.

- 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.

3.2 Dynamic Algorithm Configuration

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 in an effort 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 a **best paper award!**

- **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

- **[EimEtAl21]** 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]

Goal: Automatically detect when an AI system (such as a classifier, predictor or reasoning engine obtained from an AutoAI system) gets 'off-track' and can no longer be used safely and reliably.

The sat4j platform now supports Dynamic Algorithm Configuration in order 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 Self-monitoring AI systems for reasoning engines.

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

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

Goal: Develop multi-objective AutoAI methods to automatically determine the best tradeoffs between performance and other objectives, e.g., derived from the six dimensions of trustworthiness.

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 optimization.

- **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]

Goal: 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.

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 under review at JMLR and has a preprint on arxiv.

- **AMLB: an AutoML Benchmark. Pieter Gijsbers, Marcos L. P. Bueno, Stefan Coors, Erin LeDell, Sébastien Poirier, Janek Thomas, Bernd Bischl, Joaquin Vanschoren. https://arxiv.org/abs/2207.12560**
- **https://openml.github.io/automlbenchmark/**

Additionally, they provide several tools to the community to analyse the obtained results, including an interactive Shiny app and multiple Python notebooks.

There are future 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 centred around climate change data will be explored.

6.2. MetaDL and Meta-Album

Meta-Learning, whose goal is to learn across datasets, can be seen as some form of AutoAI. And 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. Hence 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, date of the publication of the results. More details are given in Deliverable 2.3.

- **NeurIPS'22 Cross-Domain MetaDL competition: 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-KVmHNIFCweiljbPwimmAXMvHHf7e/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 currently submitted to NeurIPS and as such available on OpenReview. Further details about Meta-Album are presented in Deliverable 2.3 (foundational benchmarks and challenges).

- **Meta-Album: 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.** ***Under review at the NeurIPS 2022 Datasets and Benchmarks Track.***

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 performances and budget limits is called hardware dimensioning. This is a challenging and complex task for many companies that need assistance from AI experts in this task. 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 which 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 in more resources than those strictly needed) and/or using heuristics, with the risk of sub-optimal solutions.

To address this issue, HADA **[DeFEtAI22]** was proposed, an automated approach for HW dimensioning where ML models are embedded within an optimization 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 data set which we then made publicly available. The data set consists of benchmarks of online/offline optimization 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 optimization models for similar tasks. The data set contains both numerical and categorical data; it is born digital. Data is mostly quantitative. The data is raw, with minimal pre-processing. The data is stored using standard format, such as CSV files and/or a binarized 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 on the accompanying paper **[DeFEtAI22]**

- **[DeFEtAI22] 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

We (UPV) introduce 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

verbatimized transcripts of all speeches based on speech data filtering and verbatimization techniques. Additionally, 18 hours of transcribed speeches were manually verbatimized 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 verbatimization 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.

- **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. Conclusion

Since the initial report on AutoAI benchmarks advancements have been made in directions covering all the tasks of the work package. To highlight a few, these advancements include benchmarking progress in neural architecture search (NAS), hyperparameter optimisation (HPO), reinforcement learning (RL), dynamic algorithm configuration (DAC), automated machine learning (AutoML), and meta-learning for deep learning (MetaDL). In addition, progress has also been made in hardware dimensioning of algorithms and ML for language processing, which are areas more loosely connected to the tasks in WP7.