Benchmark AutoML

22.september. 2022

Project M.Sc. ITTI
Sara Jamali, 218100174
Supervisor: Dr. Sebastian Bader

Contents

1 Abstract

This project presents a benchmark for two Open-source Automated Machine Learning (AutoML) tools: H2o-AutoML and Auto-keras. First, some articles about this topic will be summarized and then the characteristics of these tools (Auto-Keras and H2o-AutoML) will be analyzed. After that, the performance of these two systems will be observed on 10 different Datasets (divided into 5 Binary- and 5 Multi-Classification tasks) from OpenML. In this project, the comparison criterion of these two systems is the execution time as well as the accuracy of each classifier.

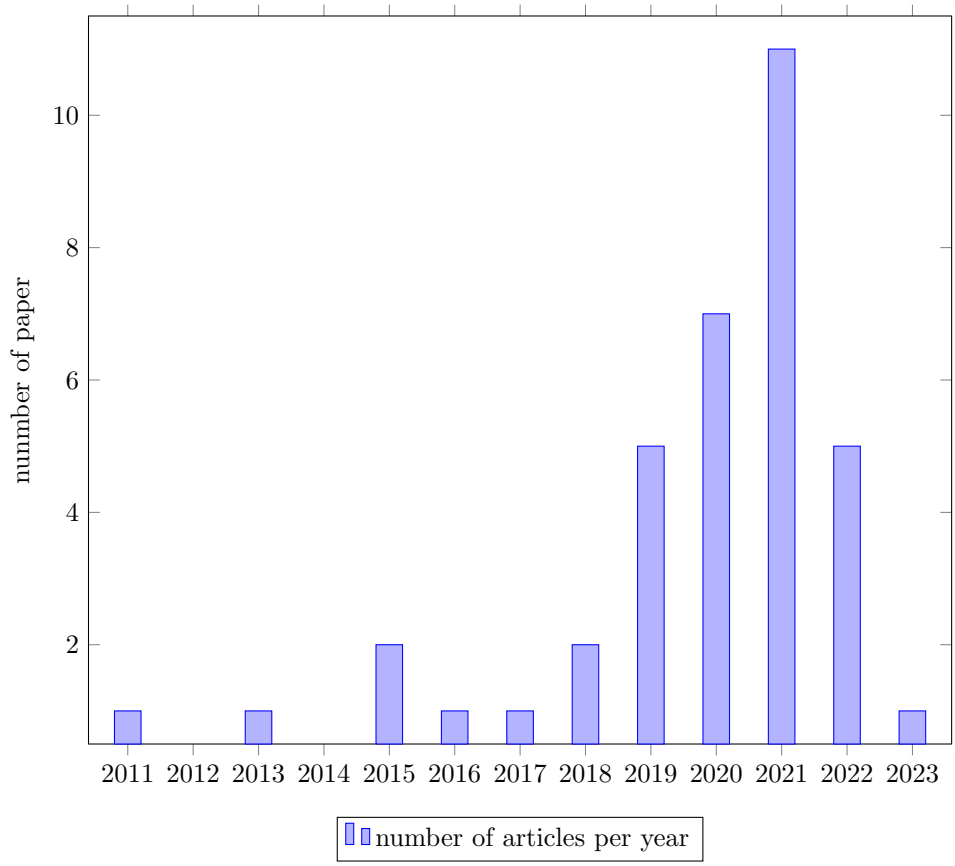
2 Introduction

In recent years, the use of machine learning methods has become prevalent in various fields. Finding the best algorithm for beginners also optimizing hyperparameters for experts can always be time-consuming. AutoML stands for Automated Machine Learning, and it is an effort to fully automate end to end machine learning process and try to find the best machine learning algorithm. This project focuses on comparing the performance of two Automated Machine learning: H2O AutoML and Auto-Keras. Also, for evaluation, these systems will use 10 different public Datasets from the OpenML for binary- and multi-class Classification for each system.
On the other hand, since we want to compare different systems, we need the same standard to ensure that the benchmarking is fair. For this purpose, various criteria have been measured to show the performance of these systems, such as the accuracy of the classifier and also the Time, that each system needs to find the best model for the dataset. Also, each dataset is divided into training and test Data with ratios of 0.80 and 0.20.

