Eight years of AutoML: categorisation, review and trends
– Review Protocol –

Rafael Barbudo[1,2], Sebastián Ventura[1,2], and José Raúl Romero[1,2]

[1]*Department of Computer Science and Numerical Analysis, University of Córdoba*

[2]*Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), Córdoba, Spain*

April 2023

Abstract

Knowledge extraction through machine learning techniques has been successfully applied in a large number of application domains. However, apart from the required technical knowledge and background in the application domain, it usually involves a number of time-consuming and repetitive steps. Automated machine learning (AutoML) emerged in 2014 as an attempt to mitigate these issues, making machine learning methods more practicable to both data scientists and domain experts. AutoML is a broad area encompassing a wide range of approaches aimed at addressing a diversity of tasks over the different phases of the knowledge discovery process being automated with specific techniques. To provide a big picture of the whole area, we have conducted a systematic literature review based on a proposed taxonomy that permits categorising 447 primary studies selected from a search of 31,048 papers. This review performs an extensive and rigorous analysis of the AutoML field, scrutinising how the primary studies have addressed the dimensions of the taxonomy, and identifying any gaps that remain unexplored as well as potential future trends. The analysis of these studies has yielded some intriguing findings. For instance, we have observed a significant growth in the number of publications since 2018. Additionally, it is noteworthy that the algorithm selection problem has gradually been superseded by the challenge of workflow composition, which automates more than one phase of the knowledge discovery process simultaneously. Of all the tasks in AutoML, the growth of neural architecture search is particularly noticeable.

1 Introduction

The extraction of useful and novel knowledge from raw data is a challenging process that requires an strong background in multiple research fields and involves different phases. Cer-

tain phases are inherently human –thus requiring from the data scientist experience and know-how– like the interpretation of the mined patterns. In contrast, others are repetitive and time consuming –thus being susceptible of being automated– like the generation of machine learning (ML) models. In this context, the term Automated Machine Learning [4] (AutoML) was recently coined to encompass those approaches that are committed to precisely automate these repetitive phases, allowing data scientists to shift the focus to those phases requiring from their know-how and intuition and, ultimately, bringing the knowledge discovery process closer to domain experts.

This document provides the supplementary material for the paper "*Eight years of AutoML: categorisation, review and trends*" to ensure its replicability. More precisely, the followed review protocol is described in detail, including both the literature search and selection strategy, and the data collection process. Regarding the literature search and selection strategy, it shows the search strings, the literature sources and the inclusion/exclusion criteria used to locate and filter the candidate papers. As for the data collection process, information about the fields considered during the review of the primary studies is provided.

2 Formulation of research questions

The following research questions (RQ) have been formulated:

RQ1: From the primary studies, is it possible to extract a common terminology in the field of AutoML?

RQ2: In quantitative terms, how does the research in AutoML evolve?

RQ3: Which are the phases of the knowledge discovery process covered by the different tasks performed in AutoML and what different techniques are applied?

RQ4: What are the emerging trends and open gaps for future research?

3 Selection of primary studies

The selection of primary studies involves both an automatic and a manual literature search process. For the automatic literature search, three search string are defined (Section 3.1) to query a number of digital libraries and citation databases (Section 3.2). Then a set of exclusion and inclusion criteria are applied to select the relevant primary studies (Section 3.3). As for the manual search, a snowballing procedure is carried out to retrieve any manuscript that could be disregarded during the automatic search (Section 3.4).

3.1 Search strategy

List of possible terms. Terms have been extracted from the title and keywords of well-known papers in the area, and complemented with synonyms and other similar words. In addition, the *Knowledge Discovery in Databases* (KDD) [3], the *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) [2] and the *Sample, Explore, Modify, Model, Assess* (SEMMA) [1] processes have been considered for deriving such terms.

Search strings. For automated search, and having considered the aforementioned terms, the following three search strings have been constructed to conduct the queries. Notice that within the manuscript content, the title, abstract and keywords are queried. In addition, these search strings have been adapted depending on the database-specific query format (see Appendix A).

