

The Technological Emergence of AutoML

A Survey of Performant Software and Applications in the Context of Industry

ALEXANDER SCRIVEN, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Australia

DAVID JACOB KEDZIORA, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Australia

KATARZYNA MUSIAL, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Australia

BOGDAN GABRYS, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Australia

With most technical fields, there exists a delay between fundamental academic research and practical industrial uptake. Whilst some sciences have robust and well-established processes for commercialisation, such as the pharmaceutical practice of regimented drug trials, other fields face transitory periods in which fundamental academic advancements diffuse gradually into the space of commerce and industry. For the still relatively young field of Automated/Autonomous Machine Learning (AutoML/AutonoML), that transitory period is under way, spurred on by a burgeoning interest from broader society. Yet, to date, little research has been undertaken to assess the current state of this dissemination and its uptake. Thus, this review makes two primary contributions to knowledge around this topic. Firstly, it provides the most up-to-date and comprehensive survey of existing AutoML tools, both open-source and commercial. Secondly, it motivates and outlines a framework for assessing whether an AutoML solution designed for real-world application is 'performant'; this framework extends beyond the limitations of typical academic criteria, considering a variety of stakeholder needs and the human-computer interactions required to service them. Thus, additionally supported by an extensive assessment and comparison of academic and commercial case-studies, this review evaluates mainstream engagement with AutoML in the early 2020s, identifying obstacles and opportunities for accelerating future uptake.

Additional Key Words and Phrases: Automated machine learning (AutoML)

1 INTRODUCTION

Societal interest in machine learning (ML), especially the subtopic of deep learning (DL), has surged within recent years. This is partially driven by the continuing success of these approaches in many application areas [240, 398, 424, 439], facilitated by both fundamental advances [147, 152, 323, 506] and the increasing availability of computational resources. Unsurprisingly, on the academic side, the field of artificial intelligence (AI) continues to dominate research outputs, as noted by the 2021 UNESCO Science Report [487]. However, it is the current level of ML engagement in industry that is truly unprecedented. For instance, the 2021 Global AI Adoption Index, commissioned by IBM, found that 80% of 5501 global businesses are either using automation software or planning to within 12 months, and 74% are exploring or deploying AI [274]. The Gartner 2019 CIO Agenda survey, with 3000 respondents from across the globe, agrees with this trend, revealing that the proportion of firms deploying AI has increased from 10% in 2015 to 37% in 2019 [268]. Similar

Authors' addresses: Alexander Scriven, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Sydney, New South Wales, 2007, Australia, alexander.scriven@uts.edu.au; David Jacob Kedziora, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Sydney, New South Wales, 2007, Australia, david.kedziora@uts.edu.au; Katarzyna Musial, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Sydney, New South Wales, 2007, Australia, katarzyna.musial-gabrys@uts.edu.au; Bogdan Gabrys, Complex Adaptive Systems Lab, Data Science Institute, University of Technology Sydney, Sydney, New South Wales, 2007, Australia, bogdan.gabrys@uts.edu.au.

conclusions are echoed in the 2020 McKinsey ‘State of AI’ report [387]. Naturally, such a rate of mainstream permeation is also accompanied by intensifying discussions on how to use ML, and AI more broadly, in a socially responsible manner [284, 348, 375, 405, 535].

Nonetheless, despite the growing desire of industry to utilise ML, talent in data science remains scarce [443, 516]. Both the Gartner and IBM studies agree that lack of expertise creates a barrier to AI adoption [268, 274], especially as, by and large, ML technology still requires specialist skills to implement and employ. Worse yet, in practice, deploying ML solutions for real-world applications requires technical skills beyond the domain of data science. Any shortfall in these broader talents will also adversely affect ML engagement in industry [230, 510]. So, faced with these realities, a business may ponder: does ML really have to rely so heavily on humans? Enter ‘automated machine learning’ (AutoML), a research endeavour that has become particularly popular over the last decade [203, 204, 257, 273, 340, 485, 582], striving to mechanise as many high-level ML operations as possible. The appeal of this emergent field is multi-faceted, driven by many of the same motivations that inspire automation in general. These include not just democratisation, enabling the broader public to leverage the power of ML approaches, but also efficiency boosts, redistributing the time and effort of existing talent to more valuable functions.

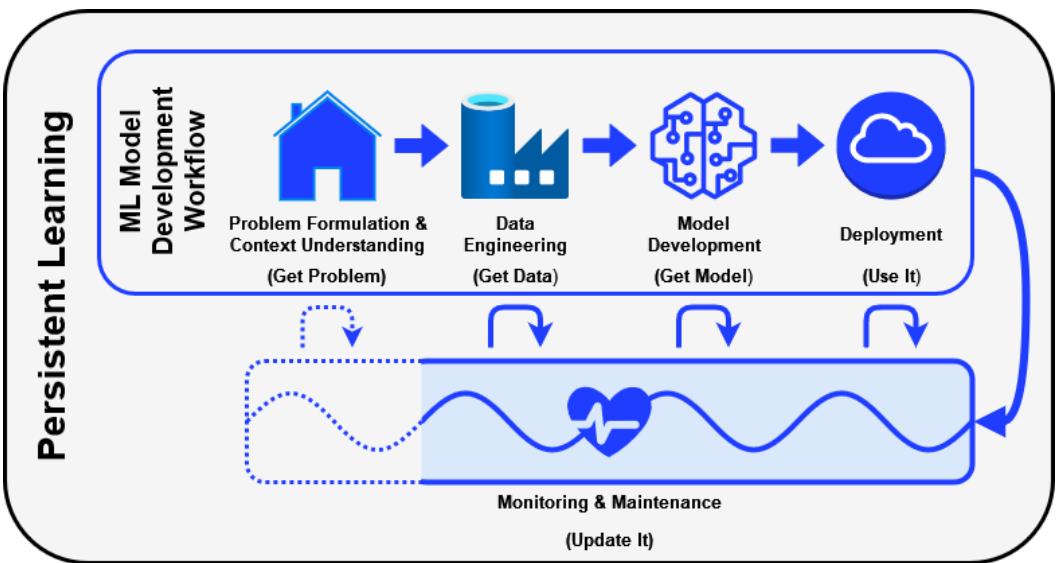


Fig. 1. Schematic of a general machine learning (ML) workflow, which captures the phases involved in designing, constructing, deploying and maintaining an ML model for a real-world application.

Notably, within the modern era of AutoML, academia has already made much progress. Admittedly, it can be challenging to contain this ever-widening field within a simple overview, and various works lean on taxonomies and categorical systems to aid this [195, 308, 313, 485]. Consider then a conceptual representation of the processes that are involved in running a real-world ML application, i.e. an ML workflow, as shown in Fig. 1. With respect to this depiction, the bulk of AutoML research has traditionally focused on automating the model-development phase. Advances in Bayesian optimisation, which continue to be employed [224, 315], are frequently credited for jump-starting this process, reducing human involvement in hyperparameter optimisation (HPO) and chipping away at the broader ‘combined algorithm selection and HPO’ (CASH) problem [532]. Since then,

this undertaking has evolved in many ways, such as by encapsulating neural architecture search (NAS) [196, 197, 460, 560], which now forms the core of the AutoML-subfield known as ‘automated DL’ (AutoDL).

However, as previously hinted, the scope of AutoML – AutoDL included [195] – has itself gradually expanded to encompass the rest of an ML workflow. For instance, data engineering has received its own fair share of research attention. Some works in this space focus on the initial stage of data preparation [311], which may involve sampling and cleaning, while others contribute to the topic of automated feature engineering (AutoFE) [331], covering both feature generation and selection. Then there are phase-agnostic methodologies, such as meta-learning [64, 216, 273, 344, 345, 546], that can theoretically be applied anywhere; these continue to free humans from micro-managing ML systems and supplying domain knowledge. Of course, there is still much further to go. Automating the phase of continuous monitoring and maintenance has recently been highlighted as a crucial prerequisite for truly autonomous machine learning (AutonoML) [308], where systems persist by adapting ML models to changes in data environment [287, 580]. Progress in this space remains relatively nascent [146, 235, 361]. Additionally, rigorous efforts to survey and benchmark state-of-the-art (SOTA) algorithms and approaches [204, 469] are relatively sparse. Nevertheless, the key takeaway from all of this is that, academically, the field of AutoML is rich with activity.

Unfortunately, the translation of pure theory to real-world practice is rarely smooth or one-to-one. That is not to suggest that AutoML has been shunned by industry; to the absolute contrary, a prior dearth of tools to assist with developing ML models – according to the IBM survey [274], one of the top three obstacles for AI uptake – has actually led to an explosion of commercial AutoML services. Alongside numerous open-source packages, these offerings provide businesses plenty of options to choose from, as of the early 2020s, should they wish to apply ML approaches to problems of interest. Yet a healthy scepticism remains warranted, especially where source code is confidential and promotional material is inherently biased. It cannot be assumed that AutoML algorithms and architectures, developed in experimental environments that are well-controlled and sanitised, will deliver optimal outcomes once applied within messy real-world contexts. Certainly, the academic case studies that exist [423, 425, 540], evaluating one or more AutoML solutions within particular industrial domains, are too few in number to make broad claims. So, it is worth asking the question: how are publicly available SOTA AutoML tools and services performing with respect to the demands of industry?

The notion of ‘performant’ ML must be central to any pioneering survey that grapples with this question. In most academic research, performance is usually gauged by purely technical metrics, such as model accuracy and training efficiency. The focus is on how well a computer, in the absence of any human, can generate predictions/prescriptions via ML techniques. On the other hand, industrial contexts are much more human-centric, where stakeholders may have a diversity of interests and obligations; the outcomes and impact of an ML application may only be very loosely correlated with technical performance. Importantly, such matters cannot be ignored by academic AutoML researchers either, as stakeholder requirements can affect the very foundations of algorithms and architectures. For instance, a need for interpretability may force ML model-selection pools to be constrained, a focus on fairness may require mechanisms for bias mitigation, and so on.

Simply put, the technological emergence of AutoML is driven by stakeholder need and the human-computer interaction (HCI) required to service it. Correspondingly, it is impossible to gauge the current state of AutoML technology, especially in terms of whether it can support the needs of industry, without the careful development of an assessment framework anchored by a comprehensive set of HCI-weighted criteria for ‘performant ML’. Certainly, the absence of such a systematic appraisal may not only obscure future directions for progress but, if deficiencies are not

identified, may also have an eventual chilling effect on technological engagement, especially in the case of unmet expectations.

With all that stated, the primary goal of this review is to present a comprehensive snapshot of how AutoML has permeated into mainstream use within the early 2020s. In contrast to two associated monographs that examined fundamental algorithms and approaches behind AutoML/AutoDL [195, 308], this work surveys both their implementation and application in the context of industry. It also defines what a ‘performant’ AutoML system is – HCI support is valued highly here – and assesses how the current crop of available packages and services, as a whole, lives up to expectation. To do so in a systematic manner, this review is structured as follows. Section 2 begins by elaborating on the notion of an ML workflow, conceptually framing AutoML in terms of the high-level operations required to develop, deploy and maintain an ML model. Section 3 uses this workflow to support the introduction of industry-related stakeholders and their interests/obligations. These requirements are unified into a comprehensive set of criteria, supported by methods of assessment, that determine whether an AutoML system can be considered performant. Section 4 then launches the survey in earnest, assessing the nature and capabilities of existing AutoML technology. This begins with an examination of open-source AutoML packages; some of these are tools dedicated to a singular purpose, e.g. HPO, while others are comprehensive systems that aim to automate a significant portion of an ML workflow. The section additionally investigates AutoML systems that are designed for specific domains, as well as commercial products. Subsequently, Section 5 assesses where AutoML technology has been used and how it has fared. Academic work focusing on real-world applications is surveyed, as are vendor-based case studies. All key findings and assessments are then synthesised in Section 6, with commentary around how mature AutoML technology is, as well as whether there are obstacles and opportunities for future uptake. Finally, Section 7 provides a concluding overview on the technological emergence of AutoML.

2 THE MACHINE LEARNING WORKFLOW

Many academic works have presented diagrams that attempt to encapsulate the high-level operations of ML within one consolidated workflow [198, 203, 257, 565], which we henceforth refer to as an MLWF. One early forerunner in this endeavour, though not exclusive to ML, is the popular CRoss Industry Standard Process for Data Mining (CRISP-DM) model [497], and several recent efforts have built upon this basis, e.g. by additionally considering quality assurance [520]. In this section, we extend this model further, diverging where necessary, to align even closer with the modern practices of data science. Such a summary will not be unfamiliar to academics and practitioners of data science, and many MLWFs found in AutoML literature are indeed similar, often only expanding/compressing one or more aspects. Nonetheless, if this monograph is to grapple with the notion of performant ML, particularly within organisational settings that operate beyond pure experimental research, a robust characterisation of an MLWF is required.

Fundamentally, many papers that depict MLWFs agree that there are certain standard phases of ML operation, as captured by the ‘ML Model Development Workflow’ component of Fig. 1. Specifically, a typical ML application will flow from ‘Problem Formulation & Context Understanding’ through ‘Data Engineering’ and ‘Model Development’ to ‘Deployment’. Some MLWFs also incorporate some form of ‘Monitoring & Maintenance’, although this is often presented almost as an afterthought. An academic focus on one-and-done projects, as well as the computational expense of developing modern DL models, means that the challenge of dynamically changing data environments is often ignored, either negligently or deliberately. However, there is a growing awareness within industry that persistent learning is essential, and we thus highlight ‘Monitoring & Maintenance’ as a unique phase within Fig. 1. Indeed, while many MLWFs, such as a CRISP-DM representation [497], provide double-headed arrows or other depictions of circularity between the

first four phases, we associate that continuum of updates with the ‘Monitoring & Maintenance’ phase. Granted, development during an ML project is frequently iterative, with previous phases of operation being revisited prior to deployment, but the primary intent of the ‘ML Model Development Workflow’ is to move forward and bring an ML solution to production. In contrast, it is the intent of ‘Monitoring & Maintenance’ to continually reassess and keep that ML solution relevant, even if – the dashed lines in Fig. 1 hint that this is rarely an academic concern – an ML problem must be partially reformulated while keeping its present solution online.