3 Related Work

In recent years, research on AutoML has increased. In the first step, to find articles related to this project, the keywords such as "Benchmarking automated machine learning, AutoML, Comparing AutoML, Evaluation different AutoML" have been used. After collecting some papers, by examining the year of publication of the articles, which can be observed in the diagram Figure 1, it can be concluded that the majority of the focus on this issue was after 2019. Also, by examining the articles, it can be supposed that in Germany, the "University of Freiburg" and "Leibniz University Hannover" [**1**] have a particular focus on this topic. By checking the website of this university, everyone can find answers to many questions in this field.
The oldest paper about Automl dated back to 2013 and introduced Auto-weka [**2**] as the first automatic system. By reviewing the found research, it can be said that most of the published articles in this field until 2021 can be divided into two major topics.
In the early years, most of the focus was on creating automatic systems for machine learning, and in recent years, this idea has been extended to other applications of machine learning, such as Automatic recommendation systems (Auto-Rec)[**3**], Automatic time series systems (Auto-series)[**4**], Automated Graph Machine Learning (AutoGL)[**5**], etc.
With the increasing number of automatic systems, measuring their performance and comparing their efficiency has become a big challenge, which forms the second part of the article. After reviewing the papers, it was found that some of them compare the performance of the different systems to each other, and in some of them, different optimization methods are evaluated. Also, in recent years, many articles have studied comparing the performance of automatic systems with humans[**6**; **7**; **12**]. After 2021, the most focus of both groups of articles has been reduced on machine learning and focused on deep learning and Automated Deep Learning (AutoDL).

Fig. 1. number of articles per year



Because this project focuses on creating a benchmarking platform, the following table summarizes the related articles. As mentioned, creating a benchmark is straightforward. Each benchmark includes an optimization metric, dataset, and specific storage space to save the comparison result. As a result, these points have been considered in the review of articles.

Title	Authors	AutoML system	Benchmark Dataset	Quality metric	Hardware
Benchmarking Automatic Machine Learning Frameworks (2018)[**8**]	Adithya Balaji et al.	auto-sklearn, TPOT, H2O, auto-ml	open source Dataset from OpenML (57 classification and 30 regression)	MSE and F1 score	Amazon Web Services (AWS) Batch framework, run a system on the run on Intel Xeon E5-2666 v3 processors.

Title	Authors	AutoML system	Benchmark Dataset	Quality metric	Hardware
An Open Source AutoML Benchmark (2019)[**9**]	Pieter Gijsbers et al.	Auto-Weka, Auto-sklearn, TPOT H2O-automl	39 classification datasets from previous papers	AUROC for binary classification log loss for multiclass classification The measures are estimated with ten-fold cross-validation.	Use standard m5.2xlarge instances available on Amazon Web Services
"Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools (2019)[**10**]	Anh Truong et al.	TransmogrifAL, H2O, Darwin, DataRobot, Google AutoML, Auto-sklearn, MLjar, TPOT, Auto-keras, Ludwig, Auto-weka, AzurML, H2o-Driverless AI	nearly 300 datasets from OpenML	'Accuracy' for binary and multiclass classification tasks 'Mean Squared Error (MSE)' for regression tasks.	run on Amazon EC2p2.xlarge instances, which provide 1 Tesla K80 GPU, 4 vCPUs (Intel Xeon E5-2686, 2.30Ghz), and 61 GiB of host memory.
On Evaluation of AutoML Systems (2020) [**11**]	Mitar Milutinovic et al.	CMU, ISI, MIT, NYU, SRI, TAMU, UCB	106 known datasets and 62 blind datasets from CIFAR, Geolife, Mice Protein, Retail sales, Amazon, Facebook..."	Different Quality metrics for different tasks and different datasets, like Accuracy, F1, MAE, RMSE, NM1	—