First search string:

> ("algorithm selection" OR "algorithm recommendation" OR ((hyperparameter OR "hyper parameter") AND (optimization OR tuning)) OR (auto* AND (workflow OR pipeline) AND (planning OR design OR generation OR composition))) AND "machine learning"

Second search string:

> "neural architectur* search" OR (auto* AND ("neural network" OR "deep learning") AND (design OR search OR evol*) AND architectur*)

Third search string:

> auto* AND "machine learning" AND ((data AND (preprocessing OR "pre-processing" OR integration OR clean* OR transformation)) OR (feature AND (engineering OR selection OR extraction)) OR postprocessing OR "post-processing" OR (knowledge AND (filtering OR integration)) OR ((pattern OR model) AND (interpretation OR explanation)))

3.2 Databases and data sources for automated search

The automatic literature search has been carried out by using the aforementioned search strings to query both digital libraries and citation databases.

3.2.1 Digital libraries

ACM LIbrary: `http://dl.acm.org/`
IEEE Xplore: `https://ieeexplore.ieee.org/`
ScienceDirect: `https://www.sciencedirect.com/`
SpringerLink: `https://link.springer.com/`

3.2.2 Citation databases

Scopus: `https://www.scopus.com/`
ISI Web of Knowledge: `https://www.webofknowledge.com/`

3.3 Study selection process

The candidate papers retrieved by the search strings are conveniently filtered according to a number of inclusion and exclusion criteria.

3.3.1 Exclusion criteria

The following exclusion criteria are used to select the candidate papers that will be considered as primary studies:

1. Papers prior to the appearance of the term AutoML, i.e., 2014, are removed.
2. Papers that are not written in English are removed.
3. Those publications for which there is no clear evidence of a blinded, peer-review process are removed.
4. Papers that are not published in conferences ranked A*–A according to the CORE ranking system are removed.
5. Papers published in journals that are not indexed in the JCR are removed.

3.3.2 Inclusion criteria

The titles and abstracts of candidate papers are scrutinised to ensure that they meet the following inclusion criteria:

1. The automation is defined for a phase of the knowledge discovery process.
2. The contribution of the paper is focused on making a real progress in the field of AutoML, not as much on the specific artificial intelligence methods used.
3. The proposal is independent of a specific dataset.
4. Theoretical proposals are allowed.

3.4 Data sources for manual search

For the manual literature search, a snowballing procedure has been conducted over the primary studies. Here, potential candidates are subjected to same inclusion/exclusion criteria that the papers obtained during the automatic literature search. Therefore, manuscripts not belonging to a peer-review source are discarded. However, technical reports and additional material will be collected from publicly available sources like `https://arxiv.org`, if these materials are properly cited in a primary study. These supplementary materials will not be cited in the review, although data extraction forms will provide links to them.

4 Data extraction form

Table 1: Data extraction form.

CATEGORY	DESCRIPTION
Automated phases	
List of phases	It lists the phases of the knowledge discovery process being automated.
Phase(s) detail	It further decomposes the phases. As an example, data reduction can be accomplished by both feature extraction or selection methods.
# of automated phases	It indicates the maximum number of phases that are jointly automated.
Accomplished tasks	
List of tasks	It lists the tasks that are conducted to perform the automation.
Task relationships	It specifies if these tasks are conducted together, in sequence or independently.
Applied techniques	
List of techniques	It lists the techniques being applied to conduct the tasks.
Technique(s) details	When more than one task and technique are considered, it specifies which particular technique is responsible for performing each task.
Experimental framework	
Type of study	It identifies the type of experimental study being conducted, if any, and differentiates between an empirical study and a sample execution.
Case study	It denotes the type of data used during the validation depending on its source: synthetic data, from literature or from an industry.
Comparison framework	It indicates if the proposal is compared against other AutoML approaches, machine learning methods or some different version of the same approach.
Statistical analysis	It indicates whether some type of statistical analysis is performed using pairwise tests, multiple tests or effect size comparisons.
Additional material	
Source code	It indicates whether the source code of the proposal is available.
Sfw. guides & manuals	It indicates whether there are any document or instructions on how to install and/or execute the software artefacts.
User interface	It indicates whether a command-line or a graphical user interface is available.
Annexes or reports	It determines whether there is some extension of the manuscript in terms of annexes, technical reports or others.
Experimental results	It indicates whether raw data and outputs from the experimental validation are provided.
Datasets	If datasets are explicitly constructed for or used in the primary study, it specifies whether they are available or not.
Others	It indicates the existence of any other type of material.

