

A Multivocal Review of AutoML Practices, Challenges, Opportunities and Open Issues

Keerthiga Rajenthiram ✉ 

Vrije Universiteit Amsterdam, The Netherlands

Faezeh Amou Najafabadi ✉ 

Vrije Universiteit Amsterdam, The Netherlands

Ilias Gerostathopoulos ✉ 

Vrije Universiteit Amsterdam, The Netherlands

Patricia Lago ✉ 

Vrije Universiteit Amsterdam, The Netherlands

Abstract

Background. Automated Machine Learning (AutoML) is increasingly embedded in software and data engineering workflows, automating the design and optimization of machine learning (ML) pipelines. AutoML has been extensively synthesized through secondary studies that map tools, algorithms, and evaluation practices. However, these reviews predominantly reflect a research-centric perspective and provide limited evidence on how AutoML is actually used, adapted, and experienced in practice. Traditional surveys and interviews capture only a narrow slice of the practitioner community, leaving open how well academic claims about AutoML align with day-to-day engineering realities.

Aims. This study aims to (i) assess to what extent AutoML practices and challenges reported in existing secondary studies reflect practitioners' experiences, and (ii) identify mismatches between research and practice by empirically comparing academic claims with practitioner experiences shared on YouTube.

Method. Following established guidelines for multivocal literature reviews in software engineering, we conduct a multivocal evidence synthesis combining 16 peer-reviewed secondary studies (systematic and multivocal literature reviews) with 30 practitioner-created YouTube videos in which engineers, data scientists, and ML practitioners describe their AutoML workflows, workarounds, and pain points.

Results. Our analysis reveals a marked imbalance: academic reviews emphasize algorithm selection, search strategies, and benchmarking, whereas practitioners foreground data quality and cleaning, resource and cost constraints, tool robustness, scalability, and integration into production pipelines. We also observe areas of alignment, for example around hyperparameter optimization and the need for reproducible workflows.

Conclusions. Treating YouTube videos as curated gray literature under established multivocal review guidelines, our study shows that they provide complementary, practice-grounded evidence about AutoML adoption. The identified gaps and alignments suggest concrete directions for empirical AutoML research that more directly address operational and organizational constraints in real-world software engineering settings.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases Automated Machine Learning (AutoML), Multivocal Literature Review, YouTube Videos, Gray Literature, Empirical Study, Research–Practice Gap

Digital Object Identifier 10.4230/LIPIcs.ESEM.2026.27

Category Technical Track Paper

Supplementary Material *Other (Online Catalog):* <https://doi.org/10.5281/zenodo.20282909> [1]

Other (Replication Package): <https://doi.org/10.5281/zenodo.20283067> [2]



© Open Access. Public;
licensed under Creative Commons License CC-BY 4.0

20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026).

Editors: Robert Feldt, Maria Paasivaara, Daniel Mendez, Stefan Wagner, and Marvin Muñoz Barón;

Article No. 27; pp. 27:1–27:20



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Machine learning (ML) has become a foundational technology across modern industries, powering applications in healthcare, finance, manufacturing, and digital services [23, 27]. Its success has led to the emergence of AI engineering, a discipline that bridges software engineering and artificial intelligence to design, deploy, and maintain ML-enabled systems [23]. Despite the rapid progress of ML, many organizations still struggle to operationalize ML systems because of their heavy dependence on scarce expert knowledge and the high cost of developing, tuning, and maintaining models [23, 21, 11]. Traditional ML workflows require practitioners to manually handle tasks such as data preprocessing, feature engineering, model selection, and hyperparameter optimization – each demanding extensive domain knowledge and iterative experimentation. This dependency on expert intervention has created a bottleneck in industrial adoption, limiting innovation and scalability in organizations lacking skilled data scientists [23]. Building an effective ML model requires navigating a large design space of algorithms, encodings, and hyperparameters under constraints of accuracy, memory, and runtime; exhaustively exploring this combinatorial search space through manual experimentation is computationally expensive and often yields suboptimal results due to the limitations of human-guided search [5, 23]. These technical and operational pressures collectively strengthen the case for automation in configuration, search, and data preparation processes.

Automated Machine Learning (AutoML) has emerged as a response to these limitations by automating the end-to-end design and optimization of ML pipelines [15, 23, 5]. AutoML systems automate processes such as data preprocessing, feature engineering, algorithm and hyperparameter selection, and performance evaluation to produce optimized models with reduced human intervention [3, 27, 21]. Unlike traditional ML, where experts manually engineer and tune models, AutoML systems algorithmically explore search spaces of models, hyperparameters, and architectures to identify high-performing configurations. This paradigm aims to democratize ML – making it more accessible to non-experts – while improving reproducibility, scalability, and efficiency [11, 14, 33].

In parallel with advances in AutoML technology, the literature has witnessed a growing number of secondary studies that survey AutoML tools, algorithms, and evaluation practices [6, 8, 15, 16]. While these reviews provide a valuable map of the research landscape, they predominantly reflect a research-centric perspective and give limited attention to how AutoML is actually used, adapted, and experienced in real projects [15, 23]. Existing studies rarely capture the fine-grained practices, workarounds, and frustrations that arise when teams integrate AutoML into complex socio-technical environments. Traditional empirical methods such as surveys and interviews, although important, typically reach only a narrow slice of the practitioner community, raising questions about how well academic claims about AutoML align with day-to-day engineering realities [12, 34].

At the same time, many ML engineers, data scientists, and practitioners now openly share their AutoML experiences on social media platforms, notably YouTube, where they discuss implementation challenges, best practices, and lessons learned from applied projects. YouTube is one of the most visited websites worldwide, with more than two billion users and hundreds of hours of video uploaded every hour, and it has increasingly been used as a data source in qualitative and mixed-methods research [32, 19]. Prior work has shown that user-generated YouTube videos can serve as a qualitative research asset and a rich source of naturally occurring data, provided that search strategies, sampling decisions, and contextual limitations are made explicit [9, 30, 17]. These studies highlight both the opportunities

and the methodological challenges of working with YouTube data, for example the need to document search terms, ranking mechanisms, and inclusion criteria to approximate real viewing behavior [30, 32]. Motivated by this body of work, we treat practitioner-created YouTube videos as curated gray literature that can complement traditional sources when studying AutoML in practice.

In software engineering, guidelines for multivocal literature reviews (MLRs) explicitly advocate combining formal literature (e.g., journal and conference papers) with gray literature (e.g., blog posts, white papers, and videos) to obtain a more complete picture of both the state of the art and the state of practice [13]. MLRs extend systematic literature reviews by incorporating such gray sources under a structured protocol for searching, quality assessment, and data extraction [13, 23]. Building on these guidelines, we conduct a multivocal evidence synthesis that integrates findings from peer-reviewed secondary studies on AutoML with practitioner experiences shared on YouTube. This design allows us to empirically examine to what extent AutoML practices and challenges reported in academic reviews are reflected in practitioners' accounts, and where they diverge.