Title	Authors	AutoML system	Benchmark Dataset	Quality metric	Hardware
"Can AutoML outperform humans? An evaluation of popular OpenML datasets using AutoML Benchmark" (2020)[**6**]	Marc Hanussek et al.	TPOT, H2O, Auto-sklearn, AutoGluon	6 Dataset for Classification 6 Dataset for Regression with the most runs on OpenML	Auc, Acc, RMSE, MAE For reliable results, Used 10-fold cross-Validation	server host locally at Fraunhofer IAO. two Intel Xeon Silver 4114 CPUs @2.20Ghz, four 64GB DIMM DDR4 Synchronous 2666MHz memory modules, and two NVIDIA GeForce GTX 1080 Ti.
Leveraging Automated Machine Learning for Text Classification: Evaluation of AutoML Tools and Comparison with Human Performance (2020)[**12**]	Matthias Blohm et al.	TPOT, H2O, Auto-sklearn, AutoGluon	13 dataset that was used in past Kaggle competition	ACC, F1, roc-auc	Local server with two Intel Xeon Silver 4114 CPUs @2.20Ghz (yielding 20 cores in total), four 64GBDIMM DDR4 Synchronous 2666MHz memory modules and two NVIDIA GeForce GTX 1080 Ti.
A Comparison of AutoML Tools for Machine Learning, Deep Learning, and XGBoost (2021)[**13**]	Luis Ferreira et al.	Auto-keras, Auto-Pytorch, Auto-sklearn, H2O, TPOT, rminer, TransmogrifAI, AutoGluon	12 Datasets from the OpenML	MAE, AUC, F1	All experiments were executed using an Intel Xeon 1.70GHz server with 56 cores and 2TB of disk space.

Title	Authors	AutoML system	Benchmark Dataset	Quality metric	Hardware
Benchmark and Survey of Automated Machine Learning Frameworks (2021)[**14**]	Marc Zöller et al.	Dummy, Random Forest, TPOT, hpsklearn, auto-sklearn, Random Search, ATM, H2O AutoML	All data sets from the AutoML Benchmark suite are used.	log loss, ROC AUC	All experiments are conducted using n1-standard-8 virtual machines from the Google Cloud Platform equipped with Intel Xeon E5 processors with 8 cores and 30 GB memory3.
MedMNIST classification Decathlon: a Light-weight AutoML benchmark for Medical image Analysis (2021)[**15**]	Jiancheng Yang et al.	Auto-sklearn, Auto-keras, Google AutoML vision	MedMNIST	AUC, ACC	—

Table 1: comparison Table

The challenges in the field of designing a benchmark system include the need for ample storage space, long computing time, and the need for powerful hardware. Most articles have used Amazon Web service (AWS) to implement the system and store the results. Also, most systems used in benchmarks can be mentioned as H2O AutoML, Auto-Sklearn, and TPOT. Additionally, the most popular source for data is OpenML. The Performance measurement criterion in most articles for the Binary Classification task is AUC, for the Multi Classification task is log-loss, and for the Regression task is MAE.

4 Benchmark Suits

To define a benchmark, we need some requirements also, metrics to compare, datasets and frameworks, which are briefly explained below.

4.1 Requirement

To implement this benchmark, in the first step, sufficient theoretical and practical knowledge about machine learning is needed. Information about the selected AutoML systems and knowing their differences can be advantageous. Also, sufficient skill in programming with Python is one of the requirements of this project. In the next step, the knowledge of using AutoML systems is necessary.
One of the important requirements for installation and using H2o-AutoML is Java. Because H2o does not work without Java. Also, Auto-Keras is based on TensorFlow, and to use Auto-Keras, we need TensorFlow 2.3.0.
In addition, some Python libraries such as Matplotlib, Pandas, Sklearn, OpenML, and Scikit-learn are needed.

4.2 Dataset

A summary of the selected datasets for this benchmark can be observed in Table 2. The focus of this project was only on tabular data. Data were selected from OpenML for Binary and Multi-classification tasks, and also selected data did not include missing values.

Dataset ID	task	Instances	Features	Classes
12	Multi	2000	217	10
23	Multi	1473	10	3
31	Binary	1000	21	2
37	Binary	768	9	2
50	Binary	958	10	2
54	Multi	846	19	4
61	Multi	150	5	3
469	Multi	797	5	6
1464	Binary	748	5	2
1494	Binary	1055	42	2

Table 2: Datasets

4.3 AutoML Tools

This study compares two open-source AutoML: Auto-Keras and H2o-AutoML. If possible, all systems are run with default values, and both systems will be focused on deep learning tasks.