Now, despite MLWF commonalities in the literature, it is essential to emphasise that perspectives are not universal, and academia often ignores matters relevant to business applications. For instance, several core AutoML papers ignore the deployment phase outright [198, 204, 257]. Others do not dwell on this phase, associating it with the production of predictions [203, 582] or, via the display of prominent social media and tech company logos, suggesting that organisations are interested in this facet of ML [565]. Detail is scant. In contrast, it is noteworthy that, when presenting MLWFs on websites, AutoML vendors frequently elaborate on aspects of deployment [457] and, also neglected by academia, monitoring and maintenance [59, 365]. As already discussed, this is a matter of focus; academia prioritises the development of high accuracy models, whereas industry cares equally, if not more, for sustainable operation.

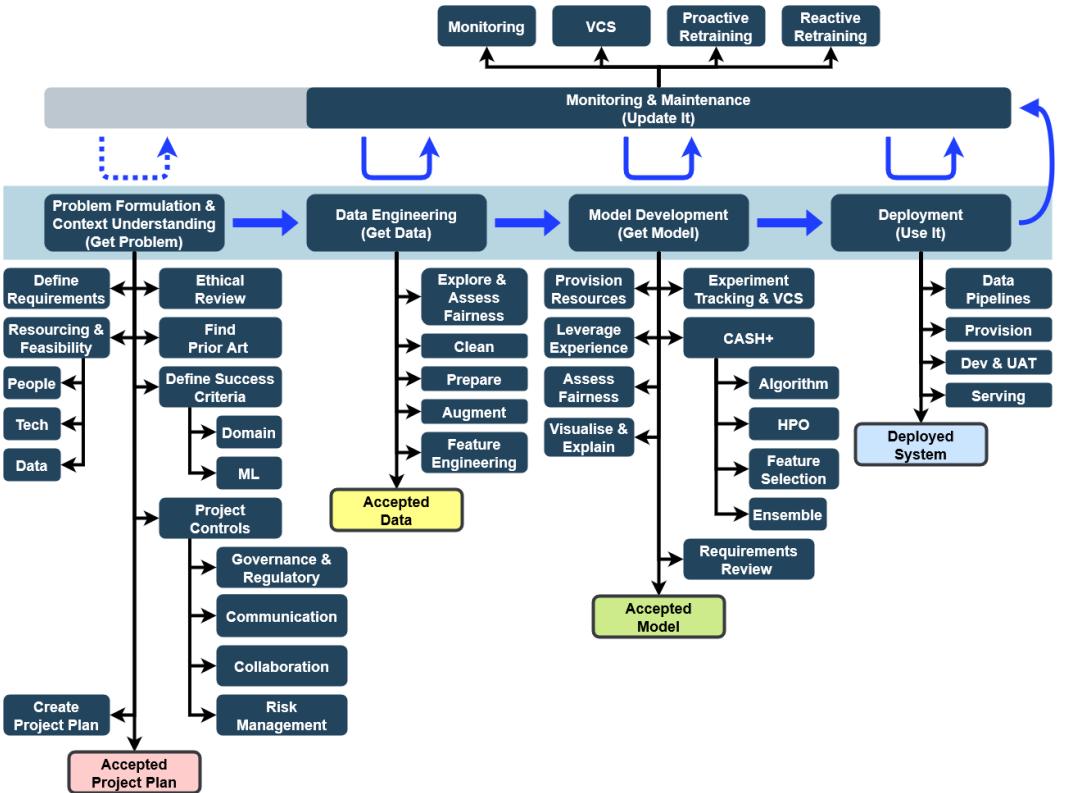


Fig. 2. Key tasks within an overarching MLWF. The lighter-coloured boxes represent the outcome of an MLWF phase that is propagated onwards.

Given this preface, we now present a deeper dive into the granular tasks that commonly constitute an MLWF within a business environment, as displayed in Fig. 2. Admittedly, these tasks are unlikely to be exhaustive for every ML application imaginable. Individual tasks may also be unnecessary in various settings, even if good practice will likely still involve consideration ahead of rejection. For instance, a financial project involving the approval/denial of credit via algorithmic means demands a greater contemplation of bias and ethics than is likely needed for the manufacturing-based prediction of machinery faults. All the same, the breakdown in Fig. 2 is sufficiently informative to support a survey of ML tools and the extent to which they automate key tasks.

The first MLWF phase of **problem formulation & context understanding** seeks to establish an agreeable plan of action, accepted by all relevant stakeholders, for undertaking the rest of an ML project. Although this certainly includes academic elements of developing/acquiring expertise around a problem context and accumulating sufficient topical knowledge to support an ML effort, much of this stage involves largely organisational considerations that ensure the ML project is appropriately defined, scoped, and resourced. First, an organisation must establish its project requirements, which are associated with why the ML application is being undertaken in the first place. For example, a business may decide to leverage ML in predicting users at risk of churn so that its customer support teams can intervene and prevent this from occurring. With these requirements formalised, this is then a good opportunity to begin considering ‘prior art’. In modern times, this might include academic references, reputable blog posts, and similar work performed internally within the organisation. Prior art can indicate how tractable the ML problem is, how best to approach solving it, and what kind of performance can be expected. At this point, stakeholders can also determine whether a SOTA deep neural network is required or whether a simpler approximator, e.g. a linear regressor, is sufficient for the established requirements. In either case, whether appetite leans towards code reuse or pushing the frontiers of ML, resourcing and feasibility checks typically ensue. Of course, the initial search for prior art already involves ensuring an ML project is conceptually sound, but this collection of sub-tasks covers other logistical matters. These considerations include identifying/acquiring people to undertake the work, technological tools to assist them, and raw data sources to form the basis of modelling work. As part of establishing a project plan, an organisation must also define how one can know that the project requirements have been satisfied, i.e. its success criteria. Determining this will generally attempt to marry organisational/domain factors with technical ML objectives and outputs. For instance, a churn-concerned business may decide that, within acceptable timelines and on the balance of projected costs and rewards, a precision score of 85% may be a satisfactory outcome. Finally, an organisation must generally establish appropriate project-management controls to support a greenlit ML application. This task includes considering governance and risk management, e.g. delegating access permissions, task responsibilities, the authority to approve work, and so on. Additionally, an organisation will typically use this phase to decide on communication channels and collaborative tools for key team members while also setting expectations and reporting processes. Furthermore, the finalisation of a project plan will often include more detailed planning and project-management artefacts, such as Gantt charts or critical paths, but these nuances of organisational practices are too varied to generalise.

The second MLWF phase of **data engineering** involves the initial exploration, processing and enhancement of data for modelling purposes. It is often professed that this stage takes up a significant portion of working time for any data scientist. Indeed, a 2016 CrowdFlower report [168] claimed that the percentage was as high as 80%, summing ‘collecting data sets (19%)’ with ‘cleaning and organising data (60%)’. However, providing limited details on the number of respondents and methodology used, the report has been contested by other surveys, despite the widespread mainstream adoption of its claims. For instance, the ‘2018 Kaggle Machine Learning & Data Science

Survey’ [294], with 23859 responses, yielded 11% for gathering and 15% for cleaning. Of course, caution is still necessary when assessing outcomes from an open and uncontrolled survey based on volunteered self-reporting. The Kaggle survey responders were diverse in occupation, e.g. including students (20%) and software engineers (13%), were dominantly skewed towards an early career, i.e. 60% under 30 years of age, and heavily represented the United States (19%) and India (19%). A 2020 version of the survey [295] could not reaffirm these claims as the question was absent; in 2020, the focus was on tools, techniques, and other questions regarding respondent skills and employment. Elsewhere, an Anaconda survey, ‘The State of Data Science 2020’ [74], found worktime proportions of 19% for ‘data loading’ and 26% for ‘data cleansing’. It had a smaller sample size of 2360 respondents, but these appeared to be more of a professional data-science background. Regardless, whether at 80% or a more moderate value, the amount of time that goes into data engineering is not insubstantial.

Regarding the task breakdown in Fig. 2, a common first step in data engineering is exploratory data analysis (EDA). Data scientists will frequently assess the numerical properties of their datasets, visually inspecting graphical representations and identifying quality issues, e.g. missing values and anomalies. Different problem contexts will shape exactly what is undertaken in this section, and many helpful guides are available in many books and blogs [385, 497, 499, 544, 556]. Crucially, this is also an appropriate time to assess bias and fairness within the data inputs themselves; see Section 4.1.2. Traditional EDA guides often ignore this facet, but recent years have seen a surge of attention and concern around ML and trustworthiness. Correspondingly, matters of bias and fairness are some of the key criteria for performant ML outlined in Section 3.2. Regardless, once EDA equips a data scientist with a sufficient understanding of a supplied data environment, they are then usually able to modify the data ahead of ML modelling. Nomenclature and ontologies vary across the literature for the tasks involved in modifying data [125, 231, 461, 545], but here we settle on the sequence of cleaning, preparing, augmenting, and feature engineering. Specifically, we define cleaning in relation to handling erroneous data, while preparation involves formatting input data so that an ML algorithm can access its information content. Accordingly, the imputation of missing values is treated here as a cleaning task, while one-hot encoding categorical variables could be considered a preparatory step. Data preparation also includes scaling, standardisation, and any preprocessing related to data type or structure, e.g. handling timestamps or freeform text. Then there is augmentation, where relevant ancillary data sources are joined to the current inputs. Within this monograph, the novel data is considered entirely external, not engineered variations of existing data as some DL literature may define it [195]. As an example of such augmentation, timestamped store-utilisation data used to predict foot traffic to a retail shop may become more informative when associated with public weather data. Finally, feature engineering aims to transform existing variables in different ways, all in the hope that informative signals in the data may surface. This task receives plenty of academic attention as it is arguably the least straightforward process.

Again, we stress that there are differing views in the data science community on how to categorise and arrange all these data-engineering tasks. For instance, while we consider one-hot encoding as a form of data preparation here, it is valid to argue that new features are being engineered, i.e. new columns are being added to tabular data. Feature engineering is an even more complex notion to organise, especially in light of the standard filter/wrapper perspective discussed in an earlier AutonoML review [308]. For instance, if feature selection is done before considering an ML model, i.e. in filter style, it would seem to belong in the data-engineering phase. An example is filtering out features according to the outcomes of Pearson correlations or chi-squared tests. However, if features are selected based on whether they improve the performance of a specific ML model, i.e. in wrapper style, the algorithms that do so may be pipelined as part of the model-development phase. An example is exclusion based on feature importance scores that the Random Forest ML algorithm

provides. Ultimately, we highlight these nuances but persist with the arrangement in Fig. 2. The tasks listed under data engineering are, in aggregate, holistically encompassing; minor variations do not significantly perturb the framework proposed by this monograph for assessing performant ML.

The third MLWF phase of **model development** is arguably the core of an ML application, proceeding once a cleaned, prepared and enhanced dataset is in hand. As the task breakdown suggests, a large-scale ML application often involves several preparatory steps. Beyond setting up requisite model-training infrastructure, an organisation may need to establish tools that track experimentation and model versioning while trials are being undertaken. Eventually, though, there comes the actual process of fitting a mathematical model to a desirable function. Even in the present time, AutoML literature predominantly focusses on selecting an ML algorithm and tuning hyperparameters, i.e. solving the CASH problem [532], so we use the term CASH+ to be more encompassing of ML solutions. Specifically, some ML applications and AutoML packages will bundle feature-selection or predictor-ensembling methods as part of an ML pipeline they pursue. Also, it is worth noting that many researchers have investigated how to ‘leverage experience’ when solving some subset of the CASH+ problem, e.g. using opportunistically derived predictor rankings to constrain search spaces for ML algorithms [309, 413]. The leveraging of experience thus covers the technical area of meta-learning [64, 65, 308, 344–346, 546], but it also refers to leaning on domain experts for assistance in contextualising, understanding and making better decisions with preliminary model results [287]. We note that prior experience can inform any phase of the MLWF, but model development has most often been the focus of such research and development.

Notably, various metrics and visualisations may be produced throughout an ML application to understand the modelling activity. However, once a preliminary final model has been produced, contextually driven tasks may be carried out to uncover what work was undertaken and what was ultimately produced. Here, revisiting the notions of bias and fairness is crucial in producing models that meet trust-based requirements. Again, Section 4.1.2 elaborates on this topic, e.g. on assessing data versus a model, but it is sufficient to note here that an ML model can be assessed and remedied if it proves problematically biased or unfair. Another topic that is also deeply entwined with trust in ML is that of explainable AI (XAI). Different organisational stakeholders will have different needs concerning XAI, as discussed in Section 3.1. However, the model-development phase of an MLWF is an appropriate time to understand how an ML solution was generated and why it produces the outputs that it does. This process often includes visualising performance metrics, global explainability artefacts such as feature importance, and the drivers of individual predictions/prescriptions. Additionally, scenario planning tools may be made available here to understand the impact of potential interventions on these metrics and visualisations. Finally, the generated ML model and all related artefacts, e.g. XAI items, are assessed against the initial requirements for the project. Some MLWFs consider this under an ‘evaluation’ phase that often refers to simple technical metrics, e.g. model accuracy and training time, but industrial applications often have many more requirements that an ML solution must satisfy before it can be approved for deployment. The dearth of academic contemplation in this space is a primary motivation for this monograph and its comprehensive framework for performant ML.