In this study, we conduct a multivocal evidence synthesis that comparatively analyzes AutoML practices and challenges across 16 literature reviews and 30 practitioner-created YouTube videos [13]. This design allows us to systematically contrast research and practice perspectives on AutoML and to derive implications for both communities. Our contributions are threefold: (i) we synthesize AutoML practices and challenges reported in existing secondary studies and in YouTube practitioner content; (ii) we assess where these two sources converge or diverge; and (iii) we derive implications for researchers and practitioners based on the observed gaps and alignments

2 Background and Related Work

Over the last decade, AutoML has evolved from research prototypes to an ecosystem of open-source frameworks (e.g., Auto-Sklearn, TPOT, FLAML) and commercial platforms offered by major providers such as Google, AWS, IBM, and Microsoft [23, 10]. These tools enable users to build, train, and deploy models through graphical interfaces or APIs while supporting scalability across on-premise and cloud infrastructures [10, 33]. Within the ML workflow, AutoML can trigger automatic retraining and monitoring, thereby improving reliability and maintainability [23]. In specialized domains such as medical imaging, AutoML techniques like Neural Architecture Search (NAS), Hyperparameter Optimization (HPO), and Automatic Data Augmentation (ADA) have produced architectures that compete with or exceed manually designed neural networks [3]. Similarly, AutoML has been applied in robotics [27], unsupervised learning [21], and other domain-specific settings, underscoring its expanding applicability.

AutoML has also been the subject of several secondary studies, including systematic literature reviews, mapping studies, and multivocal reviews; a systematic mapping of this body of work is provided in a recent tertiary study [25]. These reviews consolidate terminology regarding the scope and phases of AutoML [6, 8], provide overviews of pipeline structures and key components [15, 16], and evaluate well-known AutoML tools, frameworks, and algorithms through benchmarks and reported experiences [38, 33, 23]. Other reviews focus on specific phases of the AutoML process, such as HPO [18, 24], feature engineering [26, 22], and NAS [28, 29], or on application domains and use cases including healthcare, finance, and industrial IoT [36, 34, 4, 37, 7, 5, 3]. A few empirical studies complement these reviews with interviews and case studies on AutoML adoption and usage [35, 31]. Overall, this literature

offers a broad and deep account of AutoML concepts, techniques, and tool ecosystems, but it largely foregrounds algorithmic innovation and benchmark performance, and provides only limited, fragmented insight into how AutoML is actually practiced in day-to-day engineering work.

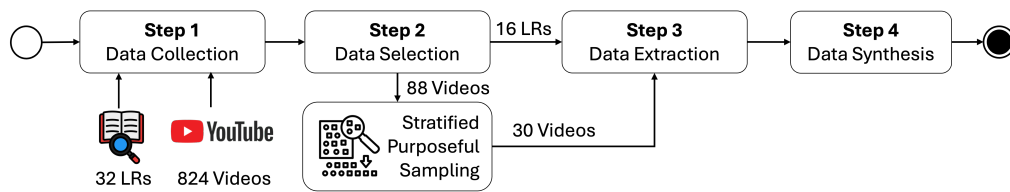
Beyond software engineering, there is a growing body of methodological work on using YouTube as a data source for qualitative and mixed-methods research. Systematic reviews of health-related YouTube studies highlight that, when search strategies, sampling procedures, and inclusion criteria are transparently reported, YouTube can serve as a reproducible and informative medium for studying practices, perceptions, and communication patterns [30, 32]. Other work explicitly positions user-generated YouTube videos as a qualitative research asset and a repository of naturally occurring data that can inform both teaching and research [9]. More broadly, reviews of social-media-based qualitative research document YouTube alongside platforms such as Twitter and Facebook as a common source for content analysis and observational studies, while also emphasizing the need to acknowledge platform-specific biases and engagement mechanisms [17, 19]. These studies collectively argue that YouTube can be used as a legitimate empirical source, provided that researchers carefully design and document their search, sampling, and analysis procedures.

In software engineering, guidelines for multivocal literature reviews explicitly advocate combining formal literature (e.g., journal and conference papers) with gray literature (e.g., blog posts, white papers, practitioner talks, and videos) to obtain a more complete picture of both the state of the art and the state of practice [13]. MLR guidelines specify concrete procedures for identifying, selecting, and assessing gray sources, and they highlight the value of practitioner material for revealing real-world constraints, trade-offs, and work practices that are under-represented in academic publications [13, 23]. So far, only one multivocal literature review has addressed AutoML in depth [23]. However, that study primarily synthesizes practitioner- and academic-oriented sources at a relatively coarse level, and it does not explicitly compare the frequency of specific practices and challenges across source types. To the best of our knowledge, no prior AutoML study has systematically treated practitioner-created YouTube videos as gray literature under these guidelines, nor explicitly contrasted the distribution of practices and challenges between academic secondary studies and YouTube content.

Against this backdrop, our work focuses on the *state of practice* of AutoML as communicated by practitioners on YouTube, and on how this state aligns with or diverges from the AutoML landscape portrayed in academic secondary studies. We build on the existing AutoML review literature and MLR guidelines, but extend them in two key ways: (i) we treat practitioner-created YouTube videos as a curated gray-literature corpus, sampled and analyzed following established recommendations for YouTube research and MLRs; and (ii) we perform a comparative, cross-source analysis that explicitly quantifies and interprets similarities and differences in reported practices and challenges between academic and practitioner sources. In doing so, we provide a more fine-grained and empirically grounded picture of how AutoML is actually used, adapted, and experienced in real-world software and data engineering projects.

3 Study Design

In this study, we conduct a comparative analysis of evidence from 16 literature reviews and 30 practitioner-created YouTube videos. Following a multivocal evidence synthesis approach [13] that bridges white and gray literature, we compare these sources to identify where research assumptions align or diverge from real-world practice.



■ **Figure 1** Overview of the multivocal literature review process, spanning data collection, selection, extraction, and synthesis across LR and YouTube corpora.

3.1 Research Questions

To explore the alignment and divergence between these perspectives, we formulate the following research questions (RQs):

RQ1: What AutoML practices are reported in the existing literature and on YouTube?

Rationale. This question identifies and categorizes practices associated with AutoML across research and practice. It examines how the literature describes methods and strategies for designing, implementing, and optimizing AutoML pipelines, and contrasts these accounts with the hands-on practices reported by practitioners in real-world projects.

RQ2: What challenges are reported in the existing literature and on YouTube?

Rationale. This question identifies and categorizes limitations, barriers, and risks encountered in AutoML development and deployment. It compares how the literature conceptualizes technical, organizational, and ethical challenges with the difficulties and pain points expressed by practitioners.

RQ3: To what extent do practices and challenges discussed in the literature align with those reported on YouTube?

Rationale. Building on RQ1 and RQ2, this question investigates the degree of convergence and divergence between academic and practitioner perspectives. It evaluates whether research priorities reflect the realities encountered in practice and pinpoints areas where theoretical assumptions may overstate, overlook, or underestimate practical concerns in AutoML adoption.

3.2 Research Process

The overall research process – from data collection and selection to data extraction and synthesis is summarized in Figure 1.

3.2.1 Data Collection

This study integrates two complementary sources of evidence: *literature reviews*, which provide theoretically grounded and aggregated knowledge, and *YouTube videos*, which capture practical, experiential, and up-to-date insights from practitioners. Together, these sources offer a broader and more grounded perspective on AutoML by combining white and gray literature [13].

Literature Review Dataset. The literature review dataset was derived from our recent tertiary study on AutoML [25], which synthesized evidence from existing systematic and multivocal literature reviews. In that study, a total of 32 reviews were identified through a

structured search and selection process that treated secondary studies as primary units of analysis. For the present work, we reused this curated dataset and focused on the subset of 16 reviews (more than 400 primary studies) that contained explicit information on AutoML practices and challenges – the key dimensions relevant to our comparative analysis. Reusing this validated dataset ensures methodological consistency with prior work and provides a robust foundation for examining how reported practices and challenges in the literature compare with those discussed by practitioners.

