

# AutoML: Automating the Machine Learning Pipeline

Ms. Aayushi Jain<sup>1\*</sup>

<sup>1</sup>Assistant Professor, Yashwantrao Chavan College of Engineering, Nagpur.

---

## Abstract

Automation of the machine learning model-building process is the aim of the emerging field of "Automated Machine Learning, or AutoML". By automating as much of the repetitive, unproductive labour that arises when machine learning is used, autoML was developed to boost productivity and efficiency. Provide a thorough overview for machine learning researchers and practitioners, as well as a foundation for future advancements in AutoML. As a consequence, the method seems to be more effective in terms of operation time and memory use, while maintaining a similar level of output quality. However, based on the estimates we made; we can say that our technique has a great deal of potential for providing an excellent automated ML pipeline design solution. As a result, our next goal will be to make it as competitive as feasible.

*Keywords: AutoML (Automated Machine Learning), Machine learning, Algorithms, Meta-learning, Neural Architecture Search (NAS).*

---

## 1 Introduction

Machine learning is often understood as the branch of AI that focusses on teaching computers new skills without having to modify their code. The need for the capability that machine learning methods offer is rapidly expanding, and effective applications of this strategy are being discovered in a growing range of scientific, technological, and societal domains. In recent years, automated machine learning (AutoML) solutions have been created to address the difficulties encountered by individuals who lack machine learning competence yet sometimes need to utilise it to solve certain issues. Additionally, this technology

---

\* ISBN No. - 978-81-974433-9-8

aims to do away with the need of searching for algorithms—their hyper parameters, for example—by sheer force [1].

It is regarded as an optimisation issue in the majority of current methods, which increases design complexity. However, another strategy known as "meta-learning" involves gathering data on how well certain algorithms perform in related situations. This method has the following drawback, though: it reduces the total complexity of pipeline search and optimisation. Meta-learning efficiency in the automation of ML pipeline design relies on the volume of task-related data kept in a given system. In order to address this problem, it seems promising to compile information on the efficiency of pipelines in particular situations that is retrieved by separate systems into a shared knowledge base. This base may then be used as a meta-learning model to forecast the optimal pipeline for any new jobs [2].

From initial integration to pipeline design, data collecting, optimisation, solution deployment, and maintenance, the purpose of automated machine learning (AutoML) is to automate every step of data analysis. The pipeline composition problem has been the subject of much prior study; however, the majority of these solutions have the drawback of producing more false positives than real ones. There are two primary methods for handling this problem. The pipeline structures in the first class of techniques employ fixed templates, which specify the sequence in which the pipeline components are performed. Preset structures may reduce the number of incorrect pipelines during the composition and optimisation process, but they also make it impossible to investigate potentially intriguing pipelines with structures different from those listed in the templates. There have been several efforts in the second class of techniques to generate particular ML pipeline architectures based on limitations, hence reducing the unpredictability of pipeline building. AI planning or context-free grammars may be used to simulate these constraints and guide pipeline construction [3].

## 2 Literature Review

(Salehin et al., 2024) [4] Provide a thorough description and a summary of the data processing needs for AutoML techniques in this semantic review study. We give "neural architecture search (NAS)" greater attention since it is now one of the most discussed subtopics in the AutoML arena. NAS techniques explore through a wide range of potential topologies using machine learning algorithms to identify the optimal one for a particular job. We provide a review of the outcomes obtained by common NAS algorithms on well-known benchmark datasets, including ImageNet, CIFAR-10, and CIFAR-100. We also explore some notable research prospects in one-shot NAS, one-two-stage NAS, and coupled hyper parameter with architecture optimisation in NAS approaches. We spoke about how the particular issue being addressed might affect the size and complexity of the search space in NAS. Finally, we evaluate many unresolved problems (SOTA problems) in the latest AutoML algorithms that ensure more research in the future.

(Etikani et al., 2024) [5] The purpose of the article is to examine machine learning process automation using cloud-based pipelines. The research also addresses how these kinds of systems get beyond basic challenges related to machine learning processes, such inefficiency, scalability limitations, and

collaboration convolutions with other systems of a similar kind. By making use of distributed computing and storage, cloud-based pipelines simplify every step of machine learning, from processing data to deploying models. The report lists benefits such improved staff integration, resource management, and process organisation that is more efficient. Methods on service deployment and maintenance, large-scale model development and training, and automated data pipeline generation have all been studied. Principal results indicate that using this framework shortens the time needed to complete machine learning tasks and improves the calibre of the models created, not to mention making experiment replication easier.