The fourth MLWF phase of **deployment** and the fifth MLWF phase of **monitoring & maintenance** incorporate all the tasks required to turn an experimental modelling project into a sustainable productionalised system embedded within some organisational process. Broadly, they encompass much of what is nowadays referred to as ‘MLOps’. Specifically, one of the first steps in deploying an ML solution is to convert all relevant data transformations into a robust pipeline. This pipeline must connect raw data sources to the ML models running inference, and this transmission must be suitable for context, e.g. batched, real-time, constrained for the internet of things, and

so on. Computational resources must also be provisioned to host the ML solution and support its inferences. Now, one common practice in software engineering is undertaking user acceptance testing (UAT) [124, 155], ensuring that a system meets the expectations of end users. Such practices are similarly relevant in organisational applications of ML where model results are intended for widespread consumption.

Eventually, once sufficiently tested and fully provisioned, an ML solution can be appropriately placed in production. However, the performance of a static ML model can decay over time due to environmental dynamics, such as data drift, concept drift [361, 373, 553, 581], and other system disruptions. Ideally, monitoring processes are established for all metrics of relevance, including technical performance, data properties of interest, and variables that assess deployed ML solution outcomes, e.g. bias and fairness metrics. An adaptation process can then be triggered automatically or after manual review if a monitored metric dips below a threshold. The simplest form of adaptation is the full retraining of an ML model, but there are numerous redeployment techniques, e.g. blue-green deployment, canary deployment [56], and many more [471]. Of course, as change occurs within a deployed solution, implementing a version control system (VCS) is advisable to mitigate the risks of unforeseen issues with updates; it is always helpful to roll back to prior safe versions.

In conclusion, we have schematised a general MLWF and, via Fig. 2, elaborated on the typical tasks that an organisation may carry out in running an ML application. As already mentioned, not everything here will be universally relevant. Some ML efforts are solely angled towards uncovering data insights, and several vendors of automated tools operate within this market. However, standard organisational use of ML involves taking a trained model from problem conceptualisation to production, generally relying on consistent delivery of business value via long-term consumer engagement. Thus, the more tasks within the full MLWF that an AutoML tool automates, accounting for the standards of performant ML, the more appealing it is to industrial stakeholders. In fact, if the extent and degree of automation sufficiently encompass the monitoring & maintenance phase, such that an ML model learns persistently and autonomously, the era of AutonoML will have truly arrived [308]. In short, the MLWF framework helps anchor assessments of AutoML tools and their scope; this monograph will detail the broad spectrum of existing AutoML services. However, to honestly assess the value of modern AutoML to industry, a simple question needs answering: who cares?

3 PERFORMANT MACHINE LEARNING

Nuances aside, a common maxim in economics is that “demand creates supply”. Need and desire drive interest, investment, and innovation. In complement to this rule, a product does not survive and thrive in an industrial setting without serving a purpose and generating a positive impact. Indeed, while the clientele for ML technology may be extensive, it is also finite. Over time, competition for stakeholder attention and engagement is an optimisation process, impelling tools and systems that support ‘performant ML’ to bubble up into prominence. Of course, as with biological evolution, this process is not perfect; odd and even detrimental ‘genotypes/phenotypes’ could arise and become entrenched within a ‘population’ of AutoML services. Nevertheless, by and large, the quest for performant ML is the driving force behind AutoML technology.

So, what is performant ML? Who decides? Traditionally, academia has a very narrow scope when defining ML performance, as exemplified by typical textbooks on the topic [281]. Its focus is predominantly on metrics that judge how well an ML model approximates a desirable function, such as classification accuracy and various ratios involving a confusion matrix, e.g. sensitivity, specificity, recall, area under the curve (AUC), and the F_1 score [213, 444]. Occasionally, these metrics may also be paired with information on how long it took to optimise them, i.e. the time costs of ML model training. These considerations are particularly pertinent within hardware-conscious

research beyond pure algorithmic advancement, where performance measures must be mindful of infrastructure [200, 486]. However, in industrial and applied contexts, it is reasonable to question whether these alone are the only metrics that matter. One experimental investigation [582] asserts that most CASH procedures perform reasonably similarly, at least in technical terms, concluding that the suitability of deploying AutoML frameworks for real-world use cases should consider factors beyond those of typical academic concern. This review supports such a perspective.

Thus, having already detailed *what* is typically involved in the practice of real-world ML, this section delves into *who* would care for automating an MLWF and how they would judge the overall process as performant. Specifically, in Section 3.1, we outline the key stakeholders involved in ML tasks within an industrial context, detailing their needs and the potential benefits they may realise from AutoML. Then, in Section 3.2, we propose a comprehensive set of criteria by which the practical application of an MLWF can be evaluated. Finally, in Section 3.3, we synthesise these considerations to assess the role that industry currently expects AutoML to play in supporting performant ML.

3.1 Key Stakeholders and Requirements

Before establishing criteria for performant ML, one must first understand who would be involved in applying an MLWF within the context of industry. Thus, Section 3.1.1 details the primary stakeholders that would be, and presently are, most engaged with the usage and outputs of AutoML. Essentially, the discussion lists what these groups care about and how AutoML may factor into meeting these needs. Secondary stakeholders are also briefly noted in Section 3.1.2, as their desires will likewise influence the continuing evolution of AutoML technology.

3.1.1 Primary Stakeholders. When discussing this collective, there are essentially two subcategories: data scientists and other technical users. The latter term encompasses those who have the potential to use AutoML technologies directly but are not expert data scientists. There naturally exists a spectrum on which such potential users can fit, from the trained professional to the technophobe with limited computer literacy. This type of stakeholder, therefore, represents users who may lack technical skills but are valuable to an ML application due to their increased domain knowledge. However, all primary stakeholders are still defined here by their close participation in ML analysis rather than involvement with any general infrastructure/architecture. Thus, we exclude IT and generic data-engineering roles that would not typically interface with an AutoML system. Also, it is understood that modern organisations are fluid, and roles may be transitional or hybrid, but the following categorisation should still be sufficient in encompassing the employment space related to ML applications.

Data Scientists. This group of technicians represents the most prominent core stakeholder in AutoML technology. Granted, democratisation is a central aim of the AutoML endeavour, but such a process is gradual and ongoing. Realistically, data scientists remain the primary users interfacing with ML tools, meaning their needs dominate any discussion of the requirements that AutoML must satisfy. Such considerations can be summarised as follows:

- **Efficiency.** One of the key appeals of automation is that, ideally, it speeds up processes substantially. After all, machines are generally better than humans at formulaic tasks, maintaining high levels of consistency and endurance. The resulting procedural fluidity can be valuable to data scientists in two main forms.
 - **Operational Efficiency.** This concept refers to saving time and effort expended by the staff of an organisation when managing a technical process. In context, data scientists will often form their own personal workflows for expediently tackling ML problems, but the manual application of these can still have high starting costs. For instance, template-driven

approaches may need to be adjusted and tweaked per ML problem, while those who have not invested in such practices will likely need to code from scratch. Accordingly, there are many points along an MLWF where the automation of existing practices can speed up operations substantially. Crucially, none of these involve the science of ML; operational efficiency merely relates to the logistics that support an ML application.

Of course, the desire to streamline work processes is common throughout industries focussed on maximising productivity and minimising cost. With the high salaries commanded by data science talent, there is an organisational impetus to mechanise operations that are high-volume and low-value, e.g. via robotic process automation (RPA), so that data scientists are employed where their technical skills will have the greatest impact. Indeed, an IBM survey [274] found that, alongside saving costs (58%) and freeing valuable time for employees (42%), driving greater efficiencies (58%) was a top reason for businesses using or considering automation tools. Nonetheless, interviews with data scientists [550] indicate that employees also appreciate the prospective benefits of increased operational efficiency that AutoML may offer.

Admittedly, because the interplay between automation and ergonomics is complicated, it can be challenging to draw bounds on the scope of AutoML under this requirement. For instance, the automation of project maintenance via Git, a VCS that a 2021 Stackoverflow developer survey [518] found was used by over 93% of 80000 respondents, will have had an undeniable impact on supporting streamlined ML. Automating collaboration between team members is another nuanced driver of efficient ML applications, almost ubiquitously ignored by academia, and data scientists have expressed a desire for tools that enhance communication and associated productivity [432].

- **Technical Efficiency.** This concept refers to a technical process running faster or with fewer resources. In context, this generally relates to the time and memory footprint involved in developing, deploying and maintaining an ML solution. At one implementational level, data scientists may appreciate efficiencies arising from reducing the time complexity of algorithmic processes, e.g. by vectorising looped tasks. However, the development and release of theoretically novel algorithms can also significantly impact the speed of ML. In fact, the field of modern AutoML launched on the back of expedient ML model selection, as reviewed previously [308]. Accordingly, while many data scientists will have some degree of reticence when adopting unfamiliar techniques, sufficient technical efficiency, exemplified by the history of convolutional neural networks, can overcome this barrier to uptake.
- **Technical Performance.** While this review does not focus on metrics related to the standard ‘correctness’ of ML models, this is primarily due to how heavily the concept has been discussed elsewhere. Certainly, it is inescapable that data scientists and dependent stakeholders seek ML solutions that are sufficiently representative of some desirable function, often a ground truth. However, while academia often seeks to push the limits of model validity, the costs and diminishing returns can be prohibitive within an industry setting. Research circles have noted such concerns, with some discussing how good is good enough [222, 332, 531].

That stated, AutoML is yet to shrug off an association with mixed technical performance. While developers of AutoML packages tend to promote the predictive power of their mechanisms and frameworks, independent benchmarks vary. Some suggest automated techniques achieve mediocre results compared to humans [582], some are more favourable [362], and yet others sit in the middle, e.g. stating that AutoML performs equal or better on 7 out of 12 tasks [249]. Although not an academic work, a recent poll on KD Nuggets [438] likewise indicates this subdued outlook among data scientists more generally, asking the following

question: “How well do current AutoML solutions work, in your opinion?” With a scaling from 1 to 5, i.e. ‘badly’ to ‘super-human’, the poll returned an average score of 2.4. However, there was a notable difference in average scores between those who tried AutoML (2.56) and those with only preconceptions (2.29). Moreover, no consideration was given to which tool was used. Ultimately, although the technical efficiency of AutoML does make it easier to reach improved technical performance, which is generally appealing to data scientists, it is not yet clear whether improved model validity should even be a primary selling point of AutoML.

- **Methodological Currency.** Essentially, other things being equal, data scientists prefer to operate as close to SOTA as possible. However, the modern AI field is progressing fast, and advances are constantly being made across the entire standard MLWF. This evolution is not a monolithic affair either. For instance, novel auto-augmentation data-engineering methods will likely have little reference to new deployment techniques for field-programmable gate arrays (FPGAs), even if both may be relevant to a DL application [195]. Accordingly, the so-called ‘unicorn’ data scientist that is an expert in all the niche skills and topics across an MLWF is extremely rare, if not outright nonexistent [112, 205]. Even keeping abreast of ML modelling alone can be challenging, noting that the industry-leading scikit-learn library – it has Sklearn as an alias – has, in version 0.24.2, 191 available estimators in the form of classifiers, regressors, clustering methods, and transformers [41]. Given the already amorphous role of a data scientist at present [396], a standard representative of this stakeholder group is likely to have varying degrees of expertise in different algorithms, HPO techniques, and other technical processes. Thus, AutoML can provide value to a data scientist by supporting access to unfamiliar techniques, whether brand new or renascent, subsequently improving operational and technical efficiencies.
- **Ease of Use.** Regardless of efficiency, a technical or operational process in an ML application loses its appeal if a data scientist cannot interface with it effectively. Of course, assessing ease of use for any computational tool is somewhat subjective, depending on who is using it and what they are using it for. Poor design or dependence on overly specialised skills can immediately hamper uptake. However, data scientists also vary in their personal preferences. Some are comfortable with code, some may seek a command-line interface (CLI) for its perceived simplicity and accessibility, and yet others will desire a graphical user interface (GUI) to interact with technical products [209]. This notion of a convenient user interface (UI) within the context of AutoML is reviewed deeper elsewhere [313].

Beyond personal preferences, technical tools score higher with data scientists if they are fit for purpose, integrating well into an existing MLWF and addressing specific use cases. For instance, when using ML to predict purchasing propensity in e-commerce, a technical stakeholder would likely appreciate any convenient method of accessing and manipulating data related to sales, customer demographics, website activity, etc. Granted, this lies within the purview of automated data engineering, but the emphasis here is on effectively configuring operational/technical processes for a specific use case. As another example, consider a data scientist working with a recommendation engine. Rather than composing a standard error measure over all samples, the stakeholder may prefer to work with a precision evaluation on some top-N recommendations [241], an exponential decay that notes users are less likely to pick items down a ranked list [130], or some other non-standard metric [261]. As of the early 2020s, most AutoML packages strive to be as generally applicable as possible, but Section 5.2 does provide limited examples of modern tools that conveniently specialise.

As a final note, while an expansive set of programming approaches does exist, data scientists have clustered around certain popular open-source languages and frameworks for

ML. Developing in these spaces automatically improves ease of use. Specifically, a 2019 KDnuggets Software Poll [437] highlighted Python and R as preferred languages from 2017 to 2019, although with a yearly decrease for R. It also listed Keras, scikit-learn and TensorFlow as popular ML libraries. These results were further corroborated by the Kaggle State of Machine Learning and Data Science report in 2020 [295], which was notably more focussed on practising data scientists. This report ranked scikit-learn, TensorFlow and Keras as the top ML frameworks, while also revealing that the top three languages regularly used by respondents – multiple selections were allowed – were Python (15530), SQL (7535), and R (4277). Additionally, at a ratio of 14241 to 1259, respondents recommended Python over R as the first language an aspiring data scientist should learn. Admittedly, it is unclear whether these programming languages will maintain a stranglehold on the mainstream in the long-term, with newer entrants like Julia acquiring small but growing fanbases [534].