1. H2O: an automated machine learning, working on the H2O framework and using grid search to find the best model. H2O AutoML can be used for both scenarios, AutoML and AutoDL. H2o can

optimize some hyperparameters, such as the number of hidden layers and hidden units per layer, the learning rate, finding activation functions, etc.

2. Auto-Keras: focused on NAS (Neural Architecture Search). Auto-Keras is an open-source library for performing AutoML for deep learning models. The search is performed using Keras models via the TensorFlow API and uses the Bayesian Optimization method to tune some Hyperparameter such as the number of dense layers, finding activation function, etc.
The description of AutoML tools can be observed in Table 3.

AutoML	Framework	language	OS	senario
H2O-AutoML	H2O	Python, R	Linux, macOS, Windows	ML, DL
Auto-Keras	Keras	Python	Linux, macOS, Windows	DL

Table 3: Description of AutoML Tools

5 The Concept of benchmarking

As mentioned, the H2o-AutoML can be used for both machine learning and deep learning purposes. Auto-Keras is specific for Neural Architecture search. Since the benchmark should be fair; therefore, this study focuses on deep learning and compare their performance to find the best classifier with higher accuracy.
It is crucial to know that Auto-Keras does not support a cross-validation method. In this project, holdout validation is used for both AutoML tools. It means, for each prediction in binary- and Multi-class classification tasks, all datasets are split with a ratio of 80:20 for training and testing. Also, the training datasets are divided with a ratio of 85:15 for training data and Validation data purposes.
Also, some considerations were taken into this benchmark, as an example: it is essential for using H2o-AutoML that the value of (nfold) should be equal to zero. Because when the user chooses another validation method such as Hold-out, cross-validation is still enabled and models will still be validated using cross-validation only.
The maximum number of models to build in each tool is considered to be 5, which means each system select 5 different Neural Network Architect and train them to find the best solution for each dataset and choose the Network with high accuracy as a final classifier.
In addition, various criteria have been considered to compare the performance of these two AutoML tools, such as the accuracy of the classifier, Confusion Matrix, and the execution time that both systems need to execute each task.
The result of the performance of each tool will be automatically saved with the name (dataset name)-framework-metrics for accuracy and execution time and (dataset name)-framework-Classification for the result of the classification report so that it can be reviewed and used later.

6 Result

Figures 2,3, and 4 summarize the result of this benchmark. Figure 2 presents the execution time of the total tasks in this project. For this benchmark is 20 AutoML tasks executed. In this comparison, H2o-AutoML always requires much more running time than Auto-Keras, in other words, Auto-Keras is always faster than H2o-AutoML in both Binary- and Multi-class Classification and requires an average time of just 24 sec, but H2o needs 161 sec execution time in same conditions with Auto-Keras.
Regarding comparing each system's performance, the classifier's accuracy is considered as a Comparison criterion. Figures 3 and 4 show the result of the classification task. After analyzing these tools,

Auto-Keras shows poor performance in comparison with H2o-AutoML in both Binary- and Multi-class classification tasks, and it has just better performance in 2 datasets. The average accuracy per dataset in the binary classification task for Auto-Keras is 0.75, but for the H2o-AutoML, this value is 0.80. Also, this observation applies to Multi-class Classification, and clearly, the accuracy of the H2o-AutoML is still higher than the Auto-Keras. The average accuracy for H2o-AutoML is 0.73 and for Auto-Keras is 0.66.