For each primary study under review, a number of fields will be collected during the data collection process. More precisely, each manuscript will be scrutinised to gather information about the phases of the knowledge discovery process being automated (Section 4.1), the

accomplished tasks (Section 4.2), the applied techniques (Section 4.3), the experimental framework (Section 4.4) and the availability of supplementary material (Section 4.5). As shown in Table 1, different elements of information are gathered for each category. It should be noted that authors have not been contacted to guarantee objectiveness and fairness, thus missing information will be reflected in the paper but not considered.

4.1 Automated phases

List of phases. It compiles the phases of the knowledge discovery process that are automated. The following phases are considered:

- Domain understanding
- Dataset creation
- Data cleaning and preprocessing
- Data reduction and projection
- Data mining
- Knowledge interpretation
- Knowledge integration

Phase(s) details. It is a text field that gives more details about the targets of the automation, i.e. the phases of the knowledge discovery process, according to the category above. As an example, data mining phase can be further decomposed into classification and clustering.

Number of automated phases. It allows to distinguish between those approaches that jointly optimise more than a phase, e.g., workflow composition, and those doing it in an isolated way. In case that only a phase is automated, "Not Applicable" is assigned. Otherwise, the number of jointly automated phases is indicated.

4.2 Accomplished tasks

List of tasks. It compiles the tasks that are responsible of conducting the automation. The following tasks are considered:

- Algorithm selection: It may select a single ("*single*"), an unordered set ("*set*") or a ranking ("*ranking*") of algorithms.
- Hyper-parameter optimisation: We differentiate whether the approach is independent of the algorithm being tuned ("*independent*") or not ("*dependent*"). Neural architecture search ("nas") is an illustrative example of the second group.
- Workflow composition: Some approaches fix the structure of the resulting workflows ("*fixed*"), while others optimises their structure too ("*variable*").

- Supporting task: It is related to those tasks supporting a primary one, i.e. algorithm selection, hyper-parameter optimisation or workflow composition.
- Ad hoc tasks: It refers to those tasks that are specially designed to conduct the automation of a certain phase, generally a pre- or postprocessing one.

Task relationship. It indicates how the accomplished tasks are related when a proposal conduct two or more of them. In case that only a task is conducted, "Not Applicable" is assigned. We consider the following scenarios:

- Together: Tasks are carried out at the same time.
- In sequence: A task require from another to be finished.
- Independent: Tasks do not depend on any other one to start.

4.3 Applied techniques

List of techniques. It compiles the techniques that are applied (e.g. evolutionary algorithm, Bayesian optimisation) to conduct the automation.

Technique(s) details. It is an unrestricted field of text that is used to retrieve any valuable information about the list of techniques. As an example, it is interesting to gather which ML algorithms are used at the meta-level by a meta-learning proposal. It is also used to specify which technique is responsible of conducting each task whenever a primary study considers two or more tasks and techniques.

Main technique(s). It highlights the main technique(s) carrying out the accomplished tasks. It is particularly common to identify as "Main technique" more than one when two or more tasks are performed.

4.4 Experimental framework

Type of study. It determines the scope of the experimental validation that is carried out. The following categories are proposed:

- Theoretical proposal: It indicates that no experiments are conducted.
- Sample execution: It is a proof of concept of the proposed approach, generally applied to solve a toy problem.
- Empirical validation: It considers an in depth validation over more real problems is conducted.