YouTube Videos Dataset. We collected practitioner-created YouTube content in two rounds. The first dataset, retrieved on 7 April 2025, served as a pilot and contained 540 videos. A second search on 16 June 2025 added 284 new videos published in 2025, resulting in a combined pool of 824 videos. Using the YouTube Search API, we queried each search term individually and retrieved up to 100 videos per term listed below.

`["AutoML","Automated Machine Learning","AutoML practices","AutoML challenges","AutoML pipelines","AutoML tools comparison","AutoML automation tools","AutoML case studies","AutoML in real-world projects"]`

To minimize personalized search bias and support reproducibility, all API calls were performed without authentication and mirrored incognito usage, and the same query parameters were applied across both collection rounds [32, 30]. No restrictions were imposed on publication year or geographic region to ensure broad coverage of AutoML-related content.

3.2.2 Data Selection

Literature reviews. The 32 reviews identified in our prior tertiary study [25] were already the result of a structured search and selection process. For this work, we further restricted this set to the 16 reviews that explicitly reported AutoML practices and/or challenges because these dimensions are central to our comparative analysis. The 16 AutoML-focused secondary studies collectively synthesize findings from more than 400 primary studies, where primary-study counts were reported.

YouTube Videos. Video filtering was carried out in two stages using both automated and manual screening. In the first stage, we applied the inclusion (I) and exclusion (E) criteria in Table 1 to the 824 collected videos. Automated filtering based on video metadata (title, description, duration, and language tags) reduced the dataset to 144 candidate videos.

■ **Table 1** Inclusion (I) and exclusion (E) criteria for YouTube videos.

Inclusion criteria (I)	Exclusion criteria (E)
I1 Content related to Automated Machine Learning (AutoML) (e.g., videos referring to AutoCAD or other non-AI automation topics were excluded).	E1 Purely promotional or marketing content (e.g., product advertisements, discount announcements, sponsored “buy now” videos).
I2 Duration longer than one minute (to exclude YouTube Shorts and very short clips).	E2 Non-informative content (e.g., vlogs, reactions, entertainment, or non-technical news coverage).
I3 Content in English.	E3 Duplicate uploads of the same video.

The rest 144 candidate videos were manually inspected for relevance. We kept only videos that presented at least one concrete AutoML workflow or project and explicit discussion of practices and challenges (e.g. rationales for design decisions, trade-offs between tools or settings, and reflections on what worked or failed); we excluded videos that were dominated by tool marketing, very brief GUI click-through demonstrations with no discussion of workflow decisions. This manual screening resulted in 88 videos.

In the second stage, we applied a *stratified purposeful sampling* strategy [20] to ensure that the final sample reflected the diversity of AutoML topics represented in the corpus. Specifically, we grouped the 88 videos into five strata according to the parent terms assigned to their search queries: *AutoML*, *AutoML pipelines*, *AutoML automation tools*, *AutoML case studies*, and *AutoML practices*. The largest strata covered general AutoML overviews and pipeline-focused content, whereas automation-tool, case-study, and practice-focused videos were fewer but conceptually central to our research questions. From this frame, we drew a stratified purposeful sample of 30 videos, aiming for a minimum number of information-rich cases in each stratum and deliberately oversampling the smaller yet theoretically important strata (automation tools, case studies, and practices). This follows Patton’s notion of stratified purposeful sampling, where purposeful samples are nested into strata along key dimensions to capture major variations and enhance the credibility and heterogeneity of qualitative findings, while acknowledging that such samples are not intended for statistical generalization [20]. The final sample of 30 practitioner-created videos has a total running time of approximately 637 minutes (around 11 hours) of content, which we subjected to detailed data extraction and coding.

3.3 Data Extraction

A unified extraction protocol was applied to both datasets to ensure consistency across sources. We systematically extracted information on AutoML practices and challenges, guided by the following operational definitions:

- ● **Practice:** Actions, processes, or workflow strategies used to design, implement, or optimize AutoML pipelines, including methods for data preparation, hyperparameter tuning, model evaluation, and human-in-the-loop integration.
- ■ **Challenge:** Limitations, risks, or barriers encountered in the use or adoption of AutoML, encompassing technical issues (e.g., scalability, reproducibility), organizational barriers (e.g., lack of expertise, integration difficulties), and ethical or regulatory concerns (e.g., bias, explainability, compliance).

For the literature reviews, one author conducted the initial extraction and two other authors independently checked the extracted items against the primary texts. Metadata such as paper ID, title, venue, year, and type of study were retained, and all extracted practices and challenges from the 16 reviews were organized in structured spreadsheets.

For the YouTube dataset, we used the automatically generated transcripts as the primary analysis material, complemented when necessary by viewing relevant video segments to clarify ambiguous statements. We first conducted a small pilot on two videos to calibrate the application of the above definitions and refine the extraction guideline. The standardized protocol was then applied across the full dataset, with each of the 30 videos independently extracted by one author and subsequently validated by a second author. Discrepancies in extracted practices or challenges were discussed in joint sessions until consensus was reached, and the guideline was updated when new decision rules emerged. All extracted data were consolidated in spreadsheets to support cross-source comparison, and the full extraction artifacts are available in our replication package [2].

3.4 Data Synthesis

Open and thematic coding were conducted separately for the literature reviews and the YouTube dataset to preserve the original framing and emphasis of each source. The initial codebook was seeded from the AutoML pipeline phases identified in [25], which provided

the core stages of AutoML workflows. In each dataset, an initial round of open coding was performed by one author, who assigned descriptive labels to the extracted practices and challenges. A second researcher then independently reviewed the assigned codes and their underlying excerpts, assessing coding consistency and the interpretation of emerging categories. Disagreements were resolved through discussion, and the shared codebook was revised accordingly.

Within each dataset, we iteratively applied a constant-comparison strategy, clustering individual codes and merging recurring or conceptually similar practices and challenges to derive higher-level thematic groups. This process reduced redundant open codes (e.g., minor wording variations referring to the same underlying practice) while retaining distinctions that were meaningful for analysis. The code list was continuously refined throughout the synthesis to ensure comprehensive coverage of all extracted practices and challenges. Conducting coding and theme development separately for academic and practitioner sources enabled a more reliable comparison between research and practice perspectives. The unified thematic codebook used in this study, along with a description of each theme and example excerpts, is provided in the replication package [2].

4 Results

This section presents the synthesized results from both datasets, integrating quantitative counts and qualitative insights on AutoML practices and challenges. The findings are organized per RQ.

4.1 RQ1: AutoML Practices

The analysis identified a total of **229 actionable AutoML practices**, organized into **12 thematic groups** and **89 subgroups**. These practices were synthesized from two complementary knowledge sources: Literature Reviews (LR) drawing on 16 peer-reviewed publications, and a practitioner-oriented corpus of 30 YouTube (YT) educational videos. The distribution across sources was markedly asymmetric. Of the 229 practices, 199 (86.9%) originated exclusively from YouTube content and 23 (10.0%) exclusively from the academic literature, with only 6 (2.6%) corroborated by both. In aggregate, YouTube videos contributed to 205 practices (89.5%) and LR papers to 29 (12.7%) – proportions that reflect not a mere difference in corpus size, but a fundamental divergence in what each knowledge channel chooses to document. Academic publications concentrate on formally defined, theoretically grounded problems – neural architecture search, meta-learning, adversarial robustness, and domain-specific clinical applications – whereas practitioner video content engages extensively with applied, tool-specific, and workflow-level concerns that constitute the daily reality of AutoML adoption. This structural difference is reflected in the thematic distribution shown in Figure 2. Optimization emerged as the most extensively documented group (34 practices, 9 subgroups), followed by Pipeline & Orchestration (29, 14 subgroups), Data Preparation (28, 11 subgroups), and Deployment (21, 7 subgroups) – all dominated by YouTube sources and collectively accounting for 48.9% of all identified practices, reflecting the operational emphasis practitioners place on search efficiency, pipeline management, and production readiness. At the other end, Task Formulation (6 practices) and Human-Centric Design (8 practices) were the least represented. Notably, Human-Centric Design was the only group drawing comparably from both sources (4 LR, 4 YT), suggesting that human–AutoML interaction remains an active concern across both communities despite its limited overall representation.