- **Explainability.** Understanding how an ML solution came to be and why it says what it says has surged in importance within academia over the last several years. However, this is not an unfamiliar requirement to data scientists who regularly interact with business stakeholders; part of the job is translating work and outputs into an understandable format. Unsurprisingly, an ability to communicate well is often cited as a core component of the skill profile for such a technician [111, 166]. Indeed, a seminal 2012 Harvard Business Review article defined a data scientist as “a hybrid of data hacker, analyst, communicator, and trusted adviser” [185]. Likewise, an IBM survey [274] found that 91% of businesses using AI say their ability to explain how a decision was arrived at is critical. Accordingly, many data scientists will appreciate computational tools that provide insight into what exactly they do. Satisfying this requirement boosts operational efficiency, but it also improves trust in an ML solution. This desire for transparency around data and ML modelling has been corroborated by recent interviews [198], albeit limited both in number and to students. Another set of interviews surveying 20 professional data scientists, limited by their association with the same organisation, likewise found a consensus need to surface what was done, e.g. what algorithms or preprocessing techniques were used, and how it was done, e.g. what hyperparameter values were chosen. Clearly, explainability and associated trust are essential issues to stakeholders, and a deeper dive into these topics is available elsewhere [313].

Analysts. This group of primary stakeholders is the first that can be considered to encompass ‘other’ technical users. Analysts typically have moderate exposure to techniques and technology involving data, possessing strong skills with popular business software, e.g. Microsoft Excel, as well as reasonable fluency in SQL and some exposure to R and Python [182, 511]. However, data visualisation will generally be enacted via popular software applications such as Tableau and PowerBI [230], rather than a technical coding library. Of course, this is a generalisation as the spectrum of proficiency is broad. Now, while the requirements of an analyst cover the same scope as a data scientist, priorities tend to differ. Rather than technical efficiencies, given that analysts do not commonly practice ML and are thus unconcerned with optimising such processes, **ease of use** becomes particularly important. Additionally, the core job function for an analyst requires proximity to business stakeholders, so one would also seek tools with a high degree of **explainability**. Unlike data scientists, who lean towards understanding ML processes to instil confidence in the rigour and validity of an ML solution, analysts generally need explainability to bridge the gap between technical and non-technical stakeholders, as required by business intelligence/analytics (BI/BA) roles [140, 188, 561]. Naturally, analysts working with AutoML tools are still likely to desire strong **technical performance** from an ML solution, but their standards will differ from that of a data scientist, who is far more likely to have benchmarked such metrics and is aware of what is currently

SOTA. Essentially, perspectives will be ‘anchored’ differently depending on stakeholder experience with a technology thus far [225].

Business Users. This group is yet another step removed from the technical expertise generally required for direct ML involvement. Admittedly, many existing AutoML vendors market themselves as operable by ML novices; one could propose that any accountant, lawyer, line manager or other business stakeholder could participate in loading data, undertaking ML and deploying performant solutions as part of their decision-making workflow. However, this is a lofty ideal even before considering the survey results in the rest of this review. Several factors also complicate matters. Firstly, business users are unlikely to have confidence using ML tools, even if ease of use is outstanding. Indeed, despite visualisation tools and other methods for supporting explainability, AutoML-assisted technical operations, e.g. the deployment of an ML solution, are likely to remain daunting. Secondly, this type of stakeholder is unlikely to consider direct ML involvement within their remit. The creation and management of analytical models typically fall to data scientists or analysts, and any organisational dearth of expertise here is probably better met by hiring talent to fill the gap. In fact, it has been argued that enabling non-technical users to run an ML application may even be harmful [126, 187, 324]. Nonetheless, business users remain critical stakeholders in an ML application, often acting as both the driving force and beneficiaries of its outputs. Certainly, manager functions within business units are those that commission bodies of work to be completed and expect results from that expenditure of effort. Thus, **technical performance** and **efficiency** are paramount, although more through a return-on-investment (ROI) lens that considers staffing time and organisational resources. Conversely, any characteristic around interfacing with an AutoML system, e.g. usability or explainability, is likely to be less of a direct concern, as organisations will usually rely on data scientists and analysts for reporting. Indirectly, though, business users will still benefit from such facets, as they require confidence in an ML product and its alignment to business objectives.

Deployment Technicians. This group encompasses those who move experimental ML solutions into production. In smaller organisations, this role may blend with that of a data scientist, but larger businesses or those dealing with more mature technical functions often have a separate department dedicated to the policies, procedures and processes behind deploying technical products. Normally, once an ML model has been created and is ready for consumption, it sits as an object within the same technical scope as other business applications, i.e. ingesting data, writing data, and interacting with other systems. How this solution is consumed will vary based on the intent behind an ML application, but modern business practices have established standardised roles focussed on deployment. Traditionally called DevOps [97], these functions, if specific to ML, are starting to be referred to as MLOps [183, 393, 501]. For those tasked with associated responsibilities, an ML pipeline is usually considered sacrosanct; its experimental accuracy is unquestioned, and matters such as explainability are irrelevant. Instead, more technical considerations related to infrastructure are essential, and these have been the focus of various studies [167, 200, 360, 486, 512, 513]. Like data scientists, deployment technicians care about **efficiency**, albeit in matters of inference and maintenance rather than model training, and they would also seek **methodological currency** from AutoML packages, given how quickly hardware and deployment techniques can evolve, e.g. FPGAs and federated ML. Other considerations relating to an ML application include the ability to scale well [582], continuously update [563], handle dirty data [246], and adapt robustly to concept drift [373].

3.1.2 Secondary Stakeholders. This collective is not generally invested in a specific ML application like the primary stakeholders listed above. However, the category remains important in discussing AutoML, as its constituents have roles and responsibilities that will both impact and be impacted

by intensifying uptake of associated technologies. Indeed, disregarding the requirements of the following organisational stakeholders would render an incomplete understanding of the dynamics that drive AutoML adoption in enterprise use cases. Importantly, as before, we do not consider specific job titles here due to the fluidity of definitions within the modern workforce, instead focussing on organisational roles and responsibilities.

Corporate Management. This group of secondary stakeholders encompasses the finance, human resources and other management units within an organisation, save for those included within a separate ‘risk and governance’ subcategory below. Crucially, for any organisation in the private sector or elsewhere, the allocation of finite resources is a strong motivator and constraint for business decisions and activities; corporate management is closely tied to those considerations. Indeed, a recent Boston Consulting Group survey of senior executives at 1034 large organisations [397] found that the most significant driver for responsible AI use related to business benefits, as declared by 40% of respondents. This motivation was followed by customer expectations (20%), risk mitigation (16%), and regulatory compliance (14%). Of course, maintaining the health of a business organisation manifests in diverse requirements of a performant ML application, not all simple and direct. For instance, an IBM survey, previously mentioned [274], found that **explainability** is important to corporate management. In fact, a CapGemini survey reveals that the proportion of executives interested in this area has increased from 32% in 2019 to 78% in 2020 [135]. However, such a stakeholder is typically not interested in understanding a specific ML model; they are frequently answerable to external entities, e.g. customers and regulatory bodies, and thus adopt their interests. Then there is **ease of use**, which corporate management is unlikely to ever avail itself of directly, but investing in accessible ML tools does benefit staff training and acquisition. Granted, other requirements that AutoML could satisfy are more straightforwardly justified. Strong **technical performance** of ML solutions can provide a competitive advantage within an industry and generate revenue. Good **efficiency** can similarly save money, e.g. operational efficiency frees time for existing staff and technical efficiency may save server costs.

Risk and Governance Entities. This group is essentially dedicated to avoiding harm related to business practices, which, in context, refers specifically to the processes and outcomes of an ML application. Given that conditions of uncertainty are inevitable within real-world settings, it is up to these entities, alongside management, to understand and mitigate associated risks. Data governance, for instance, is increasingly of corporate interest, with the biannual Information Governance ANZ group survey finding in 2021 that 64% of organisations had adopted a formal Information Governance (IG) framework, implementing associated policies and procedures, and 74% had IG projects underway or planned for the following year [82]. Two years earlier, only 51% were using a formal IG framework [81]. Evidently, there is a growing consensus that the implementation of an advanced technology should be subject to IG considerations and risk-based oversight. Specifically, ML applications and the automation of their higher-level processes are expected to align with data governance and security practices via controllable access to data, models, and technical functionality. Essentially, the running of an MLWF should be auditable.

Notably, there are many ways related to ML in which risk may arise, and some have nothing to do with its technical processes. For instance, the field of data science is presently beset with significant variability in the skills and preferred approaches of its practitioners. Admittedly, one could argue this is an issue in any industry that is heavily dependent on human expertise, e.g. medicine or law. However, the relative immaturity of industrial data science means that no educational or experiential thresholds are commonly agreed upon to signify that someone is a data scientist [191, 262, 325, 555]. In fact, there is currently a proliferation of online courses and boot camps to assist people transitioning into the field [139], many of debatable quality, and the absence of regulation means that it is not uncommon for prospective employees to simply change their job titles. This

inconsistency in skills and approaches can damage the quality of ML outputs, which is particularly dangerous in high-stakes problem contexts. Moreover, even if every professional data scientist were a genuine expert, the lack of standardisation can still cause issues with reproducibility, which is a prominent concern across all sciences, including ML [116, 245, 435, 467]. Thus, for stakeholders dedicated to risk and governance, AutoML has the appealing potential, in theory, to provide consistency and transparency in the application of ML, establishing a robust baseline in the practice of data science. On the other hand, AutoML packages must themselves have appropriate safeguards for such an ideal to be realised, as ease of use erodes the accessibility barrier that prevents non-experts from inducing errors and possible harm via their ignorance.

Now, when discussing overlapping requirements with a data scientist, a naive expectation is that solid **technical performance** will stimulate trust in an ML solution simply by virtue of generating valid ML predictions/prescriptions. However, the standard application of ML, even with a substantially accurate model, does not consider many nuanced drawbacks that can make an ML solution a poor fit for a real-world context. These nuances can be very subtle, which is why any auditing bodies predominantly require **explainability** from the tools and processes used in an MLWF, if only to seek trust through transparency [535]. Indeed, trust in AI has recently surged in importance within both academic and industrial circles of discussion [198, 519, 535, 551]. Chief among the factors that can threaten trust are issues of bias and fairness, of which the public has become more aware and critical as data science progressively seeps into the lives of common people [410, 475]. The aforementioned IBM report [274] found 87% of respondents professed that “ensuring applications and services minimise bias” is an essential aspect of AI. However, the report also noted that skills shortages and a lack of assistive tools are the most considerable barriers to developing/managing trust in AI. Naturally, this pressure for trustworthy ML has a financial motivation for many businesses, passed on from customers; a recent CapGemini survey of 800 organisations and 2900 consumers [135] found that 71% of the latter want a clear explanation of results and 66% expect AI to be fair and free of bias. This expectation has increased awareness of AI discrimination among surveyed executives, from 35% in 2019 to 65% in 2020. Sure enough, modelling in the literature posits that ignoring the societal requirement for debiasing can adversely impact business demand and associated profits [541]. Additionally, surveys [388], taxonomies [401] and instructional research [175] are accumulating on this topic, sometimes concerning specific fields and applications, such as medicine [236] and hiring practices [184], respectively. Simply put, risk and governance entities are likely to desire increasing capabilities of assessing/managing bias and fairness within ML applications, and this will hold true of AutoML as well.

Finally, we note that this set of requirements, traditionally neglected by frontier ML theory and technology, cannot be ignored for long. While extensive laws lag substantially behind the pace of ML progress, calls for regulation are gaining political traction worldwide [234]. For instance, consideration of the issue has appeared in the Australian 2021-2022 Budget Fact sheets [421], while the US White House is launching a task force, i.e. the National AI Research Resource, with a partial eye to such matters [267]. As some other examples, Standards Australia has recently developed a roadmap related to AI practices [88] and the Office of the Australian Information Commissioner has published a ‘Guide to data analytics and the Australian Privacy Principles’ [420]. Elsewhere, in April 2021, the European Union released a proposal for the regulation of AI among member states [161]. In essence, the societal context in which organisations employ ML is evolving, and business leaders are highlighting trust and explainability in ML as vital for meeting regulatory and compliance obligations [274]. This review will not delve deeper into the international laws and policies being established for AI; the key takeaway is that an MLWF and its automation will be subject to increasing regulatory oversight in the coming years, especially within industries such as finance, law enforcement, and medicine. Accordingly, organisations would likely appreciate

transparency from associated tools and processes, as well as considerations of bias and fairness, should tension between business objectives and regulatory requirements ever arise.

3.2 Unified Criteria

Having established the core requirements of primary/secondary stakeholders that engage with an ML application and associated tools, we can now distil and outline the key criteria by which AutoML in an industry setting can be assessed as supporting performant ML. This proposed framework will anchor the subsequent review of open-source packages, both specialised in Section 4.1 and holistic in Section 4.2, as well as commercial offerings in Section 4.3. To best aid such an effort, each criterion below has been broken down further into several questions and associated scoring methods. The questions have been designed to be answerable with publicly available information, if it exists, i.e. source code and documentation for open-source tools or vendor websites for commercial products. Additionally, these questions are slanted to recognise major challenges that face the ongoing uptake of AutoML technology, with one academic work [340] suggesting that obstacles fall into three main areas: search, technical speed/performance, and HCI. In fact, given that grappling with these challenges is a continuous process, the responses to the proposed questionnaire are not always binary, e.g. ‘no automation’ or ‘full automation’. Convenience functions and features that assist a user with an ML task, which suggest partial progress towards automation, warrant acknowledgement. With that all stated, we now proceed to list the criteria.