Case study. It aims to figure out the source and the scope of the used datasets during the experiments. The following categories are considered:

- Synthetic: The datasets were artificially generated.
- Controlled environment: The used datasets come from the literature.
- Industrial case: The datasets are provided by an industry partner.

Comparison framework. It analyses the proposals that are used for the comparison. Here, we consider the following types of comparisons:

- AutoML comparison: The comparison is conducted between similar AutoML proposals, generally conducting the same task.
- Baseline comparison: The AutoML proposal is validated against a machine learning method. An example of this would be to compare neural architecture search against a hand-crafted architecture.
- Variant comparison: The comparison is conducted between different configurations of the same proposal and/or preliminary versions of the same approach.

Statistical analysis. It indicates if the finding of the experimental framework are accompanied of a statistical analysis. More precisely we differentiate between the following categories:

- Pairwise comparison: It is applied to identify whether there are differences between two approaches or not (e.g. t-test).
- Multiple comparison: It is applied to identify whether there are differences between multiple approaches or not (e.g. Friedman test).
- Effect size: It measures the magnitude of the differences (e.g. Cliff's delta).

4.5 Additional material

For this category, we analyse the existence of the following elements:

- **Source code** It indicates whether the source code of the proposal is available.
- **Software guides or manual.** It indicates whether there are any document or instructions on how to install and/or execute the software artefacts.
- **User interfaces.** It indicates whether a command-line or a graphical user interface is available.
- **Annexes or reports.** It determines whether there is some extension of the manuscript in terms of annexes, technical reports or others.

- **Experimental results.** It indicates whether raw data and outputs from the experimental validation are provided.
- **Datasets.** If datasets are synthetically constructed for or used in the primary study, it specifies whether they are available or not.
- **Others.** It indicates the existence of any other type of additional material like videos or interactive visualisation tools.

5 Data extraction

Extracting information of the primary studies (Section 5.1) generate a set of documents (Section 5.2) whose outcomes are summarised in a spreadsheet provided together with this document. To avoid conflicts during the overall process, a well-defined methodology has been followed (Section 5.3).

5.1 Procedure

1. Search for papers using the aforementioned search strings over each digital library and citation database.
2. Obtain a unique global list of papers by removing duplicated entries and irrelevant manuscripts (proceedings, special issues).
3. Apply inclusion and exclusion criteria.
4. Review each primary study to fill the "Data extraction form".
5. Check the paper bibliographies to recover potential missing references in Step 2. If any, apply Step 3 and, if criteria are met, add to the list of selected primary studies. This process is not repeated for those manuscripts obtained during the snowballing procedure.

5.2 Extraction strategy

Every step of the extraction process will be recorded using forms, spreadsheets and BibTex files. Different versions will be conveniently named and the final forms will be made publicly available. The following documents will be created during the process:

- **Search results**: list of papers found per source (see Annex A).
- **Primary studies**: list of papers after applying the inclusion and exclusion criteria.
- **Data extraction forms**: information extracted for each primary study during the data collection process.

5.3 Validation process

1. Before collecting the search results:
    - To reach an agreement regarding the Review protocol [discussion meeting].
    - To check a pilot search to validate the search strings, all reviewers performing this task independently.
2. After collecting the search results, to check that the list of search results is complete, i.e. it includes all the distinct papers returned by each different database [all reviewers].
3. After selecting the primary studies:
    - Having the inclusion and exclusion criteria been evaluated by more than one reviewer, to check inconsistencies and solve disagreements [discussion meeting].
    - To perform data extraction to a random sample of primary studies (up to 3papers) and check that the "data extraction form" is consistently valid, i.e. all the reviewers interpret its fields in the same way.
4. After extracting all the data from primary studies:
    - To extract the relevant information of each paper, and a third reviewer to double-check both forms in order to solve disagreements.
    - To check the final information extracted from each study and to solve any disagreement. If necessary, to meet for discussion.