■ **Figure 2** Taxonomy of AutoML practices: 229 practices organized across 12 thematic groups and 89 subgroups, synthesized from 16 LR papers and 30 YouTube videos. Group labels indicate total practice count per group.

Table 2 presents representative actionable practices for the three most densely documented thematic groups – Optimization (34 practices), Pipeline & Orchestration (29 practices), and Data Preparation (28 practices) – which together account for 39.7% of all identified practices. These three groups were selected not only on the basis of volume, but because they illustrate the breadth of the theory–practice divergence most clearly: Optimization is the group with the strongest academic footprint, while Pipeline & Orchestration and Data Preparation are almost entirely practitioner-driven. For each group, three subgroups are presented, each populated with three representative practices drawn from the full catalog, accompanied

by their source attribution (LR = literature; YT = YouTube). Within the Optimization group, the subgroups span from foundational hyperparameter tuning strategies and search space design to more specialized concerns such as neural architecture search, meta-learning warm-starting, and business-constrained objective design. The Pipeline & Orchestration subgroups highlight the operational scaffolding practitioners build around AutoML runs – modular component design, artifact provenance tracking, and end-to-end pipeline structure – none of which appear meaningfully in the academic literature. Data Preparation practices similarly reflect practitioner-led concerns: systematic imputation strategies, column-type-aware preprocessing, and structured data loading workflows that academic treatments rarely operationalize at this level of specificity. In compliance with ESEM’s open science policy, the complete catalog of all 229 actionable practices across all 12 thematic groups and 89 subgroups is provided as a structured online appendix [1].

■ **Table 2** Representative AutoML practices from the two dominant thematic groups. Subgroups selected based on practice count and source coverage (LR = literature review; YT = YouTube). LR/YT[*] denote source indices; the complete list of source list is provided in the online appendix [1].

	Subgroup	Actionable Practice
Optimization	Search Space Design & Acceleration	P1 Design search spaces balancing compactness and comprehensiveness, incorporating human prior knowledge to reduce wasteful exploration. Use cell-based search and structured approaches as practical defaults. [LR5]
		P2 Apply acceleration techniques (e.g., weight sharing, NASP) to reduce search time, especially in AutoDL where final architecture evaluation often far exceeds the discovery cost. [LR5]
		P3 Use NAS-Bench datasets (NAS-Bench-101, NAS-Bench-201, NAS-Bench-NLP) to decouple algorithm design from expensive evaluation, enabling faster iteration on optimization strategies. [LR8]
	Neural Architecture Search (NAS)	P4 Use differentiable architecture search (DARTS) for gradient-based NAS, and explore joint HPO + architecture optimization to find both architecture and hyperparameters in a single pass. [LR8, YT29]
		P5 Combine meta-learning with NAS to reuse and transfer architectures across tasks, enabling fast adaptation to new problems. Extend NAS to few-shot and unsupervised settings where labelled data is scarce. [LR8]
		P6 Apply NAS with reinforcement learning to automate the discovery of optimal layer counts and neuron configurations for deep learning models, replacing iterative manual architecture design. [YT21]
	Business-Aware & Constraint-Driven Optimization	P7 Design or select AutoML systems that explicitly balance accuracy, efficiency, scalability, and flexibility – these four criteria should guide both tool selection and evaluation. [LR13]
		P8 Incorporate business constraints into modelling – e.g., monotonicity requirements from regulations, prediction range constraints, or weighted learning when some examples matter more than others. [YT18]
		P9 Automate target distribution analysis to recommend an appropriate mathematical objective suited to the data. Apply multi-objective fine-tuning where relevant. [YT18, YT22]
Pipeline & Orchestration	Artifact Storage & Provenance Tracking	P10 Store all pipeline artifacts (datasets, models, hyperparameters, code, accuracy scores) in cloud storage to support reproducibility, debugging, and model deprecation. Enable artifact reuse across pipeline components. [LR2, YT25]
		P11 Document and maintain provenance records for error tracing and auditing. Store artifacts from original model runs in tracking servers (e.g., MLflow) for documentation and regulatory compliance. [LR2, YT9]
		P12 Track full model lineage – log all inputs, outputs, hyperparameters, and training data references throughout the pipeline to answer provenance questions and identify what changed between runs. [YT25]
	Component-Based Pipeline Design	P13 Design pipelines as modular, component-based workflows where each step is defined by exactly one component with well-specified inputs and outputs. Components can perform preprocessing, training, evaluation, or deployment. [YT25]
		P14 Chain transformers and estimators sequentially to build pipeline models: transformers operate row-by-row on columns, estimators learn from full columns, and outputs feed into the next stage. [YT8]
		P15 Reuse prebuilt pipeline components (e.g., Google Vertex AI prebuilt components) before building custom ones. Share components across projects and teams via repositories (e.g., AI Hub). [YT25]

■ **Table 2** Representative AutoML practices (continued).

End-to-End ML Pipeline Structure	P16	Structure ML projects as end-to-end pipelines: data collection → cleaning/feature engineering → model selection → evaluation → deployment. Automate as many stages as possible while maintaining human oversight at key decision points. [YT10, YT29, YT17, YT9]
	P17	Use comprehensive AutoML frameworks that automate model selection, HPO, data cleaning, and feature engineering as part of one unified pipeline, reducing the need for separate tool integration. [YT29]
	P18	Build complete pipelines that include imputation, feature selection, classification, and calibration – not just prediction models. Enable pipeline-aware feature engineering matched to the specific pipeline being searched. [YT4, YT18]

4.2 RQ2: AutoML Challenges

The analysis identified a total of **177 AutoML challenges**, organized into **12 thematic groups** and **75 subgroups**, drawing from the same corpus of 16 LR papers and 30 YouTube videos. The source distribution for challenges reverses the pattern observed for practices: the academic literature contributed to 103 challenges (58.2% of all challenges, counting shared entries once per source) and YouTube content to 83 (46.9%), with 9 challenges (5.1%) corroborated by both. This reversal is structurally revealing. Whereas practitioners are the primary documenters of *what to do* in AutoML workflows, the formal literature more extensively captures *what goes wrong* – particularly in areas that require systematic evaluation, theoretical analysis, or longitudinal observation to surface. This asymmetry suggests that failure modes in AutoML are not always immediately visible to practitioners during hands-on adoption; many manifest only at scale, under rigorous experimental conditions, or across multiple deployment cycles.