- **Technical Performance.** Any AutoML product that supports performant ML will always be judged by certain core metrics, i.e. its potential to set/improve the ‘correctness’ of an ML model. This review aims to extend beyond such considerations, as necessary levels of solution validity differ dramatically between industries, use cases, risk profiles, and organisational agendas. Some businesses can operate at the SOTA frontier, while others are dabbling in ML technologies for the first time. Likewise, 25% accuracy for a music recommendation service may be fantastic, while 75% accuracy for a tumour classification system may be abysmal. In short, predictive/prescriptive ‘correctness’ is undoubtedly essential, but it is far from the be-all and end-all of ML requirements. It is also the criterion that we do not delve into within this assessment framework for performant ML; experimental research is required to validate the technical performance of any AutoML system, and this is out of scope for this review. Such attempts at benchmarking are also already numerous within the literature [114, 190, 237, 536, 582].
- **Efficiency (22 Questions).** As outlined in Section 3.1.1, the pace and cost of running an MLWF are set on two fronts: operational and technical. The former relates to processes that determine how effectively/productively the work of employees can be translated into advancing tasks within an MLWF. The latter relates to the speed and resource consumption involved in developing, deploying and maintaining the technical ML solution itself. Thus, several categories of questions have been established to evaluate how well an AutoML tool assists with overall efficiency. First, there is an assessment of the effort required in tracking and managing experimentation during ML modelling. Second, there is consideration around how easy it is to utilise prior art/work. Packages rate highly on this sub-criterion if they are (1) capable of storing/managing a history of previous ML applications and (2) able to leverage that previous experience for future recommendation, e.g. via meta-learning. Efficient collaboration also aids in awareness of prior art, so evaluating its presence is included in this category. Third, there is a determination of how much effort can be saved along work-intensive portions of an MLWF, i.e. data exploration/preparation, feature engineering/selection, and actual modelling. Because data preparation is a particular time-sink, it merits an extended set of spin-off

questions at this point. Finally, there is an appraisal of AutoML features, e.g. configuration control, that may support technical efficiencies beyond those intrinsically linked to technical performance. We now explicitly list the questions on efficiency.

Table 1. Assessment Framework for Efficiency. Questions: E1-E3.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	Effort in experiment management & tracking	Does it provide a model repository?	E1	0/1	0: No 1: Yes
Efficiency	Effort in experiment management & tracking	Does it provide model VCS?	E2	0/1	0: No 1: Yes
Efficiency	Effort in experiment management & tracking	Does it provide experiment tracking features?	E3	Scale 0:2	0: No 1: Yes for log storage/access, but with limited automation and/or visuals 2: Yes, with automatic log visualisation

The sub-criteria in Table 1 can be mapped to the MLWF in Fig. 2 as follows: E1 to *Find Prior Art* within *Problem Formulation & Context Understanding* and E2/E3 to *Experiment Tracking & VCS* within *Model Development* and *VCS* within *Monitoring & Maintenance*.

Table 2. Assessment Framework for Efficiency. Questions: E4-E6.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	Effort in leveraging prior work & collaboration	Does it offer a template/code repository?	E4	Scale 0:2	0: No 1: Yes, templates/code can be generated by users 2: Yes, templates/code can automatically kickstart projects
Efficiency	Effort in leveraging prior work & collaboration	Does it suggest prior work?	E5	0/1	0: No 1: Yes

Table 2. Assessment Framework for Efficiency. Questions: E4-E6.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	Effort in leveraging prior work & collaboration	Does it facilitate project collaboration?	E6	Scale 0:2	0: No 1: Yes, with basic features such as shared access to folders with project artefacts 2: Yes, with advanced features

The sub-criteria in Table 2 can be mapped to the MLWF in Fig. 2 as follows: E4/E5 to *Find Prior Art* within *Problem Formulation & Context Understanding* and E6 to *Collaboration* within *Problem Formulation & Context Understanding*.

Table 3. Assessment Framework for Efficiency. Questions: E7–E13.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	MLWF Effort: Data Exploration	Does it automatically generate visualisations to assist in data exploration?	E7	Scale 0:3	0: No 1: No, but convenience features are available 2: Yes, to some degree 3: Yes, and with automatic notification of issues or points of interest
Efficiency	MLWF Effort: Data Preparation	Does it automatically prepare data for modelling?	E8	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree
Efficiency	MLWF Effort: Feature Engineering	Does it automatically engineer features?	E9	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree

Table 3. Assessment Framework for Efficiency. Questions: E7–E13.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	MLWF Effort: Feature Engineering	Does it store features for later use by others?	E10	0/1	0: No 1: Yes
Efficiency	MLWF Effort: Feature Selection	Does it automatically select features?	E11	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree
Efficiency	MLWF Effort: Modelling	Does it specify HPO search spaces and algorithms by default?	E12	0/1	0: No 1: Yes
Efficiency	MLWF Effort: Modelling	Does it optimise an entire ML pipeline?	E13	0/1	0: No 1: Yes

The sub-criteria in Table 3 can be mapped to the MLWF in Fig. 2 as follows: E7 to *Explore & Assess Fairness* within *Data Engineering*, E8 to *Clean* and *Prepare* within *Data Engineering*, E9 to *Feature Engineering* within *Data Engineering*, E10 to *Find Prior Art* within *Problem Formulation & Context Understanding*, E11 to *Feature Selection* within *Model Development*, E12 to *HPO* within *Model Development*, and E13 to *Data Engineering* and *Model Development* generally.

Table 4. Assessment Framework for Efficiency. Questions: E8A–E8F.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	MLWF Effort: Data Preparation	Does it automate categorical feature Processing?	E8A	0/1	0: No 1: Yes
Efficiency	MLWF Effort: Data Preparation	Does it automate standardisation and normalisation?	E8B	0/1	0: No 1: Yes
Efficiency	MLWF Effort: Data Preparation	Does it automate bucketing and binning?	E8C	0/1	0: No 1: Yes
Efficiency	MLWF Effort: Data Preparation	Does it automate text preprocessing?	E8D	0/1	0: No 1: Yes

Table 4. Assessment Framework for Efficiency. Questions: E8A–E8F.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	MLWF Effort: Data Preparation	Does it automate time-period extraction?	E8E	0/1	0: No 1: Yes
Efficiency	MLWF Effort: Data Preparation	Does it assist with class imbalance via sampling techniques?	E8F	0/1	0: No 1: Yes

The sub-criteria in Table 4 can be mapped to the MLWF in Fig. 2 as follows: E8A–E8F to *Prepare* within *Data Engineering*.

Table 5. Assessment Framework for Efficiency. Questions: E14–E16.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Efficiency	Technical Efficiency	Does it undertake workload optimisation?	E14	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree
Efficiency	Technical Efficiency	Does it allow time limits for modelling?	E15	0/1	0: No 1: Yes
Efficiency	Technical Efficiency	Does it allow iteration/trial limits for modelling?	E16	0/1	0: No 1: Yes

The sub-criteria in Table 5 can be mapped to the MLWF in Fig. 2 as follows: E14 to *Provision Resources* within *Model Development* and E15/E16 to *CASH+* within *Model Development*.

- **Dirty Data (5 Questions).** This criterion specifically considers how robust an AutoML system is in the face of messy data, e.g. format issues, missing values, outliers, etc. It deserves its own category due to the considerable time and effort that tends to be invested into related tasks; see Section 4.1.1. We now explicitly list the questions on dirty data.

Table 6. Assessment Framework for Dirty Data. Questions: DD1–DD5.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Dirty Data	Dirty Data	Does it automatically clean dirty data?	DD1	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree
Dirty Data	Dirty Data	Does it automatically infer data types?	DD2	0/1	0: No 1: Yes
Dirty Data	Dirty Data	Does it automatically find and deal with missing values?	DD3	Scale 0:2	0: No 1: Partially, as it finds missing values 2: Yes
Dirty Data	Dirty Data	Does it automatically find and deal with outliers?	DD4	Scale 0:2	0: No 1: Partially, as it finds outliers 2: Yes
Dirty Data	Dirty Data	Does it undertake other domain-specific or advanced data cleaning operations?	DD5	0/1	0: No 1: Yes

The sub-criteria in Table 6 can be mapped to the MLWF in Fig. 2 as follows: DD1–DD5 to *Clean* within *Data Engineering*.

- **Completeness & Currency (13 Questions).** This criterion considers completeness through the lens of technical domain coverage, i.e. the types of ML problems that an AutoML package can handle. For instance, those constrained to binary classification tasks will only ever be able to assist organisations with a subset of business problems. Thus, the major subset of questions under this criterion determines the ML applications suitable for an AutoML system. Admittedly, an associated low score is not a problem for specialist tools, but it reflects poorly on any general applicability claims. The most ‘complete’ AutoML systems should additionally be configurable for arbitrary business domains, i.e. by enabling custom evaluation metrics for ML solutions. Beyond such a focus, there is an appraisal concerning the integration of HPO methods and libraries, the latter being chosen for evaluation over individual ML algorithms to ensure a degree of abstraction. After all, as mentioned earlier, an interface to scikit-learn 0.24.2 immediately provides access to 191 estimators. Finally, this criterion also assesses methodological currency, ensuring technical domain coverage uses up-to-date techniques and approaches. However, notably, this monograph surveys only open-source and commercial AutoML packages that are in popular use as of the early 2020s, meaning that

a lack of currency typically applies only to ‘faded’ tools; see Appendix A and Appendix B. We now explicitly list the questions on completeness & currency.

Table 7. Assessment Framework for Completeness & Currency. Questions: CC1–CC13.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Completeness & Currency	Technical Domain Coverage	How does it handle unsupervised learning?	CC1	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Technical Domain Coverage	How does it handle regression on tabular data?	CC2	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Technical Domain Coverage	How does it handle classification on tabular data?	CC3	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Technical Domain Coverage	How does it handle multi-class classification on tabular data?	CC4	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Technical Domain Coverage	How does it handle time series and forecasting?	CC5	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Technical Domain Coverage	How does it handle image-based problems?	CC6	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Technical Domain Coverage	How does it handle text-based problems?	CC7	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Technical Domain Coverage	Does it handle multi-modal tasks?	CC8	0/1	0: No 1: Yes

Table 7. Assessment Framework for Completeness & Currency. Questions: CC1–CC13.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Completeness & Currency	Technical Domain Coverage	How does it handle ensemble strategies?	CC9	Scale 0:2	0: Not at all 1: Via convenience features or platform extensions 2: Full AutoML
Completeness & Currency	Customisation	Does it allow custom evaluation metrics?	CC10	0/1	0: No 1: Yes
Completeness & Currency	HPO Coverage	Which HPO techniques does it offer?	CC11	Points 1:N	Grid Random Bayesian Multi-Armed Bandit Genetic Meta-learning
Completeness & Currency	Library Coverage	Which popular libraries does it interface with?	CC12	Points 1:N	Sklearn Keras TF XGBoost LightGBM Catboost Pytorch Ax R
Completeness & Currency	Currency	Is it actively maintained?	CC13	0/1	0: No 1: Yes

The sub-criteria in Table 7 can be mapped to the MLWF in Fig. 2 as follows: CC1–CC10 to *Data Engineering* and *Model Development* generally and CC11–CC13 to *CASH+* and *Requirements Review* within *Model Development*.

- **Explainability (7 Questions).** This criterion is core to any assessment of an AutoML system, even if the requirements manifest in different ways for different stakeholders [330]. Indeed, for practising data scientists, this primarily encompasses explaining how an ML solution arises, how it arrives at an output, and why its performance level is what it is. For other technical users, that insight into drivers for technical performance, e.g. feature importance, remains essential. As for corporate stakeholders, explainability must be present to ensure compliance with governance, regulatory and corporate social-responsibility requirements. Of course, beyond the standard questions, scenario-building capabilities are also essential to note, as they expand the value of an ML solution beyond predictive power to insight generation. Such a component would be especially desirable to semi-technical and business users that care more about understanding a problem context than any particular deployed model. Finally, this criterion includes an evaluation of whether an AutoML package considers bias and fairness. Specific tools are dedicated to this topic, so any appraisal must not only

consider the identification of associated flaws but also the capacity for their remediation; see Section 4.1.2. We now explicitly list the questions on explainability.

Table 8. Assessment Framework for Explainability. Questions: EX1–EX7.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Explainability	Data Lineage	Are data lineage & processing steps clear?	EX1	0/1	0: No 1: Yes
Explainability	Model Understanding	Is it clear what modelling steps were undertaken?	EX2	0/1	0: No 1: Yes
Explainability	Model Understanding	Does it automatically explain global model characteristics?	EX3	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree
Explainability	Model Understanding	Does it automatically explain local prediction-level artefacts?	EX4	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree
Explainability	Scenario Modelling	Does it support scenario exploration?	EX5	0/1	0: No 1: Yes
Bias & Fairness	Metrics	Does it automatically generate best-practice bias/fairness metrics for the model/data?	EX6	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree
Bias & Fairness	Metrics	Does it automatically mitigate and/or remediate bias/fairness flaws in the model/data?	EX7	Scale 0:2	0: No 1: No, but convenience features are available 2: Yes, to some degree

The sub-criteria in Table 8 can be mapped to the MLWF in Fig. 2 as follows: EX1–EX5 to *Visualise & Explain* and *Requirements Review* within *Model Development* and EX6–EX7 to *Explore & Assess Fairness* within *Data Engineering* and *Assess Fairness* within *Model Development*.