A Results from search queries

A.1 General information

This appendix summarises the outcomes obtained from querying each database by means of the aforementioned search strings. It also compiles how the search strings have been adapted to each literature source:

- Queries have been conducted between 26 and 29th of December 2018.
- Queries have been revised to include those manuscripts published during 2019, 2020 and 2021:
  - First update: 23^{rd} of April 2019
  - Second update: 10^{th} of January 2020
  - Third update: 20^{th} of January 2021
  - Fourth update: 18^{th} of January 2022
- Query results are exported to BIBTEX or CSV files according to the source limitations: *YYMMDD-Source-SearchString.bib* and *YYMMDD-Source-SearchString.csv*
- Basic information about each search result (author, title, publication, year) has been saved in the corresponding spreadsheet.

A.2 ACM Library

Webpage: http://dl.acm.org/

Field: Abstract

Notes: The queries have been adapted according to the following observations:

- The search has been restricted to "The ACM Guide to Computing Literature".
- Special characters (*) are not supported within exact matches ("").
- This database does not support more than one field (abstract, title, keywords).
- It discerns between British and American English (e.g. optimization ≠ optimisation).

Search string transformation:

- #1: optimization → optimization OR optimisation
- #2: "neural architectur* search" → "neural architecture search"

ID	Query	# Papers
1	recordAbstract:((“algorithm selection” OR “algorithm recommendation” OR ((hyperparameter OR “hyper parameter”) AND (optimization OR optimisation OR tuning)) OR (auto* AND (workflow OR pipeline) AND (planning OR design OR generation OR composition))) AND “machine learning”)	389
2	recordAbstract:(“neural architecture search” OR (auto* AND (“neural network” OR “deep learning”) AND (design OR search OR evol*) AND architectur*))	547
3	recordAbstract:(auto* AND “machine learning” AND ((data AND (preprocessing OR “pre-processing” OR integration OR clean* OR transformation)) OR (feature AND (engineering OR selection OR extraction)) OR postprocessing OR “post-processing” OR (knowledge AND (filtering OR integration)) OR ((pattern OR model) AND (interpretation OR explanation))))	1512
	Total (distinct papers)	2327

Table 2: Search results from ACM Library

A.3 IEEE Xplore

Webpage: `https://ieeexplore.ieee.org/`

Field: Metadata only

Notes: Information from the webpage:

- Command search is conducted.
- Only manuscript metadata is queried.
- There is a maximum of 40 search terms

Search string transformation: No string transformations were required.

A.4 ISI Web of Knowledge

Webpage: `https://www.webofknowledge.com/`

Field: Topic (TS)

Notes: Information from the webpage:

- Field Tags used: Topic (TS), Research Area (SU).
- Topic Tag includes: Title, Abstract and Keywords.

Search string transformation: No string transformations were required.

ID	Query	# Papers
1	("algorithm selection" OR "algorithm recommendation" OR ((hyperparameter OR "hyper parameter") AND (optimization OR tuning)) OR (auto* AND (workflow OR pipeline) AND (planning OR design OR generation OR composition))) AND "machine learning"	816
2	"neural architectur* search" OR (auto* AND ("neural network" OR "deep learning") AND (design OR search OR evol*) AND architectur*)	2083
3	auto* AND "machine learning" AND ((data AND (preprocessing OR "pre-processing" OR integration OR clean* OR transformation)) OR (feature AND (engineering OR selection OR extraction)) OR postprocessing OR "post-processing" OR (knowledge AND (filtering OR integration)) OR ((pattern OR model) AND (interpretation OR explanation)))	8501
	Total (distinct papers)	10988