Figure 3 presents the full taxonomy of AutoML challenges across all 12 groups and 75 subgroups. The most extensively documented challenge groups are Process (23 challenges), Optimization (22 challenges), and Deployment (20 challenges), together accounting for 36.7% of all identified challenges and collectively capturing the central difficulty of AutoML adoption: despite advances in automation, building complete, reliable, and maintainable pipelines remains fundamentally hard. Scope & Applicability (18 challenges) and Data Preparation (18 challenges) follow closely. The latter is noteworthy: all 18 Data Preparation challenges originate exclusively from YouTube, confirming that data-level friction – fragmented sources, imputation complexity, format diversity, and business logic drift – is a practitioner reality the academic literature has largely left untheorized. Transparency (8 challenges), Feature Engineering (6 challenges), and Ethics, Safety & Compliance (9 challenges), while smaller in volume, carry disproportionate weight for real-world deployment: the interpretability deficit directly undermines clinical adoption and regulatory compliance, while unresolved ethical and fairness risks grow more consequential as AutoML systems are deployed in high-stakes domains.

Table 3 presents representative challenges for the three most extensively documented groups – Process (23 challenges), Optimization (22 challenges), and Deployment (20 challenges) – accounting for 36.7% of all identified challenges. These groups were selected because they represent distinct failure modes across the AutoML lifecycle: process-level breakdowns in pipeline automation and team coordination, optimization-level failures in search efficiency and model selection, and deployment-level gaps in production monitoring and CI/CD integration. Within the Process group, challenges are shared across both sources but articulated differently – the literature frames them as systemic design failures, while practitioners describe them as operational pain points. The Optimization group reveals a persistent tension between what AutoML promises and what it delivers under time, memory,

27:12 A Multivocal Review of AutoML Practices and Challenges



■ **Figure 3** Taxonomy of AutoML challenges: 177 challenges organized across 12 thematic groups and 75 subgroups, synthesized from 16 LR papers and 30 YouTube videos. Group labels indicate total challenge count per group.

and budget constraints. The Deployment group exposes a blind spot common to both sources: drift detection, monitoring failures, and missing CI/CD abstractions are identified by both communities, yet neither offers mature solutions. The complete catalog is provided in the online appendix [1].

■ **Table 3** Representative AutoML challenges from the two dominant thematic groups. Subgroups are selected based on challenge count and source coverage (LR = literature review; YT = YouTube). LR/YT[*] denote source indices; the complete list of source list is provided in the online appendix [1].

Subgroup		Challenge
Process	Incomplete End-to-End Pipeline Automation	C1 AutoML systems fail to automate the entire ML pipeline – many stages still require human effort. Task Formulation is not currently automated, and defining the right prediction goal remains more important than algorithm choice. [LR6, LR9]
		C2 AutoML may deliver suboptimal performance due to poor automation decisions in some tasks. Even when automation is available, existing tools require human interventions to achieve production-quality results. [LR6, YT24]
		C3 Predictive modelling cannot be completely automated – some detection and risk assessment tasks require human judgment even when AutoML provides partial support. [YT18]
	Scalability & Deployment Process Limitations	C4 Deploying a single model traditionally requires many data engineers, scientists, and engineers – this process does not scale as each new model demands a large team. Building integrations across data collection, training, and deployment remains the central challenge. [YT8]
		C5 The training stage is highly iterative and effort-intensive, requiring extensive time and experimentation to converge on a good solution. ML pipeline implementation is consistently time-consuming across organisations. [YT17, YT25]
		C6 Organisational and process barriers – including expertise scarcity, process sluggishness, and diverse stakeholder expectations – compound the difficulty of delivering ML projects at scale. [LR10]
	Incomplete & Fragmented Pipeline Coverage	C7 Most AutoML libraries only focus on parts of the pipeline – preprocessing and post-processing remain underexplored. Integration gaps exist across ML pipeline stages, with a lack of intelligent automation in decision-making. [LR8, LR7, LR6]
		C8 AutoML tools lack transparency, flexibility, adequate documentation, and regulatory compliance support, making their integration into existing organisational processes complex. [LR6]
		C9 ML workflow acceleration focuses on one-time execution, neglecting gradual and continuous development. Long development cycles (months to years) persist even for experienced teams. [LR10]
	Compute & Memory Burden at Scale	C10 Scaling AutoML to handle big data efficiently is an ongoing challenge. AutoML tools face scalability issues with large datasets due to training inefficiencies and high compute requirements – especially for hyperparameter tuning. [LR22, LR6]
		C11 AutoML can crash or fail due to memory limitations on large or complex datasets. For deep learning with millions of rows, more powerful servers are required to complete training in reasonable time. [LR6, YT20]
		C12 AutoML training is excessively time-consuming across frameworks. Excessive training time is a universal complaint regardless of tool, dataset size, or model type. [YT21, YT29, YT28]
Deployment	Missing CI/CD & Automated Post-Deployment Support	C13 AutoML lacks native support for modern CI/CD pipelines, limiting reliability and maintainability. AutoML tools also do not provide automated solutions for monitoring and updating models after deployment. [LR6, LR7]
		C14 Organisations face deployment challenges due to high compute requirements, infrastructure setup costs, and limited cross-environment deployment support across different real-world environments. [LR6]
		C15 RL-trained models are often impractical to deploy due to infeasible answers in real use cases. Deploying ML systems on the cloud raises difficulties around portability, cost management, and operations. [LR10]
	Production Monitoring & Data Drift	C16 Model evaluation in AutoML platforms is typically done offline, raising concerns about performance monitoring once models are in production. Continuous dashboards and alerts for tracking performance over time are often absent. [YT16, YT7]
		C17 Detecting data drift, transform inconsistencies, and latency pattern shifts during execution is critical for reliability. Feedback loop detection in production – where deployed models influence future training data – remains an open challenge. [YT7, YT18]
		C18 The world changes over time – models will eventually encounter data they have never seen during training. Automated drift detection and adaptation are essential but remain largely underbuilt in current AutoML tools. [YT18]

4.3 RQ3: Alignment and Divergence across Literature and Youtube

Building on the taxonomies established in RQ1 and RQ2, this section examines convergence and divergence between academic and practitioner AutoML knowledge, drawing on source attribution data across all 406 coded items (229 practices, 177 challenges) and supported by Figures 4 and 5. The overarching finding is one of structural misalignment: only 6 shared practices (2.6%) and 9 shared challenges (5.1%) were corroborated by both sources – reflecting a genuine and systematic divergence in what each community attends to, how it frames problems, and what it considers worth documenting.

Alignment and Divergence in Reported Practices. Figure 4 reveals that meaningful alignment between the literature and YouTube is confined to a small number of thematic groups. *Optimization* is the clearest point of convergence: both sources document practices related to hyperparameter search, neural architecture search, and meta-learning warm-starting, reflecting a shared recognition that search efficiency is central to AutoML. *Evaluation* and *Tool Usage* also appear across both sources, indicating that validation reliability and operational framework use represent common ground – practitioners and researchers alike are invested in knowing whether a pipeline works and which tools make it work. *Transparency* represents a subtler form of alignment: academic treatments frame it as explainability and model interpretability as design properties, while YouTube content addresses it through practical tooling (e.g., SHAP values, feature importance dashboards), but both communities recognize that AutoML outputs must be interpretable to be trustworthy. Beyond these areas, the divergence is sharp and consistent. YouTube content dominates *Deployment* (21 practices, 0 from LR), *Data Preparation* (28 practices, 0 from LR), and *Pipeline & Orchestration* (29 practices, 2 from LR) – groups that collectively account for 34.1% of all identified practices yet are almost entirely absent from the academic literature. These are not peripheral concerns: they represent the operational scaffolding without which no AutoML system reaches production. The literature, by contrast, concentrates its practice-level contributions in *Human-Centric Design* (4 of 8 practices from LR), *Model Selection* (5 of 16 from LR), and *Optimization* (8 of 34 from LR) – areas characterized by theoretical framing and algorithmic novelty. In short, academic publications tend to document AutoML improvement as a methodological pursuit, while practitioners approach it as an engineering discipline of efficiency, maintainability, and operational reliability.