(Baratchi et al., 2024) [6] In order to produce models with optimal performance, this is accomplished by automatically recognising and addressing every design decision made when building a machine-learning model. We provide a thorough summary of AutoML's history, current state, and future prospects in this post. First, we introduce AutoML, clearly state the problems it aims to tackle, and describe the three core components of AutoML approaches: search space, performance evaluation, and search strategy. The hyperparameter optimisation (HPO) techniques that are often used in the design of AutoML systems are then discussed. In conclusion, we provide an overview of neural architecture search, a particular use case of AutoML that automatically creates deep learning models. We go over and contrast the current AutoML solutions. Finally, we offer a list of unresolved issues and possible directions for further research. All things considered, we provide a thorough review of machine learning for scholars and practitioners and lay the groundwork for future advancements in AutoML..

(Sravani, 2022) [7] Since machine learning is being used more and more in a variety of fields, it is more crucial than ever to have an efficient pipeline for building models. Nevertheless, the process of creating and honing models is still mostly conventional, requiring the assistance of domain experts and requiring laborious data manipulation procedures. This hinders the advancement of machine learning models in both academic and industrial settings. This requirement promotes AutoML, the new research age focused on completely autonomously fitting “machine learning models”. Automating this whole model development process without the need for outside assistance is the aim of AutoML. First, we provide some AutoML observations. Second, we analyse each AutoML pipeline component and provide a short overview of their workings. Furthermore, highlighting AutoML's value in a business context, we offer an instance study on its use and impact in academia. Finally, we wrap off by discussing open research questions and potential future study areas.

(Aliev & Baimuratov, 2022) [8] The issue of automating machine learning (ML) pipeline design is examined in this article. Techniques for automating the ML pipeline design were examined. An ontology seems to have promise in resolving the aforementioned issue, according to the study completed. An ontology engineering-based approach to automate the design of machine learning pipelines was put forward. To build pipelines, an ML ontology was developed. Based on the provided ontology, a pipeline construction application was generated automatically. The suggested solution's efficacy has been empirically evaluated in comparison to one of the most advanced automated pipelines building tools, TPOT. In addition to seeming more effective in terms of yielding high-quality results in the shortest amount of time, the proposed approach is also equivalent to the aforementioned tool at all running times.

(Nguyen et al., 2021) [3] Finding the most promising machine learning pipelines (ML) with the time, CPU, and memory resources at hand is the aim of automated ML composition and optimisation. These approaches sometimes take a long time to optimise and compose pipelines, which limits their ability to explore complicated pipelines in search of more accurate prediction models. Provide an innovative approach that uses an AVATAR surrogate model to assess the correctness of machine learning pipelines without actually running them. By automatically picking up on the features of datasets and the impact that machine learning algorithms have on them, the AVATAR creates a knowledge base. This knowledge base is used by developing a simplified transition from an original machine learning workflow to a surrogate model, a Petri net-based pipeline. Instead of evaluating the validity of the machine learning pipeline itself, the AVATAR evaluates a surrogate model that is constructed utilising input/output simplified mappings and the capabilities and impacts of the pipeline's various components. It takes fewer resources to evaluate this surrogate model than it does to run the original pipeline. Therefore, the AVATAR enables the pipeline composition and optimisation approaches to analyse more pipelines by quickly rejecting invalid pipelines. The AVATAR is included into "the sequential model-based algorithm (SMAC) configuration". Our tests demonstrate that SMAC discovers better answers with AVATAR than it does when it works alone. This is because the AVATAR reduces the time, effort, and resources needed to evaluate pipelines while increasing their capacity.

(Yang et al., 2021) [9] Provide a collection of ten MedMNIST open, pre-processed medical datasets. With no prior information needed, MedMNIST is standardised to carry out classification jobs on lightweight 28 x 28 pictures. It covers the most common types of medical picture data and ranges from 100 to 100,000 records, depending on the job (binary/multi-class, ordinal regression, or multi-label). MedMNIST may be used to medical image analysis using AutoML, multi-modal machine learning, fast prototyping, and teaching. MedMNIST Classification Decathlon also aims to benchmark AutoML methods on all 10 datasets. To this end, we have evaluated a number of baseline techniques, such as paid or open-source AutoML products.