Table 3: Search results from IEEE Xplore

ID	Query	# Papers
1	TS=(("algorithm selection" OR "algorithm recommendation" OR ((hyperparameter OR "hyper parameter") AND (optimization OR tuning)) OR (auto* AND (workflow OR pipeline) AND (planning OR design OR generation OR composition))) AND "machine learning") AND SU=Computer Science	563
2	TS=("neural architectur* search" OR (auto* AND ("neural network" OR "deep learning") AND (design OR search OR evol*) AND architectur*)) AND SU=Computer Science	1079
3	TS=(auto* AND "machine learning" AND ((data AND (preprocessing OR "pre-processing" OR integration OR clean* OR transformation)) OR (feature AND (engineering OR selection OR extraction)) OR postprocessing OR "post-processing" OR (knowledge AND (filtering OR integration)) OR ((pattern OR model) AND (interpretation OR explanation)))) AND SU=Computer Science	2335
	Total (distinct papers)	3720

Table 4: Search results from ISI Web of Knowledge

A.5 ScienceDirect

Webpage: https://www.sciencedirect.com/

Field: Article Title, Abstract, Keywords

Notes: The queries have been adapted according to the following observations:

- Special characters (*) are not supported.
- Advanced search does not support more than 8 Boolean operators.

Search string transformation:

- #1-3: auto* → automated
- #2: evol* → evolution
- #2: architectur* → architecture
- #3: clean* → cleaning
- Search strings #1 and #3 were split due to the restriction of not using more than 8 Boolean operators.

ID	Query	# Papers
1a	("algorithm selection" OR "algorithm recommendation" OR ((hyperparameter OR "hyper parameter") AND (optimization OR tuning))) AND "machine learning"	193
1b	(automated AND (workflow OR pipeline) AND (planning OR design OR generation OR composition)) AND "machine learning"	29
2	"neural architecture search" OR (automated AND ("neural network" OR "deep learning") AND (design OR search OR evolution) AND architecture)	135
3a	automated AND "machine learning" AND (data AND (preprocessing OR "pre-processing" OR integration OR cleaning OR transformation))	117
3b	automated AND "machine learning" AND (feature AND (engineering OR selection OR extraction))	225
3c	automated AND "machine learning" AND (postprocessing OR "post-processing" OR (knowledge AND (filtering OR integration)))	41
3d	automated AND "machine learning" AND ((pattern OR model) AND (interpretation OR explanation))	40
	Total (distinct papers)	706

Table 5: Search results from ScienceDirect

A.6 Scopus

Webpage: `https://www.scopus.com/`

Field: Article Title, Abstract, Keywords

Notes: The queries have been adapted according to the following observations:

- Scopus allowed to restrict the subject area (SUBJAREA) to computer science (COMP).
- Year range (PUBYEAR) and document type (DOCTYPE) restriction are embedded within the search strings.

Search string transformation: No string transformations were required.

A.7 SpringerLink

Webpage: `https://link.springer.com/`

Field: All (Advanced search)

Notes: The queries have been adapted according to the following observations:

- The advanced search does not allow to specify the search fields.
- It discerns between British and American English (e.g. optimisation $\neq$ optimization).
- The original search strings reported and intractable number of manuscripts. Thus, the NEAR operator was used

Search string transformation:

- #1: optimization $\rightarrow$ optimization OR optimisation
- #1-3: AND $\rightarrow$ NEAR. The AND operator has been replaced in those situation where the joined terms are expected to appear close (e.g. *feature* and *selection*).