Alignment and Divergence in Reported Challenges. The challenge data presents a partial but important reversal of this pattern. As shown in Figure 5, the literature and YouTube converge more meaningfully around challenges than around practices, particularly in *Optimization*, *Process*, and *Human-Centric Design*. The shared recognition of HPO difficulty, incomplete pipeline automation, and the persistent need for domain expertise suggests that certain failure modes are visible to both communities, even if they are encountered through different routes – researchers through systematic benchmarking, practitioners through hands-on deployment. *Scope & Applicability* also emerges as a shared concern: both sources acknowledge that AutoML is not one-size-fits-all, that it overemphasizes HPO and feature engineering relative to the full pipeline, and that terminology confusion around AutoML, AutoAI, and AutoDS creates misaligned expectations. Divergence persists, however, and follows a predictable pattern. The literature contributes disproportionately to *Data Quality* (10 challenges, 0 from YouTube), *Ethics, Safety & Compliance* (6 LR vs. 4 YT), and *Evaluation* (8 LR vs. 4 YT) – challenge areas that require sustained empirical investigation, controlled experimental conditions, or regulatory awareness to surface. Practitioners, by contrast,

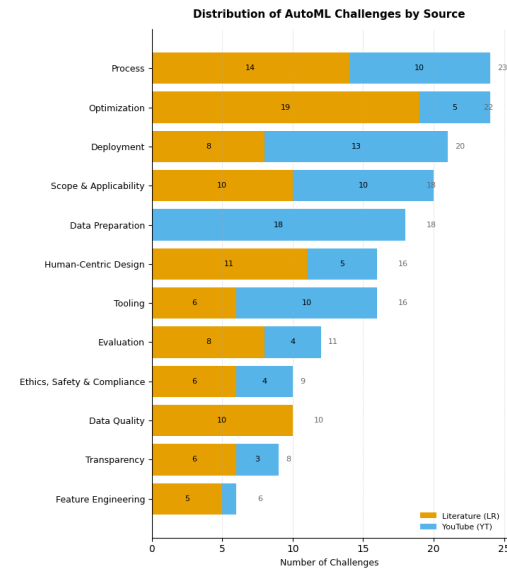
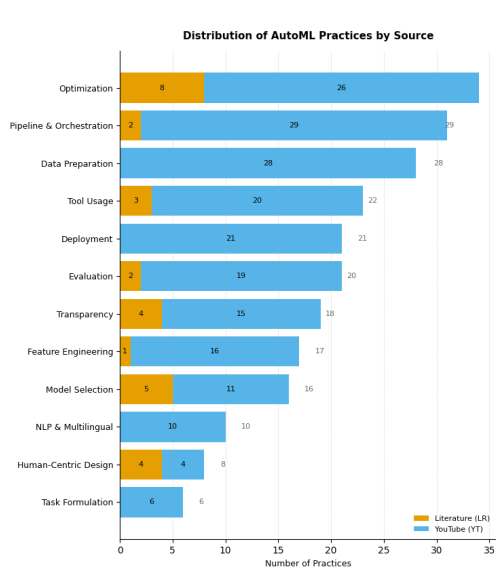


Figure 4 Distribution of AutoML practices per thematic group by source (LR vs. YouTube). Groups ordered by total practice count. **Figure 5** Distribution of AutoML challenges per thematic group by source (LR vs. YouTube). Groups ordered by total challenge count.

are the sole documenters of *Data Preparation* challenges (18 challenges, 0 from LR), and contribute substantially to *Tooling* (10 YT vs. 6 LR) and *Deployment* (13 YT vs. 8 LR). The complete absence of Data Preparation challenges from the academic literature is particularly striking: issues such as fragmented data sources, business logic drift, imputation complexity, and format diversity are reported consistently and in granular detail by practitioners, yet remain essentially untheorized in the formal literature. This represents one of the clearest instances of the theory–practice gap this study documents.

Cross-theme Reflections. Examining practices and challenges together reveals a recurring and important structural pattern: the thematic areas where practitioners are most active in developing operational knowledge are precisely the areas where the literature most underrepresents challenges – and vice versa. *Optimization* illustrates this duality most clearly. It is simultaneously the group with the broadest practice documentation (34 practices, spanning HPO, NAS, meta-learning, and edge-aware deployment) and a source of persistent, well-documented challenges (22 challenges around search cost, generalization failures, and performance–cost trade-offs). Both communities recognize optimization as central to AutoML, but the literature frames it as a frontier research problem while practitioners treat it as a configurable workflow component – a distinction that shapes what each community considers a solution. *Data Preparation* presents the inverse pattern. Practitioners document it extensively as both a practice domain (28 practices) and a challenge domain (18 challenges), yet the literature contributes to neither. This bilateral absence from academic sources is not explained by the topic’s simplicity – the granularity and specificity of practitioner-reported data preparation challenges (MNAR imputation, schema drift, multilingual NLP preprocessing) indicate that these are non-trivial engineering problems deserving formal treatment. *Human-Centric Design* exhibits yet another pattern: it is documented as a challenge primarily in the literature (11 of 16 challenges from LR), yet appears as a practice

relatively infrequently and predominantly in academic sources (4 of 8 practices from LR). This suggests that human-centered concerns are well-theorized but rarely operationalized – practitioners are not yet systematically building the human-in-the-loop mechanisms that the literature identifies as necessary.

Taken together, these findings confirm that AutoML research and practice operate in largely parallel, non-communicating worlds. The literature theorizes optimization, robustness, and human-AI interaction at a conceptual level, while practitioners have independently developed an extensive, operationally grounded body of knowledge around deployment, data preparation, and pipeline orchestration that research has scarcely acknowledged. This gap is not merely a communication problem – it reflects a structural misalignment in research priorities that risks making AutoML scholarship increasingly disconnected from the practitioners it purports to serve.

5 Discussion

Drawing on the synthesis across RQ1–RQ3, this section identifies open issues and research opportunities, and practical implications for AutoML researchers and practitioners.

5.1 Open Issues and Research Opportunities

Several open issues emerge from our analysis that warrant sustained attention from the research community. *Theory-practice gap.* The most immediate open issue is the absence of academic engagement with a large and coherent body of practitioner knowledge. Entire thematic areas – Data Preparation, Deployment, and Pipeline & Orchestration – are documented almost exclusively through practitioner sources, yet the challenges within these areas are neither trivial nor solved. Research that formalizes, evaluates, and builds upon this practitioner knowledge base – through empirical studies, tool evaluations, and replication studies in real-world settings – would substantially reduce this gap.

Human-centered AutoML remains under-operationalized. The literature consistently identifies human oversight, interpretability, and trust as critical requirements, yet practitioner content reveals that these properties are rarely built into AutoML workflows in practice. This gap calls for research into concrete mechanisms – human-in-the-loop interfaces, explainability-by-design frameworks, and participatory evaluation protocols – that translate human-centered principles into implementable AutoML components.