Fig. 2. Execution Time

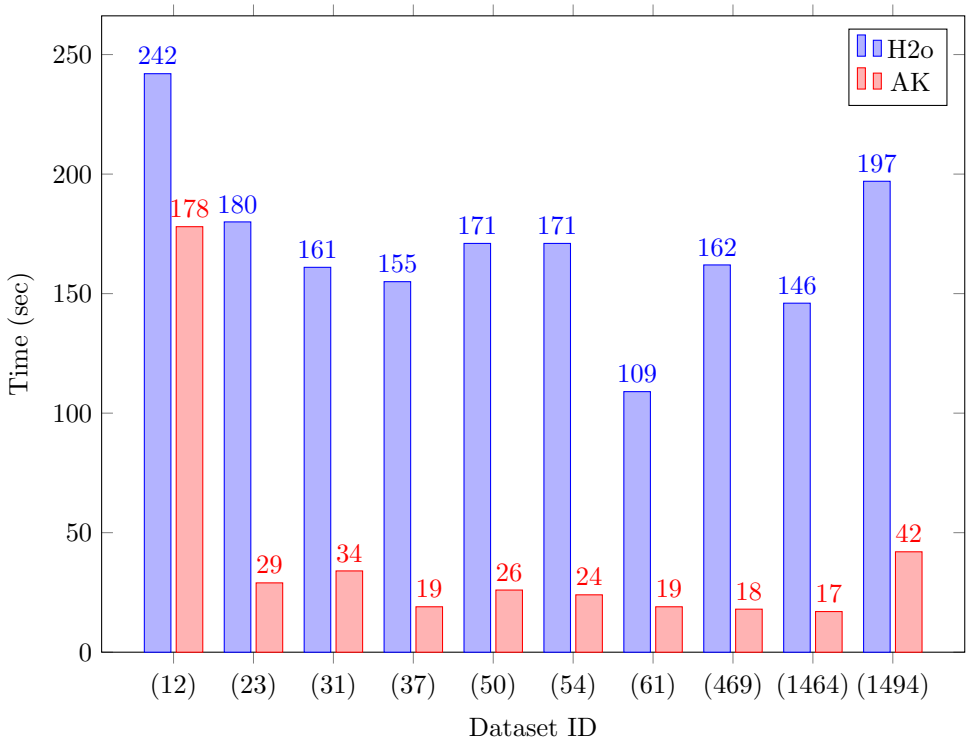


Fig. 3. Accuracy of Multi-Classification task

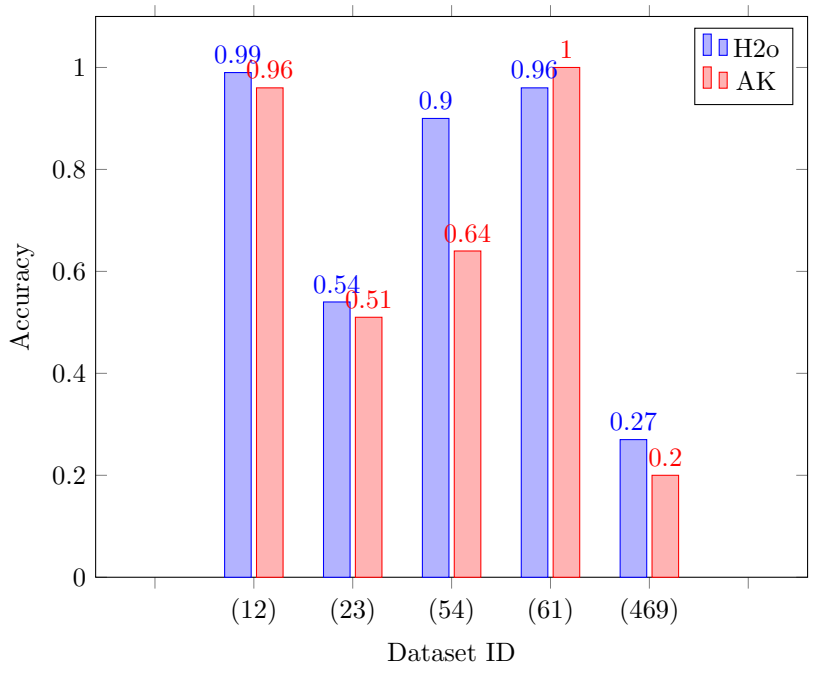
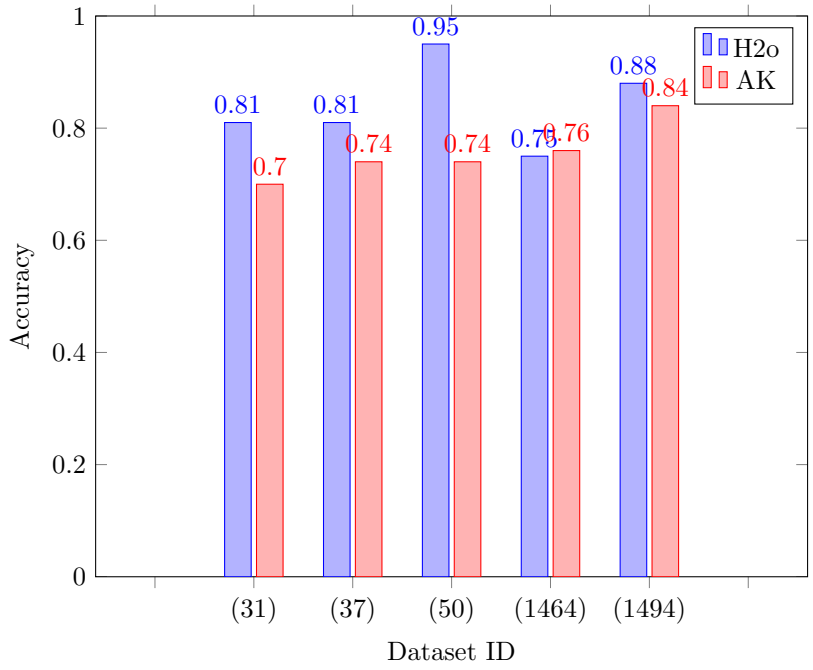