ID	Query	# Papers
1	TITLE-ABS-KEY((“algorithm selection” OR “algorithm recommendation” OR ((hyperparameter OR “hyper parameter”) AND (optimization OR tuning)) OR (auto* AND (workflow OR pipeline) AND (planning OR design OR generation OR composition))) AND “machine learning”) AND (LIMIT-TO(PUBYEAR,2020) OR LIMIT-TO(PUBYEAR,2019) OR LIMIT-TO(PUBYEAR,2018) OR LIMIT-TO(PUBYEAR,2017) OR LIMIT-TO(PUBYEAR,2016) OR LIMIT-TO(PUBYEAR,2015) OR LIMIT-TO(PUBYEAR,2014)) AND (LIMIT-TO(DOCTYPE,“cp”) OR LIMIT-TO(DOCTYPE,“ar”) OR LIMIT-TO(DOCTYPE,“cr”) OR LIMIT-TO(DOCTYPE,“ip”) OR LIMIT-TO(DOCTYPE,“re”) OR LIMIT-TO(DOCTYPE,“sh”)) AND (LIMIT-TO(SUBJAREA,“COMP”)) AND (LIMIT-TO(LANGUAGE,“English”))	1455
2	TITLE-ABS-KEY(“neural architectur* search” OR (auto* AND (“neural network” OR “deep learning”) AND (design OR search OR evol*) AND architectur*)) AND (LIMIT-TO(PUBYEAR,2020) OR LIMIT-TO(PUBYEAR,2019) OR LIMIT-TO(PUBYEAR,2018) OR LIMIT-TO(PUBYEAR,2017) OR LIMIT-TO(PUBYEAR,2016) OR LIMIT-TO(PUBYEAR,2015) OR LIMIT-TO(PUBYEAR,2014)) AND (LIMIT-TO(DOCTYPE,“cp”) OR LIMIT-TO(DOCTYPE,“ar”) OR LIMIT-TO(DOCTYPE,“re”) OR LIMIT-TO(DOCTYPE,“ip”)) AND (LIMIT-TO(SUBJAREA,“COMP”)) AND (LIMIT-TO(LANGUAGE,“English”))	2482
3	TITLE-ABS-KEY(auto* AND “machine learning” AND ((data AND (preprocessing OR “pre-processing” OR integration OR clean* OR transformation)) OR (feature AND (engineering OR selection OR extraction)) OR postprocessing OR “post-processing” OR (knowledge AND (filtering OR integration)) OR ((pattern OR model) AND (interpretation OR explanation)))) AND (LIMIT-TO(PUBYEAR,2020) OR LIMIT-TO(PUBYEAR,2019) OR LIMIT-TO(PUBYEAR,2018) OR LIMIT-TO(PUBYEAR,2017) OR LIMIT-TO(PUBYEAR,2016) OR LIMIT-TO(PUBYEAR,2015) OR LIMIT-TO(PUBYEAR,2014)) AND (LIMIT-TO(DOCTYPE,“ar”) OR LIMIT-TO(DOCTYPE,“cp”) OR LIMIT-TO(DOCTYPE,“ip”) OR LIMIT-TO(DOCTYPE,“re”)) AND (LIMIT-TO(SUBJAREA,“COMP”)) AND (LIMIT-TO(LANGUAGE,“English”))	6657
	Total (distinct papers)	10202

Table 6: Search results from Scopus

ID	Query	# Papers
1	"machine learning" AND ("algorithm selection" OR "algorithm recommendation" OR ((hyperparameter OR "hyper parameter") NEAR (optimization OR optimisation OR tuning)) OR (auto* NEAR (workflow OR pipeline) NEAR (planning OR design OR generation OR composition)))	2088
2	"neural architectur* search" OR (auto* NEAR ("neural network" OR "deep learning") NEAR (design OR search OR evol*) NEAR architectur*)	488
3	"machine learning" AND auto* NEAR ((data NEAR (preprocessing OR "pre-processing" OR integration OR clean* OR transformation)) OR (feature NEAR (engineering OR selection OR extraction)) OR postprocessing OR "post-processing" OR (knowledge NEAR (filtering OR integration)) OR ((pattern OR model) NEAR (interpretation OR explanation)))	672
	Total (distinct papers)	3048

Table 7: Search results from SpringerLink

References

[1] Azevedo AIRL, Santos MF (2008) Kdd, semma and crisp-dm: a parallel overview. IADS-DM

[2] Chapman P, Clinton J, Kerber R, et al (2000) Crisp-dm 1.0: Step-by-step data mining guide. SPSS inc 16

[3] Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, Menlo Park, CA, USA, chap From Data Mining to Knowledge Discovery: An Overview, p 1–34

[4] Hutter F, Kotthoff L, Vanschoren J (2019) Automatic machine learning: methods, systems, challenges. Springer, Cham