- **Ease of Use (5 Questions).** As with explainability, this criterion also manifests differently for different stakeholders, primarily due to the variability in technical skills and operational requirements. For instance, an AutoML tool only available via Python or R scripting immediately limits the userbase to technicians familiar with coding constructs, e.g. variables. Even amongst technicians, programming languages and data-science libraries that are less common will further restrict utility. Thus, one of the pertinent questions to ask is whether a CLI is available, given that it is somewhat more accessible to general users. Indeed, a CLI ideally requires nothing more than simple commands to be typed in and executed with a press of a return key. Of course, future work may further evaluate the usability of individual UIs, but, for this monograph, it is sufficiently informative to delineate between AutoML technologies that have an accessible UI and those that do not. As an aside, any particular quirks that assist with the translation of business problems to the system undertaking analytical work are also worth noting under this criterion, e.g. natural language processing (NLP) engines and other exotic forms of HCI. However, deeper discussions about HCI in AutoML are deferred to other reviews [313]. We now explicitly list the questions on ease of use.

Table 9. Assessment Framework for Ease of Use. Questions: EU1–EU5.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Ease of Use	Interface	Can it be interacted with via coding?	EU1	0/1	0: No 1: Yes
Ease of Use	Interface	Is there a CLI with simple commands?	EU2	0/1	0: No 1: Yes
Ease of Use	Interface	Is there a GUI?	EU3	0/1	0: No 1: Yes
Ease of Use	Interface	Is it desktop-based or browser-based?	EU4	Scale 0:2	0: Desktop only 1: Browser only 2: Both
Ease of Use	Learning	Is there clear and extensive documentation and guidance available?	EU5	Scale 0:2	0: No 1: Partially 2: Yes

The sub-criteria in Table 9 are not technically mappable to individual tasks/phases of the MLWF in Fig. 2. They refer to the ways in which one can interact with an AutoML system, as well as how convenient these forms of HCI are. Fundamentally, ease of use can impact every aspect of an MLWF, i.e. wherever a user must interact with an AutoML system to complete a task.

- **Deployment & Management Effort (11 Questions).** This criterion is perhaps the one that most distinguishes industrial concerns from academic considerations. Indeed, once experimentation results in an ML solution, it takes significant technical effort to embed the object within a business decision-making process [491]. Often, a production environment

must continue to feed an ML model with transformed data, potentially in real-time and/or streaming fashion, then transfer generated predictions/prescriptions to an end user or other downstream systems. Once deployed, an ML solution should also ideally be monitored for changes in defined metrics [129], e.g. those related to technical performance and fairness. Whether reactively, in response to monitored information, or proactively, optimal modelling may additionally necessitate reapplying earlier MLWF processes as part of maintenance, such as model retraining. Therefore, characteristics of an AutoML tool that assist or hinder this important criterion are crucial to appraise here. We now explicitly list the questions on deployment & management effort.

Table 10. Assessment Framework for Deployment & Management Effort. Questions: DM1–DM4.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Deployment & Management Effort	Deployment	Does it use model compression techniques?	DM1	0/1	0: No 1: Yes
Deployment & Management Effort	Deployment	Can it be deployed on-premise and/or in the cloud?	DM2	Scale 0:2	0: No 1: Yes, only in the cloud 2: Yes, both
Deployment & Management Effort	Deployment	Does it offer advanced deployment testing mechanisms, e.g. A/B or champion-challenger?	DM3	0/1	0: No 1: Yes
Deployment & Management Effort	Deployment	Does it offer advanced deployment update mechanisms, e.g. blue-green or canary?	DM4	0/1	0: No 1: Yes

The sub-criteria in Table 10 can be mapped to the MLWF in Fig. 2 as follows: DM1/DM2 to *Provision* within *Deployment*, DM3 to *Serving* within *Deployment* and *Proactive Training* and *Reactive Training* within *Monitoring and Maintenance*, and DM4 to *Serving* within *Deployment*.

Table 11. Assessment Framework for Deployment & Management Effort. Questions: DM5–DM11.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Deployment & Management Effort	Management	Does it automatically set up monitoring?	DM5	Scale 0:2	0: No, not present at all 1: No, manual setup and/or configuration is required 2: Yes
Deployment & Management Effort	Management	Does it monitor hardware usage and performance?	DM6	0/1	0: No 1: Yes
Deployment & Management Effort	Management	Does it monitor model performance metrics?	DM7	0/1	0: No 1: Yes
Deployment & Management Effort	Management	Does it monitor data/concept drift?	DM8	0/1	0: No 1: Yes
Deployment & Management Effort	Management	Does it monitor bias/fairness metrics?	DM9	0/1	0: No 1: Yes
Deployment & Management Effort	Management	Does it reactively retrain based on monitoring triggers?	DM10	Scale 0:3	0: No 1: No, but convenience features can assist manual retraining 2: Yes, with triggers defined by user 3: Yes, with triggers provided by developers
Deployment & Management Effort	Management	Does it proactively retrain?	DM11	0/1	0: No 1: Yes

The sub-criteria in Table 11 can be mapped to the MLWF in Fig. 2 as follows: DM5–DM9 to *Monitoring* within *Monitoring and Maintenance*, DM10 to *Reactive Training* within *Monitoring and Maintenance*, and DM11 to *Proactive Training* within *Monitoring and Maintenance*.

- **Governance (3 Questions).** In an organisational context, an ML tool must align to existing data governance and security considerations to ensure regulatory and internal-policy

compliance. Therefore, this brief but important category assesses whether an AutoML tool aligns with relevant practices. Questions include whether data access is appropriately managed [277], although, for ML, it is also essential to evaluate whether an organisation can control access to functionality. Such system features are particularly pertinent during model deployment, as this phase bears significant risks around exposing organisational artefacts to external parties and inadvertently embedding immature projects into core business processes/environments. Even with the best intentions to ensure security and controlled access, the ability to audit activity on internal systems is likewise desirable to confirm that appropriate activities are being undertaken by authorised entities [474]. We now explicitly list the questions on governance.

Table 12. Assessment Framework for Governance. Questions: G1–G3.

Criteria	Sub-Criteria	Question	QCode	Scoring	Rubric
Governance & Security	Governance	Does it offer auditing of activity?	G1	0/1	0: No 1: Yes
Governance & Security	Security	Does it offer artefact access controls for the model/data?	G2	0/1	0: No 1: Yes
Governance & Security	Security	Does it offer function access controls for training, deployment, etc.?	G3	0/1	0: No 1: Yes

The sub-criteria in Table 12 are not technically mappable to individual tasks/phases of the MLWF in Fig. 2. They refer to how an organisation integrates with and accesses an AutoML system, as well as how secure these forms of HCI are. Fundamentally, governance can impact every aspect of an MLWF, i.e. wherever a user can interact with an AutoML system to influence a task.

3.3 The Role of AutoML

The assessment framework proposed in this monograph reflects the desires of major stakeholder groups when engaging with ML in an industry setting. Although the constituent questions and their measurement rubrics are diverse, there is a loose general trend: the higher an AutoML package scores for the proposed criteria, the more it is seen to support performant ML. So, after such an amalgamation of varied requirements, it is worth asking whether the framework can be condensed into succinct insights about what drives AutoML uptake, now and in the future. Basically, what does industry see as the role of AutoML, both present and prospective?

There appear to be several answers, as follows:

- **Enhancing Data Science Practices.** The use of AutoML provides the potential to engage in ML processes that are, compared to manual efforts, more efficient, technically performant,

robust, and explainable. Implementations with ongoing developer support will likely adopt the best available approaches for relevant MLWF tasks and stay up-to-date with the latest technologies. Moreover, beyond simply creating a model object, the ongoing AutoML endeavour to be genuinely end-to-end promises scalable deployment capabilities and an easier way to monitor/maintain performance subject to real-world data dynamics. Progress in these directions represents advancement towards true AutonoML and next-generation technical abilities [308, 313]. At the very least, from a purely financial perspective, efficiently generating and conveniently deploying ML solutions that potentially perform better than manual selections will offer both cost savings and revenue maximisation.

- **Democratising Data Science Practices.** Beyond pushing the limits of capabilities that ML practitioners are familiar with, AutoML offers a gateway to ML techniques and approaches for those who are not trained technicians. While not without risks and other implications, this inclusive ‘democratisation’ stands to knock down existing skill barriers, with benefits flowing both ways. Specifically, organisations will likely leverage the power of data science with greater ease, while ML applications will profit from a more fluid influx of domain knowledge.
- **Standardising Data Science Practices.** Each AutoML system acts as a wrapper for a collection of methods that deal with targeted phases of an MLWF. There may be many services on offer as of the early 2020s, but there are still far fewer packages than individual techniques. Thus, centralising work efforts into operations framed by a common system provides many potential benefits, reproducibility among them. Standardisation of such practices also supports stronger security mechanisms and access controls alongside enhanced auditability and thus governance. Indeed, as society continues to expect increasingly more from its AI engagement, particularly on the ethical front, the existence of AutoML may make it easier to certify compliance. At the very least, the technology cannot exist at odds with data governance practices lest corporate decision makers lean towards caution, hindering industrial uptake and successful business integration.

Table 13. Key criteria that an ML application should satisfy according to stakeholders. Considerations are marked F for fundamental and C for contextual.

Criteria	Data Scientist	Analyst	Deployment Technicians	Corporate	End Users
Technical Performance	F		F	C	F
Efficiency	F		F	C	
Dirty Data	F			C	
Completeness & Currency	F				
Explainability	C	C		C	F
Bias / Fairness	C	C		C	F
Ease of Use	F	F			
Deployment & Management Effort	F		F	C	

In light of these overarching industrial expectations of AutoML, it is worth presenting one last condensed overview, specifically around which stakeholders care most about particular elements of the aggregated framework presented in Section 3.2. Table 13 does so, marking criteria by F if they are fundamental considerations to a stakeholder during engagement with an ML application. Additionally, C denotes a contextual criterion, i.e. one that is conditionally important to a stakeholder depending on organisational context. Of course, as discussed earlier, different stakeholders have varying requirements and degrees thereof concerning each criterion; some will find certain concepts virtually irrelevant to their role.

Unsurprisingly, data scientists care about the greatest number of listed criteria, given the ongoing centrality of a technical role in an ML application. After all, a core purpose of AutoML is to assist

such practitioners closely with their technical tasks. However, management is also broadly invested in AutoML technology supporting performant ML. Beyond corporate stakeholders typically needing to sign off on software purchases, they also act as the bridge between the fundamental work of a data scientist or analyst and the business context in which an ML solution is to be deployed. Thus, while managers usually do not operate AutoML technology, they still have wide-ranging conditional requirements that must be satisfied.

As for deployment technicians, these stakeholders become interested in AutoML once associated systems extend beyond the traditional core of ML model selection. Specifically, such technicians will likely be keen to employ AutoML systems that incorporate and simplify MLOps processes. Accordingly, their requirements are primarily infrastructural. In a large enough analytics team, data scientists and deployment technicians will be distinct, so, while both care about technical performance and efficiency, the foci of both groups are different. The former seeks to efficiently train up an ML pipeline with a high degree of validity, while the latter seeks to ensure a deployed solution services queries efficiently and maintains a baseline level of ‘correctness’.

Finally, end users may not interact directly with the development of an ML solution, but they remain affected by the entire process. This statement holds whether the user is a target of low-stakes product recommendation or a more impactful application, e.g. an AI-based hiring decision or loan determination. Accordingly, trust – this requires explainability – is crucial for their engagement with an organisation that delivers the product. If the end user is not satisfied that an ML solution is technically performant, unbiased and fair, they remove their financial support or, depending on the application, pursue recompense. In effect, a happy client/customer indirectly validates using an AutoML service.

Notably, in the summation within Table 13, there is no explicit mention of the business user detailed in Section 3.1.1. Such a stakeholder without technical skill can be considered to sit within the corporate section, supplying domain knowledge and engaging with core technicians via an analyst as required. In some framings of this structure, the role is considered a ‘domain expert’, while others may refer to a ‘project sponsor’. Regardless of terminology, this stakeholder does not traditionally have any direct involvement with an MLWF beyond perhaps the initial phase of problem formulation & context understanding. However, given the potential of AutoML to drive organisational value through democratisation, business users remain important; any AutoML system that facilitates HCI for these technical non-experts accelerates technological adoption.

The question then is as follows: if industrial uptake of AutoML is contingent on how well it enhances, democratises and standardises data science practices, how far along is the technology when meeting expectations? In seeking to answer this question, incoming sections of the review examine how well modern offerings support performant ML, additionally attempting to glean whether AutoML has begun making a societal impact.

4 AUTOML SOFTWARE

The cataloguing of open-source and commercial AutoML services is splintered across various blogs, tutorials, social media posts, and GitHub-based collections. Currently, there is no centralised index for this technology that is both reliable and up-to-date. Thus, this review sifts through a compilation of sources to present one such encompassing snapshot of AutoML products as of the early 2020s. Investigations supporting this survey, both broad and deep, have involved, for instance, popular search engines, blogs on data science and analytics [26], and curated repositories [218, 554].

First, however, we detail the constraints around the scope of the review, i.e. what software packages can be considered under the banner of AutoML. Here, we define an AutoML ‘tool’ as a computer program that, through some form of HCI, automates some element of the MLWF in Fig. 2. Crucially, the initial high-level process to receive concerted focus under the banner of AutoML was

model selection, specifically HPO. The field has gradually expanded its attention in the years since, even though the model-development phase remains a strong priority. Given the variety of software packages currently accessible to the public, a further categorisation is thus possible.

If an AutoML tool allows a user to train an ML model on a given dataset as one of its internal functions, which is the core process of ML, we call it ‘comprehensive’. Such a tool will also often be referred to here as an AutoML system, acknowledging the integrated nature of multiple mechanisms [308]. This definition does not mean that all comprehensive tools are identical in scope. Some may or may not include ensemble mechanisms, data cleaning, etc. However, they all act as wrappers around an ML solution object and, in some way, manage it.