Ethics, fairness, and compliance are structurally underserved. With only 9 challenges identified in Ethics, Safety & Compliance and limited practitioner engagement with fairness and bias, this study finds that ethical dimensions of AutoML remain at the periphery of both research and practice. As AutoML systems are increasingly deployed in high-stakes domains – healthcare, finance, criminal justice – the absence of standardized fairness evaluation protocols, privacy-preserving automation techniques, and accountability mechanisms represents a risk that the community must proactively address.

5.2 Implications for Practice

The findings carry concrete implications for practitioners integrating AutoML into production.

Treat AutoML as a continuous process, not a one-time automation task. The dominance of Deployment, Data Preparation, and Pipeline & Orchestration in practitioner-reported practices reflects a mature understanding that AutoML value is realized over time, through

structured experimentation, versioned artifacts, and automated retraining pipelines – not through a single tool invocation. Practitioners should invest in CI/CD integration, metadata capture, and automated monitoring from the outset.

Resource management is a first-class concern. Optimization-related challenges – excessive training time, memory failures, and cost–performance trade-offs – are reported consistently across practitioner sources. Rather than relying on tool defaults, practitioners should define explicit computational budgets, employ adaptive early stopping, and conduct cost-effectiveness analysis throughout the AutoML lifecycle. Hardware-aware scheduling and parallelization strategies represent immediately actionable mitigations.

Operationalize interpretability before deployment, not after. The literature’s strong emphasis on transparency and human-centric design offers practitioners a clear prescriptive signal that is currently underheeded. Embedding explainability dashboards, model lineage tracking, and human-in-the-loop validation checkpoints throughout the AutoML lifecycle – rather than as post-hoc additions – is particularly critical in regulated or high-stakes domains where auditability is not optional.

6 Threats to Validity

This study follows established guidelines for multivocal literature reviews in software engineering [13], and addresses construct, internal, external, and conclusion validity threats as deliberate methodological considerations rather than residual limitations. **Construct validity.** The primary threat concerns operationalizing *practices* and *challenges* across heterogeneous sources that differ in terminology, abstraction, and communicative intent. We mitigated this by defining a unified extraction protocol and shared codebook before data collection and refining constructs through pilot coding. Multiple researchers independently assigned codes, and thematic groupings were reconciled through structured discussion, with disagreements resolved by returning to the source material to keep codes evidence-grounded rather than merely negotiated. **Internal validity.** Two sources of selection bias warrant acknowledgment. First, the LR dataset is restricted to secondary studies (i.e., literature reviews and surveys) meeting our inclusion criteria. Second, the gray literature corpus is intentionally scoped to YouTube, excluding other practitioner channels such as blogs, forums, and webinars. The YouTube dataset is subject to platform visibility bias, as search rankings reflect engagement metrics and recency rather than epistemic quality. We mitigated this by conducting all searches in incognito mode to minimize personalization effects, and by making all search strings, inclusion and exclusion criteria, and sampling decisions fully explicit and reproducible. **External validity.** Findings characterize AutoML knowledge as represented in peer-reviewed secondary literature and publicly available English-language practitioner videos. Generalizability is bounded in three respects: results may not transfer to proprietary industrial environments; they exclude non-English-speaking communities that may exhibit distinct adoption patterns; and they omit other gray-literature channels such as blogs, forums, and webinars. The multivocal design partially mitigates this by triangulating white and gray literature, but we make no claim of statistical representativeness. **Conclusion validity.** Reported counts are frequency indicators from a qualitative coding process, not statistical prevalence estimates. Thematic distributions are presented through bar charts and taxonomy figures to reveal structural patterns rather than confirm hypotheses. We mitigated confirmation bias through independent coding, iterative codebook refinement, and full documentation of the analysis, all available in the replication package [2].

7 Conclusion

This study conducted a multivocal literature review synthesizing 16 peer-reviewed publications and 30 practitioner YouTube videos, identifying 229 actionable AutoML practices across 12 thematic groups and 177 challenges across 12 thematic groups. The central finding is a striking structural asymmetry: only 6 practices (2.6%) and 9 challenges (5.1%) were corroborated by both sources. Practitioners are the near-exclusive documenters of operational knowledge – Deployment, Data Preparation, and Pipeline & Orchestration – while the literature disproportionately covers theoretically grounded challenges in Data Quality, Ethics & Compliance, and Evaluation. The result is two coherent but largely non-communicating knowledge bases: one theoretically sophisticated, the other operationally rich. This structural misalignment risks making AutoML scholarship increasingly disconnected from the practitioners it purports to serve. Key open issues include incomplete end-to-end pipeline automation, underdeveloped human-centered design, inadequate benchmarking infrastructure, and a near-total lack of academic engagement with data preparation challenges. In future work, we plan to extend the corpus to complementary gray-literature channels (blogs, forums, conference talks) and validate findings through interviews with AutoML practitioners.

Data Availability Statement

The replication package [2] includes the search strings, inclusion and exclusion criteria, extracted data, and the codebook with definitions and representative examples, made openly available to support transparency, auditability, and replication. The online appendix [1] provides the complete browsable catalog of all 229 actionable practices and 177 challenges across all 12 thematic groups and subgroups, intended as an open reference for researchers and practitioners.

References

- 1 Online Appendix for this review. <https://doi.org/10.5281/zenodo.20282909>, 2026.
- 2 Replication package for this review. <https://doi.org/10.5281/zenodo.20283067>, 2026.
- 3 Muhammad Junaid Ali, Mokhtar Essaid, Laurent Moalic, and Lhassane Idoumghar. A review of AutoML optimization techniques for medical image applications. *Computerized Medical Imaging and Graphics*, 118:1–35, 2024. doi:10.1016/J.COMPAMEDIMAG.2024.102441.
- 4 Florian Bachinger, Jan Zenisek, and Michael Affenzeller. Automated machine learning for industrial applications – challenges and opportunities. *Procedia Computer Science*, 232:1701–1710, 2024-01-01. doi:10.1016/j.procs.2024.01.168.
- 5 Maroua Bahri, Flavia Salutari, Andrian Putina, and Mauro Sozio. AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. *Int. J. Data Sci. Anal.*, 14(2):113–126, 2022-08-01. doi:10.1007/s41060-022-00309-0.
- 6 Rafael Barbudo, Sebastián Ventura, and José Raúl Romero. Eight years of AutoML: categorisation, review and trends. *Knowl. Inf. Syst.*, 65(12):5097–5149, 2023. doi:10.1007/s10115-023-01935-1.
- 7 Bo Chen, Xiangyu Zhao, Yejing Wang, Wenqi Fan, Huifeng Guo, and Ruiming Tang. A Comprehensive Survey on Automated Machine Learning for Recommendations. *ACM Trans. Recomm. Syst.*, 2(2):13:1–13:38, April 2024. doi:10.1145/3630104.
- 8 Yi-Wei Chen, Qingquan Song, and Xia Hu. Techniques for automated machine learning. *SIGKDD Explor.*, 22(2):35–50, 2021-01-17. doi:10.1145/3447556.3447567.
- 9 Ronald Chenail. YouTube as a Qualitative Research Asset: Reviewing User Generated Videos as Learning Resources. *The Qualitative Report*, 16(1):229–235, January 2011. doi:10.46743/2160-3715/2011.1048.