Fig. 4. Accuracy of Binary-Classification task



7 Conclusions

Two AutoML systems were investigated in this Project: H2o-AutoML and Auto-Keras. A set of 20 computational experiments consist of 10 datasets from OpenML for two different tasks (Binary- and Multi-class Classification) and for two AutoML tools. In this project, each tool is benchmarked by measuring the execution time and accuracy of the classifier.

As it was apparent from the results, the H2o-AutoML has higher accuracy, but it requires more computing time, which may be very problematic for large datasets. On the other hand, the Auto-Keras has a shorter execution time compared to H2o-AutoML and provides acceptable results.

However, for an exact comparison of the performance of these two systems, it would be better to use a large dataset as well as a system with GPU so that the benchmark results can be more reliable and accurate.

8 References

[1] "AutoML.org", url = "https://www.automl.org/"

[2] Frank Hutter et al., "Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms", https://doi.org/10.48550/arXiv.1208.3719

[3]Ruiqi Zheng et al., "AutoML for Deep Recommender Systems: A Survey", https://doi.org/10.48550/arXiv.2204.01390

[4]Zhen Xu et al., "AutoML Meets Time Series Regression, Design and Analysis of the AutoSeries Challenge", https://link.springer.com/chapter/10.1007/978-3-030-86517-7$_3$

[5]Chaoyu Guan et al., "AutoGL: A Library for Automated Graph Learning", https://openreview.net/forum?id=0yHwpLeInDn

[6]Rizka Purwanto et al., "Man versus Machine: AutoML and Human Experts' Role in Phishing Detection", https://arxiv.org/abs/2108.12193

[7]Marc Hanussek et al., "Can AutoML outperform humans? An evaluation on popular, OpenML datasets using AutoML Benchmark", https://arxiv.org/abs/2009.01564

[8]Adithya Balaji, "Benchmarking Automatic Machine Learning Frameworks", https://arxiv.org/pdf/1808.06492

[9]Pieter Gijsbers et al., "An Open Source AutoML Benchmark", https://doi.org/10.48550/arXiv.1907.00909

[10]Anh Truong et al., "Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools", https://doi.org/10.48550/arXiv.1908.05557

[11]Mitar Milutinovic et al., "On Evaluation of AutoML Systems", https://www.automl.org/wp-content/uploads/2020/07/AutoML$_2$020$_p$*aper*$_5$9.*pdf*

[12]Matthias Blohm et al., "Leveraging Automated Machine learning for text classification: Evaluation of AutoML tools and comparison with human performance", https://doi.org/10.48550/arXiv.2012.03575

[13]Luis Ferreira et al.,"A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost", DOI: 10.1109/IJCNN52387.2021.9534091

[14]Marc Zöller et al., Benchmark and Survey of Automated Machine Learning Frameworks, https://doi.org/10.48550/arXiv.1904.12054

[15]Jiancheng Yang et al, "MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis", https://jiancheng-yang.com/assets/publication/MedMNIST