In contrast, if an AutoML tool automates an aspect of an MLWF without directly training an ML model, we call it ‘dedicated’. The term is used because this survey finds these ancillary tools predominantly focus on but a single element of the MLWF. The reasons for this specialisation are numerous. For instance, many HPO tools are actually built agnostically, such that they can be applied to arbitrary optimisation problems as formulaically as to selecting hyperparameter configurations for ML pipelines. Hence, the dedication to the HPO task is, in some sense, a fruitful application of a more general theory. Other times, a dedicated tool is simply designed to prioritise challenges beyond model development without becoming bogged down in unnecessarily comprehensive implementation. Regardless, despite the diversity of dedicated tools in existence, this review finds that they are most concentrated in three areas: (1) data and feature engineering, (2) bias, fairness and explainability, and (3) HPO.

Naturally, whether dedicated or comprehensive, all AutoML tools were considered under exclusion criteria to ensure a meaningful analysis of industrial uptake. The mere presence of a GitHub repository did not automatically qualify a tool for this review. Exclusion criteria include the following:

- A lack of open-source or commercial implementation ready for organisational use. This criterion immediately cuts out numerous academic proposals and experiments. Of course, this does not devalue the novel/interesting AutoML science that these projects engage in, e.g. the reductivist evolutionary principles of AutoML-Zero [206, 459]. However, whether such theoretical concepts gain greater traction in mainstream usage is a matter for future assessment; it remains out of scope for this monograph.
- Association with a repository that possesses minimal commits and stars, often appearing as a quick burst within a short window of time and paired with the publication of an academic paper. Such a scenario suggests the repository serves more as an informational webpage or code storage facility than a means of hosting genuinely usable software for the global data-analytics community.
- Indication that a project is overly casual, i.e. done for personal interest or publicity by an individual or small group.
- A lack of updates or maintenance activity for over 18 months, occasionally paired with an announcement of the tool being deprecated. However, some of these ‘faded’ packages are still useful to mention as part of historical commentary; see Appendix A and Appendix B.

In essence, these exclusion criteria all aim to reflect the subset of AutoML technology with which industry is seriously engaging at present.

Finally, we make one more relevant note; this review treats AutoDL as largely out of scope. Here, we define an AutoDL tool as being both (1) focussed on NAS and (2) not primarily designed for tabular data. Fortunately, this exclusion has not been an issue for the most part. While AutoDL receives substantial interest as a subtopic of DL, the extreme computational resources required have stymied any significant dissemination of the technology throughout mainstream industry. Granted,

this state of affairs may change quickly at any time, and some blur in the discussion within this monograph is inescapable. However, at this point in the early 2020s, AutoDL remains extremely focussed on the model-development phase of an MLWF and, overall, has not translated far beyond the academic proof-of-principle stage. Deeper discussions are available elsewhere [195].

We now proceed to discuss the surveyed AutoML tools from the perspective of performant ML, for which an assessment framework was introduced in Section 3.2. Open-source software is explored first, with Section 4.1 for dedicated tools and Section 4.2 for comprehensive systems. Commercial AutoML products, which tend to be relatively opaque, are then analysed within Section 4.3.

4.1 Dedicated AutoML Tools

This section surveys and discusses open-source tools that, to some degree, automate a specific element of an MLWF. Several of these programs are occasionally incorporated as mechanisms within other holistic software packages. However, in complete isolation, they do not enable an ML application to be run; the core ML processes of model generation and management are typically absent.

Broad exclusion criteria have already been listed, but it is still worth providing examples of projects beyond the scope of this review. Doing so raises awareness of how much research and development is ongoing within the field of AutoML. For instance, conceptual tools lacking a substantive open-source implementation at the time of review include One Button Machine (OBM) [331], Persistant [120], and BigDansing, a system for big data cleansing [311]. Then there are codebases that, while existing, have been given limited developmental/maintenance attention, often serving to supplement the publication of an academic paper. Examples include ExploreKit [305, 306], the Data Civilizer System [193, 452], and NADEEF [176, 177]. In fairness, NADEEF exemplifies software that appears to have been actively maintained at one point, but the lack of update in over five years implies it has since ‘faded’. Finally, Cognito [158] is an example of exclusion based on appearing to be a small-scale personal/group project. Such codebases tend to have lower levels of popularity, quality or maintenance compared to software entrenched within industry, although there is no such insinuation regarding Cognito specifically; some of these projects may eventually attain sufficient recognition to be more broadly adopted by organisations and their stakeholders.

Now, the incoming discussion around dedicated AutoML tools has been organised into three categories that encompass the majority of surveyed tools. Section 4.1.1 covers data and feature engineering, Section 4.1.2 relates to bias, fairness and explainability, and Section 4.1.3 addresses HPO. Notably, data visualisation, arguably essential throughout an MLWF, is conspicuous in its absence. This omission is not because there is a dearth of tools dedicated to such a form of HCI. In contrast, there are simply too many packages in this space to consider for this survey. Such consideration is also inappropriate for scope, as designing visual representations for data can be done entirely without ML in mind. There are exceptions, however, e.g. if a visualisation package is customised to optimally communicate specific ML metrics. In practice, such products are still commonly dedicated to specific tasks, such as evaluating bias and fairness, thus slotting organically into one of the categories mentioned above.

4.1.1 Data and Feature Engineering. The first category of dedicated AutoML tools in this review encompasses those that focus solely on data preparation tasks. The surveyed open-source packages classified as such are listed in Table 14.

Arguably, the most well-known within this set of codebases is Featuretools [70], which can generate numerous features for subsequent ML model training. As is typical for automated feature-engineering efforts, Featuretools provides a pool of ‘primitives’, i.e. basic feature transformations,

Table 14. Dedicated AutoML Tools: Data and Feature Engineering.

Name	Categorisation	Github	Ref.
Compose	Data Processing	https://github.com/FeatureLabs/compose	[68, 303]
Featuretools	Feature Engineering	https://github.com/FeatureLabs/featuretools	[70, 304]
Feature-engine	Feature Engineering	https://github.com/solegalli/feature_engine	[227]
Boruta	Feature Engineering	https://github.com/scikit-learn-contrib/boruta_py	[68, 327]
tsfresh	Feature Engineering	https://github.com/blue-yonder/tsfresh	[153, 239]

such as ‘mean’ and ‘sum’. These primitives can then be stacked into even more complex transformations of data. In contrast, Feature-engine [227] avoids this form of deep feature synthesis, but it does provide an expansive variety of operations for data cleaning and preparation alongside feature generation. The package, however, does not automatically apply every operation, so some manual involvement is still required in making appropriate choices.

In both cases, an ML model never has to enter the picture. Mechanisms that prioritise feature generation are primarily sold on the basis of conveniently, and hopefully intelligently, providing many new perspectives on a dataset. On the other hand, Boruta [265] is solely dedicated to automated feature selection, which is often a more difficult task. The challenge is deciding which features are actually helpful for generating an accurate ML solution. In the case of Boruta, the package appears to act in wrapper style, as defined in Section 2. Specifically, it relies on the technical performance of a model to guide the search for good features, although this dependence on an ML model object does not mean the mechanism is considered a comprehensive AutoML system under the definitions in this review.

The remaining two packages indicate the diversity of available dedicated tools operating at this phase of an MLWF. The Compose program [68] deserves its own data-processing subcategory in Table 14, operating upstream relative to the other tools. It assists a stakeholder in constructing a training set from data, assuming one has not been previously prepared, and thus provides valuable labelling and extraction operations. As for ‘tsfresh’ [239], it is unique for being dedicated not only to a specific task, i.e. feature engineering, but also to a specific technical domain. The tool essentially generates features for problems and data that involve time series. Such customisation to distinguish an AutoML service from the rest of the pack is a recurrent theme and is examined further in Section 5.2. However, despite the variety of dedicated tools on offer, it is still notable that only a handful can arguably be considered adequate for organisational use. Certainly, there is a lot of research activity in automating data preparation, but, as an endeavour trailing HPO and automated model selection, perhaps it should not be surprising that technological translation here is nascent.

4.1.2 Bias, Fairness and Explainability. The second category of dedicated AutoML tools in this review encompasses those focussing solely on bias and fairness metrics, effectively grappling with explainability. The surveyed open-source packages classified as such are listed in Table 15.

Given the recent surge in importance that the topic of ML trustworthiness has attained, this section warrants a more extensive discussion than the other two categories of dedicated AutoML tools. First, it is essential to note that the bias of societal concern does not refer to the typical concept that technicians are familiar with, i.e. the bias-variance trade-off that explains the high-bias underfitting and high-variance overfitting of ML models [232]. From an ethical standpoint, bias is instead defined as an inclination/prejudice for or against a person or group, and several academic works have attempted to further systematise this notion, e.g. categorising six sources of bias [388, 522]. Unsurprisingly, bias is intrinsically linked to fairness, defined elsewhere as “the

Table 15. Dedicated AutoML Tools: Bias and Fairness.

Name	Categorisation	Github	Ref.
AI Fairness 360	Bias-Fairness	https://github.com/Trusted-AI/AIF360	[350]
FairLearn	Bias-Fairness	https://github.com/fairlearn/fairlearn	[122, 210]
Audit AI	Bias-Fairness	https://github.com/pymetrics/audit-ai	[450]
aequitas	Bias-Fairness	https://github.com/dssg/aequitas	[179, 478]
LIME	Viz (Bias-Fairness)	https://github.com/marcotcr/lime	[464, 465]
SHAP	Viz (Bias-Fairness)	https://github.com/slundberg/shap	[363, 364]
AIExplainability360	Viz (Bias-Fairness)	https://github.com/Trusted-AI/AIX360	[83, 349]
What If Tool	Viz (Bias-Fairness)	https://github.com/PAIR-code/what-if-tool	[436]

absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics” [151]. In turn, a lack of fairness in ML can have adverse impacts, including ‘allocation harms’, where resources and opportunities are withheld from a particular group, and ‘quality-of-service harms’, where predictive/prescriptive outcomes may be relatively poor for a particular group [210]. Of course, the topic of bias and fairness is highly complex, and a more in-depth treatment of such issues in AutoML is available elsewhere [313]. It is sufficient to note that bias can creep into an ML solution at many stages of an MLWF, e.g. within data, ML algorithms, human decisions, and various processes. Accordingly, the tools in this category support, to varying degrees, automatically diagnosing and mitigating bias-and-fairness issues.

The most popular of the dedicated tools in Table 15 is arguably AI Fairness 360 (AIF360) [350], created by IBM. At the time of survey, AIF360 was able to calculate three ‘individual fairness’ metrics [221, 514, 574] and apply 13 bias mitigation algorithms based on external research [54, 55, 133, 141, 214, 250, 297–299, 307, 441, 566, 568]. The package also supports additional metrics evaluating ‘group fairness’ and sample distortion. However, while these complex calculations and procedures are automated, their selection is not. Sample tutorial notebooks are provided by AIF360, but, within them, it is experts that decide on a metric and remedy. Likewise, befitting an AutoML tool that is not considered comprehensive, integration into an ML application also requires manual involvement. Additionally, this package has specific developer-environment requirements, which can be a barrier to entry for stakeholders with alternative operational practices. However, this particular comment is more of a general reminder than a specific critique, given that few software applications approach genuine cross-platform universality.

Now, the way that HCI works for such a tool sparks a broader discussion: for an MLWF task dedicated to the careful and deliberate review/amendment of bias/fairness flaws, how much of the process should be automated? One might argue that, at the very least, a bias-and-fairness tool should automatically calculate and visualise all relevant metrics, in much the same way as automated feature engineers in Section 4.1.1 systematically generate multiple perspectives of a dataset. The counterargument is that an embarrassment of riches may still not make it any easier for a data scientist, let alone a non-technical user, to make the best operational decisions. Metric and remedy selection is as important here as feature selection is to data engineering. Thus, if full automation is deemed unwise in bias-and-fairness space, then some form of recommendation prompts would still likely provide added value to a user. For now, the automation capability of most dedicated tools in Table 15 does not extend far in this direction. FairLearn [210], developed and released by Microsoft, is akin to AIF360 in the level of development, documentation, and maintenance. The package contains some overlap in terms of mitigation algorithms [54, 55, 250], but it also provides a dashboard with a more convenient overview of metrics and comparative

analyses for both models and potentially discriminated groups. This UI, along with very detailed documentation covering software usage and topical elaboration, probably marks the current limit of mechanisation for this aspect of the MLWF, at least among non-comprehensive tools.

Comparatively, Audit AI [450] is a less mature library focussed solely on bias-and-fairness diagnosis. It supports the calculation of several metrics under techniques named ‘4/5th’, ‘fisher’, ‘z-test’, ‘bayes factor’, and ‘chi squared’. Additionally, it provides ‘sim beta ratio’ and ‘classifier posterior probabilities’ tests for classification tasks, as well as analysis of variance (ANOVA) and a test of “group proportions at different thresholds” for regression tasks. Overall, the software is light on documentation, which essentially comes as a readme file on GitHub. However, it is notable for discussing regulatory needs in detail, e.g. referring to a set of guidelines for employee selection procedures [419] while motivating its 4/5ths rule. Likewise, albeit without citation/justification, the package mentions implementing the Cochran-Mantel-Hanzel statistical test [159, 378], which is very common in regulatory practices. In effect, Audit AI highlights the surging importance of governance to real-world ML, as stressed in Section 3.1.2. Finally, ‘aequitas’ [179] also makes the cut for this survey, despite a relatively diminished amount of activity. It enables calculating metrics derived from a confusion matrix, although applied to subgroups within data. Moreover, it can be used as a code library or interacted with via a CLI, generating reports with simple graphs. Thus, despite being relatively basic, it rounds out the tools dedicated directly to bias and fairness in this section.