(Gijsbers et al., 2019) [2] In recent years, automated machine learning, or AutoML, has become a thriving field of research. Regretfully, comparing several AutoML systems is challenging and often done poorly. Provide an open, continuous, and expandable benchmark system that adheres to industry best practices and steers clear of typical pitfalls. Utilising public datasets, the framework is open-source and has a webpage with the most recent findings. We apply the paradigm to a comprehensive comparative study of 4 AutoML systems across 39 datasets and interpret the findings.

(Liang et al., 2019) [10] This work advances the use of AutoML. It presents an evolutionary AutoML system called LEAF, which optimises network designs and network size in addition to hyperparameters. LEAF leverages both cutting-edge distributed computing frameworks and evolutionary algorithms (EAs). Results from experiments on natural language processing and medical picture classification demonstrate that the framework may be utilised to attain state-of-the-art performance. Specifically, LEAF shows that architectural optimisation outperforms hyper parameter optimisation and that networks

may be minimised simultaneously with no performance loss. Thus, LEAF lays the groundwork for democratising, enhancing, and practicalizing AI for use in future applications.

### **3 Conclusion**

Making high-performance machine learning techniques accessible to a diverse user base is the aim of the nascent discipline of AutoML. Methods on service deployment and maintenance, large-scale model development and training, and automated data pipeline generation have all been studied. Principal results indicate that using this framework shortens the time needed to complete machine learning tasks and improves the calibre of the models created, not to mention making experiment replication easier. As a consequence, the method seems to be more effective in terms of operation time and memory use, while maintaining a similar level of output quality. However, based on the estimates we made; we can say that our technique has a great deal of potential for providing an excellent automated ML pipeline design solution. As a result, our next goal will be to make it as competitive as feasible.

### **Reference**

- [1] A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, and R. Farivar, “Towards Automated Machine Learning : Evaluation and Comparison of AutoML Approaches and Tools,” no. 2017, pp. 1471–1479, 2019, doi: 10.1109/ICTAI.2019.00209.
- [2] P. Gijssbers, E. Ledell, J. Thomas, S. Poirier, B. Bischl, and J. Vanschoren, “An Open Source AutoML Benchmark,” pp. 1–8, 2019.
- [3] T. Nguyen, K. Musial, and B. Gabrys, “AutoWeka4MCPS-AVATAR : Accelerating automated machine learning pipeline composition and optimisation,” *Expert Syst. Appl.*, vol. 185, no. April, p. 115643, 2021, doi: 10.1016/j.eswa.2021.115643.
- [4] I. Salehin, S. Islam, P. Saha, S. M. Noman, and A. Tuni, “AutoML : A systematic review on automated machine learning with neural architecture search,” *J. Inf. Intell.*, vol. 2, no. 1, pp. 52–81, 2024, doi: 10.1016/j.jiixd.2023.10.002.
- [5] P. Etikani, V. Venkata, S. Rama, S. Nuguri, R. Saoji, and K. Shiva, “Automating Machine Learning Workflows with Cloud-Based Pipelines,” vol. 11, no. 1, pp. 375–382, 2024.
- [6] M. Baratchi, C. Wang, S. Limmer, J. N. Van Rijn, and H. Hoos, *Automated machine learning : past , present and future*, vol. 57, no. 5. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10726-1.
- [7] N. Sravani, “AUTOMATED MACHINE LEARNING THE NEW WAVE OF MACHINE LEARNING,” vol. 13, no. 07, pp. 17–21, 2022.
- [8] M. R. Aliev and I. R. Baimuratov, “Automation of Machine Learning Pipeline Design by an Ontology as an Integrative Meta-Learning Model,” no. December 2021, 2022.

Ms. Aayushi Jain

- [9] J. Yang, R. Shi, and B. Ni, “MEDMNIST CLASSIFICATION DECATHLON: A LIGHTWEIGHT AUTOML BENCHMARK FOR MEDICAL IMAGE ANALYSIS,” pp. 191–195, 2021.
- [10] J. Liang, E. Meyerson, B. Hodjat, D. Fink, K. Mutch, and R. Miikkulainen, “Evolutionary Neural AutoML for Deep Learning,” pp. 401–409, 2019.