- 10 Liam Connor, Rebecca Kelly, Saoirse Murphy, Eoghan McCarthy, and Edward Murray. Analysis of AutoML tools in the world of automated deep learning. *Fusion of Multidisciplinary Research, An International Journal*, 5(1):541–555, 2024-01-14. Number: 1. URL: <https://fusionproceedings.com/fmr/1/article/view/61>.
- 11 Aline Marques Del Valle, Rafael Gomes Mantovani, and Ricardo Cerri. A systematic literature review on AutoML for multi-target learning tasks. *Artificial Intelligence Review*, 56(2):2013–2052, 2023-11-01. doi:10.1007/s10462-023-10569-2.
- 12 Radwa El Shawi, Mohamed Maher, and Sherif Sakr. Automated machine learning: State-of-the-art and open challenges. *CoRR*, 2019. doi:10.48550/arXiv.1906.02287.
- 13 Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology*, 106:101–121, February 2019. doi:10.1016/j.infsof.2018.09.006.
- 14 Dhruvi Gosai and Minal Patel. Advancing the machine learning workflow: A comprehensive analysis of auto machine learning (auto-ML) techniques and future investigations. In Ashish Kumar Luhach, Dharm Singh Jat, Uttam Ghosh, Xiao-Zhi Gao, and Pawan Lingras, editors, *Advanced Informatics for Computing Research*, pages 308–323. Springer Nature Switzerland, 2025. doi:10.1007/978-3-031-84062-3_25.
- 15 Xin He, Kaiyong Zhao, and Xiaowen Chu. AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021-01. doi:10.1016/j.knosys.2020.106622.
- 16 Shubhra Kanti Karmaker (“Santu”), Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. AutoML to date and beyond: Challenges and opportunities. *ACM Comput. Surv.*, 54(8):175:1–175:36, 2021-10-04. doi:10.1145/3470918.
- 17 Brad McKenna, Michael D. Myers, and Michael Newman. Social media in qualitative research: Challenges and recommendations. *Information and Organization*, 27(2):87–99, June 2017. doi:10.1016/j.infoandorg.2017.03.001.
- 18 Alejandro Morales-Hernández, Inneke Van Nieuwenhuyse, and Sebastian Rojas Gonzalez. A survey on multi-objective hyperparameter optimization algorithms for machine learning. *Artificial Intelligence Review*, 56(8):8043–8093, August 2023. doi:10.1007/s10462-022-10359-2.
- 19 Mohamed M. Mostafa, Ali Feizollah, and Nor Badrul Anuar. Fifteen years of YouTube scholarly research: knowledge structure, collaborative networks, and trending topics. *Multimedia Tools and Applications*, 82(8):12423–12443, 2023. doi:10.1007/s11042-022-13908-7.
- 20 Michael Quinn Patton. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. Sage, Thousand Oaks, CA, 3 edition, 2002.
- 21 Yannis Poulakis, Christos Doukeridis, and Dimosthenis Kyriazis. A survey on AutoML methods and systems for clustering. *ACM Transactions on Knowledge Discovery from Data*, 18(5):1–30, 2024-06-30. doi:10.1145/3643564.
- 22 Fernando F. Prado and Luciano A. Digiampietri. A systematic review of automated feature engineering solutions in machine learning problems. In *Proceedings of the XVI Brazilian Symposium on Information Systems*, SBSI '20, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3411564.3411610.
- 23 Luigi Quaranta, Kelly Azevedo, Fabio Calefato, and Marcos Kalinowski. A multivocal literature review on the benefits and limitations of industry-leading AutoML tools. *Information and Software Technology*, 178:107608, 2025. doi:10.1016/j.infsof.2024.107608.
- 24 Mohaimenul Azam Khan Raiaan, Sadman Sakib, Nur Mohammad Fahad, Abdullah Al Mamun, Md. Anisur Rahman, Swakkhar Shatabda, and Md. Saddam Hossain Mukta. A systematic review of hyperparameter optimization techniques in convolutional neural networks. *Decision Analytics Journal*, 11:100470, 2024. doi:10.1016/j.dajour.2024.100470.
- 25 Keerthiga Rajenthiram, Pauline Delparte, and Patricia Lago. AutoML: A tertiary study of phases, methods, tools, and frameworks. In Domenico Bianculli, Hassan Sartaj, Vasilios Andrikopoulos, Cesare Pautasso, Tommi Mikkonen, Jennifer Perez, Tomáš Bureš, Martina De Sanctis, Henry Muccini, Elena Navarro, Mohamed Soliman, and Uwe Zdun, editors, *Software Architecture. ECSA 2025 Tracks and Workshops*, pages 313–328. Springer Nature Switzerland, 2026. doi:10.1007/978-3-032-04403-7_26.

- 26 S Ravishankar and Battineni Gopi. A survey on recent advancements in auto-machine learning with a focus on feature engineering | journal of computational and cognitive engineering. URL: <https://ojs.bonviewpress.com/index.php/JCCE/article/view/720>.
- 27 Imrus Salehin, Md Islam, Pritom Saha, S M Noman, Azra Tunj, Md.Mehedi Hasan, and Md Baten. AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, 2, 2023-10-01.
- 28 Imrus Salehin, Md Shamiul Islam, Pritom Saha, S. M. Noman, Azra Tunj, Md Mehedi Hasan, and Md Abu Baten. AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, 2(1):52–81, 2024. URL: <https://www.sciencedirect.com/science/article/pii/S2949715923000604>.
- 29 Sasan Salmani Pour Avval, Nathan D. Eskue, Roger M. Groves, and Vahid Yaghoubi. Systematic review on neural architecture search. *Artificial Intelligence Review*, 58(3):73, January 2025. doi:10.1007/s10462-024-11058-w.
- 30 Margaret Sampson, Jordi Cumber, Claudia Li, Catherine M. Pound, Ann Fuller, and Denise Harrison. A systematic review of methods for studying consumer health YouTube videos, with implications for systematic reviews. *PeerJ*, 1:e147, September 2013. doi:10.7717/peerj.147.
- 31 Alex Serban, Koen Van Der Blom, Holger Hoos, and Joost Visser. Adoption and effects of software engineering best practices in machine learning - supplementary material. Publisher: Zenodo Version Number: 0.0.9. doi:10.5281/ZENODO.3946453.
- 32 Wuyou Sui, Anna Sui, and Ryan E Rhodes. What to watch: Practical considerations and strategies for using YouTube for research. *Digital Health*, 8:20552076221123707, September 2022. doi:10.1177/20552076221123707.
- 33 Junchi Tian and Chang Che. Automated machine learning: A survey of tools and techniques. *Journal of Industrial Engineering and Applied Science*, 2(6):71–76, 2024-12-01. doi:10.70393/6a69656173.323336.
- 34 Jonathan Waring, Charlotta Lindvall, and Renato Umeton. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, 104:1–12, 2020. doi:10.1016/J.ARTMED.2020.101822.
- 35 Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. Whither AutoML? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–16. Association for Computing Machinery, 2021. doi:10.1145/3411764.3445306.
- 36 Han Yuan, Kunyu Yu, Feng Xie, Mingxuan Liu, and Shenghuan Sun. Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare. *Medicine Advances*, 2(3):205–237, 2024. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/med4.75>. doi:10.1002/med4.75.
- 37 Ruiqi Zheng, Liang Qu, Bin Cui, Yuhui Shi, and Hongzhi Yin. AutoML for Deep Recommender Systems: A Survey. *ACM Trans. Inf. Syst.*, 41(4):101:1–101:38, March 2023. doi:10.1145/3579355.
- 38 Marc-André Zöller and Marco F. Huber. Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research*, 70:409–472, 2021-01-27. doi:10.1613/jair.1.11854.