Of course, there is an inherent dependence of ML trustworthiness on explainability; one cannot confidently assess bias and unfairness without understanding how an ML solution works. This model-comprehension process is often best assisted by transforming technically arcane code into a more interpretable format. Thus, although we keep the survey constrained, the handful of visualisation (viz) tools in Table 15 is worthy of acknowledgement. For instance, LIME [464, 465] is a package that pursues model-agnostic visualisations to assist in explaining black-box ML classifiers. In turn, it is referenced by SHAP [363, 364], which allows users to visualise and better understand numerous ML solutions, e.g. whether structured as a linear model, tree, or a DL network. Neither of these two tools explicitly provide instructions for bias-and-fairness application; it is up to a user to integrate these procedures into auditing operations. Nonetheless, the packages are well maintained and documented, and the calculations/processes they automate simplify such tasks for ML stakeholders. Indeed, AIExplainability360 [349] is an example of a tool leveraging both LIME and SHAP alongside other software to promote explainability, despite not being updated recently and facing the common issue of users needing to gauge the utility/appropriateness of different functions themselves. Finally, the What If Tool [436] by Google does return focus to fairness, allowing users to easily select and view five fairness metrics calculated for their ML model: ‘group unaware’, demographic parity, equal opportunity, equal accuracy, and group thresholds. There is an obvious crossover with AIF360 and FairLearn, but the library does retain a performance mindset alongside bias-and-fairness awareness, e.g. when enabling users to analyse comparative statics for ML models.

One overall conclusion is that there is a surprising level of activity under this category of dedicated AutoML tools. The survey already narrowly excludes several packages, such as Parity Fairness [448], which appears to be an incremental wrapper of AIF360 and FairLearn, FairML [51], which seems to be a personal project for feature-importance graphs that faded in 2017, and Scikit-Fairness [170], which has been merged into FairLearn. However, only AIF360 and FairLearn, both corporately sponsored, can be considered outliers in terms of maturity. There is also much overlap amongst the packages. For instance, AIF360 lists both FairLearn and LIME as dependencies. Granted, the motivations that drive development in this space are likely to ensure continued progress; both

AIF360 and FairLearn appear to be driven by socioethical considerations, while Audit AI presages the rise of regulatory obligations in industrial ML.

Nevertheless, all of the tools are presently constrained in terms of automation, requiring both upskilling and careful manual interaction for full engagement with an ML auditing task. A thorough review of some of these tools reaches similar conclusions [342], noting that, despite stakeholder interest, it remains difficult to acquire the necessary skills/knowledge to make informed choices. In essence, this topic is faced with the challenge of balancing accessibility with deep technicality. The reviewing paper also found commonly held concerns around how to integrate these tools with existing MLWFs and grant them appropriate coverage. Indeed, the challenge of integrating any dedicated mechanism into a larger automated system is nontrivial [308]. Regardless, given the increasing public awareness and scrutiny around ML trustworthiness, time will tell how this technological subspace evolves.

Table 16. Dedicated AutoML Tools: HPO.

Name	Requirements	GUI	CLI	HPO Mechanisms	Ref.
Auptimizer	Sklearn	Y	Y	Random, Grid, Multi-armed Bandit (MAB), Bayesian, Evolutionary	[282, 356]
Bayesian Optimisation	Sklearn	N	N	Bayesian	[416]
BayesOpt	NONE	N	N	Bayesian	[379, 380]
BoTorch	Torch	N	N	Bayesian	[106, 451]
BTB	Sklearn	N	N	Bayesian	[84, 509]
DEAP	NONE	N	N	Evolutionary, Particle Swarm Optimisation (PSO)	[219, 220]
Dragonfly	NONE	N	Y	Bayesian	[301, 302]
GPflowOpt	TF	N	N	Bayesian	[316, 317]
Hyperopt	Sklearn, Extras (LightGBM)	N	N	Bayesian	[117, 118]
mlrMBO	NONE	N	N	Bayesian	[123, 529]
Nevergrad	Bayesian Optimisation, Torch, TF, Keras	N	Y	Evolutionary, Random, PSO, Bayesian, Genetic, Mathematical	[208]

Table 16. Dedicated AutoML Tools: HPO.

Name	Requirements	GUI	CLI	HPO Mechanisms	Ref.
Optuna	Sklearn, Scikit-Optimize, MLflow, Extras (LightGBM, Torch, CatBoost, MXNet, XGBoost, Keras, TF, Dask-ML, BoTorch, fastai)	N	Y	Bayesian, Evolutionary, Grid, Random, MAB	[60, 61]
RBFOpt	NONE	N	Y	Mathematical	[160, 165]
Scikit-Optimize	Sklearn	N	N	Bayesian	[489]
SMAC3	Sklearn	N	Y	Bayesian	[90, 272]
Tune-sklearn	Sklearn, Ray[Tune]	N	N	Random, Grid, Bayesian	[458]

4.1.3 Hyperparameter Optimisation. The third category of dedicated AutoML tools in this review encompasses those that focus solely on HPO. The surveyed open-source packages classified as such are listed in Table 16.

Given that the birthplace of modern AutoML is set in the topic of HPO [308], with plenty of prior discussion about optimisation methods and their implementations in the literature, an extensive commentary is not needed here. Every tool in the table is considered dedicated, so they do not house an ML solution during its life cycle in an end-to-end MLWF. Instead, the typical operation is for a user to provide a searchable hyperparameter space to the HPO package, then iteratively try out recommended configurations and update the package on how well those choices technically perform. Even speed-up mechanisms such as successive halving are often presented to technicians as cues for when they need to train their ML models on larger subsets of data. It is noted here that whilst Early Stopping can be applied to a variety of HPO mechanisms, it is also included given it is offered as a method by which an end user aims to arrive at an optimal set of hyperparameters. As a definitional note, those which fall under numerical optimisation are grouped as 'Mathematical'. This includes, for example, Covariance matrix adaptation evolution strategy (CMA-ES). Essentially, integrating these detached mechanisms into an automated model-development process requires some manual labour from stakeholders.

Tools that internalise ML models/algorithms, e.g. ones wrapping scikit-learn or other packages, are naturally excluded from this section on account of leaning towards being comprehensive systems, i.e. allowing a user to convert raw data into a trained ML solution. Other exclusions are due to a lack of currency, often because meaningful activity has not been present in the repository within the last 18 months. Sometimes, as with GPyOpt [528], the software has been explicitly announced as deprecated. Consequently, given that the first popularised systems of the modern AutoML era were developed around ten years ago, Appendix A notes a relatively long history of faded HPO tools compared to other dedicated categories.

Several insights can be gleaned from Table 16, which notes package names and associated characteristics. For instance, the table lists package dependencies, often pulled from a requirements.txt or

setup.py file in a GitHub repository. Notably, despite the efforts of this review to avoid delving into AutoDL, DL frameworks such as TensorFlow (TF), Keras and PyTorch (Torch) occasionally appear as such dependencies. However, half of the packages operate around scikit-learn, even though a user must still manually code the training of an ML model. A couple of subsequent columns in the table then assess how stakeholders interact with these packages. In all cases beyond Auptimizer, the technical threshold for users is high; dedicated HPO tools are overwhelmingly designed to be paired with coding scripts. That said, several packages do provide CLIs, which eases accessibility slightly. Stakeholders avoid serious programming in such cases, simply editing a configuration file instead.

Finally, we note that the dedicated tools cover a range of HPO procedures in Table 16. However, it is clear that Bayesian optimisation techniques have embedded themselves in widespread usage, even if this is partially a consequence of promotion by the research groups that initially researched this space, i.e. a first-mover advantage. For the same reason, bandit-based methods also have a strong representation, e.g. via an associated Hyperband implementation offered by Auptimizer and Optuna. In fact, recent years have seen research groups fuse the benefits of both strategies, such as with the Bayesian optimisation and Hyperband (BOHB) algorithm that Auptimizer and Optuna both provide. Naturally, different implementations do attempt to improve upon existing methods, so there are often a variety of efficiency-based tweaks augmenting the fundamental techniques. For instance, Tune-sklearn employs early stopping, while Auptimizer leverages successive halving, iteratively testing smaller sets of hyperparameter configurations on larger training-data samples. Optuna does both. Then, moving beyond these most popular mechanisms, evolutionary algorithms and particle swarm optimisation are also available, as well as other techniques termed ‘mathematical’. These remnants include the use of radial basis functions in RBFOpt and sequential quadratic programming (SQP) in Nevergrad. Ultimately, it is evident that the HPO category of dedicated AutoML tools is substantially mature. Although the average non-technical stakeholder is unlikely to engage with such software, the technical user will probably appreciate the flexibility associated tools provide in their ‘detachable’ form.

4.2 Comprehensive AutoML Systems: Open-Source

While dedicated tools certainly have their place in industrial use, the holy grail of AutoML is arguably to develop a framework that comprehensively automates *every* relevant task from one end of an MLWF to the other [308]. The field is nowhere near that goal yet. Nonetheless, within the last decade, many implementations have arisen that can automate the core facet of automated model development, i.e. CASH, while managing the fundamental ML process of model training. It is this internalised operation of converting data into an ML model that differentiates such packages from the dedicated HPO tools in Section 4.1.3. Some of these architectures have even extended their scope further out to neighbouring tasks and phases within an MLWF.

Table 17. The list of surveyed comprehensive AutoML systems that are open-source and active.

Name	Ref.
Auto_ViML	[495]
AutoGluon	[72]
AutoKeras	[310]
AutoML Alex	[343]
Auto-PyTorch	[92]
auto-sklearn	[93]

Table 17. The list of surveyed comprehensive AutoML systems that are open-source and active.

Name	Ref.
carefree-learn	[258]
FLAML	[395]
GAMA	[238]
HyperGBM	[181]
Hyperopt-sklearn	[321]
Igel	[98]
Lightwood	[399]
Ludwig	[353]
Mljar	[402]
mlr3automl	[121]
OBOE	[562]
PyCaret	[449]
TPOT	[527]

In this section, we analyse 19 open-source tools, listed in Table 17, that the survey found to be comprehensive AutoML systems. Each tool is assessed according to the criteria for performant ML introduced in Section 3.2, with information sourced from both available documentation and associated codebases. Importantly, not all parts of the evaluation framework apply to every tool, e.g. if the system is provided in the form of Python libraries, and these cases are highlighted by commentary when relevant. Furthermore, this review finds a broad spectrum of capability, ranging from holistic implementations that assist stakeholders for several MLWF phases to those that barely meet the definition of comprehensive. Additionally, we re-emphasise that the listed tools are considered current; historically notable faded projects are included in Appendix B for completeness.

With that context provided, the following findings are typically presented in tabular format. They may note:

- How each tool scores across a grouped set of criteria, thus providing a comparison between tools. Scores are also summed across all tools, thus assessing the general maturity of the AutoML ‘market’ per sub-criterion.
- The number of tools that score at different levels for a particular sub-criterion. This aggregation is another insight into ‘market’ maturity.

As Section 3.2 details, all scores are usually binary, i.e. 1/0 for yes/no, or on an integer scale. Occasionally, a sub-criterion involves listing tool features instead, and these are expanded into their own tables.

We now proceed to discuss the criteria, organised as follows: efficiency in Section 4.2.1, dirty data in Section 4.2.2, completeness & currency in Section 4.2.3, explainability in Section 4.2.4, ease of use in Section 4.2.5, and the remaining elements of the assessment framework in Section 4.2.6.

4.2.1 Efficiency. Most open-source AutoML tools arise from academic research. So, concerning the assessment framework for performant ML, it is almost immediately evident that operational efficiency is not the highest of priorities. Table 18 shows that none of the surveyed systems provides a model repository (E1) or model VCS (E2), which would require a persistent data storage of some sort. As for experiment tracking (E3), Table 19 affirms that only two tools have considered this notion, namely AutoML Alex and HyperGBM. Both are automated to the point of providing

Table 18. Scores for open-source comprehensive AutoML systems (E1–E3). Evaluates the existence of a model repository (E1), a model VCS (E2), and experiment tracking (E3). See Table 1 for rubric.

Name	E1	E2	E3	Sum (out of 4)
AutoML Alex	0	0	2	2
HyperGBM	0	0	2	2
Auto_ViML	0	0	0	0
AutoGluon	0	0	0	0
AutoKeras	0	0	0	0
Auto-PyTorch	0	0	0	0
auto-sklearn	0	0	0	0
carefree-learn	0	0	0	0
FLAML	0	0	0	0
GAMA	0	0	0	0
Hyperopt-sklearn	0	0	0	0
Igel	0	0	0	0
Lightwood	0	0	0	0
Ludwig	0	0	0	0
Mljar	0	0	0	0
mlr3automl	0	0	0	0
OBOE	0	0	0	0
PyCaret	0	0	0	0
TPOT	0	0	0	0
Sum	0/19	0/19	4/38	

Table 19. Distribution of scores for open-source comprehensive AutoML systems (E3). Evaluates the existence of experiment tracking. Scores: 0 for none, 1 for storage/access with limited automation/visuals, 2 for storage/access with automatic log visualisation, and U for unclear.

E3	
Score	# Tools
0	17
1	0
2	2
U	0

convenient visuals of experimental logs. However, AutoML Alex benefits here by plugging into the Optuna dashboard, a dedicated tool covered in Section 4.1.3.

Now, because all surveyed open-source systems lack a centralised form of persistent data storage, they cannot provide several other features for operational efficiency. For instance, we identified no repository of templates/code generated either by users or developers that could automatically kickstart an ML application (E4). That is not to say that tutorials are unavailable; all packages were found to provide example code that a stakeholder can manually leverage. They vary in substance and quality from minimal working examples to extensive end-to-end notebooks. Next, the suggestion of prior work (E5), which essentially requires a recommendation mechanism on top of a template/code repository, is also absent. Admittedly, a few surveyed systems claim to offer meta-learning capabilities, such as warm-starting, but such procedures do not typically kickstart the

operational side of an ML application, nor do they have flexible VCS in mind. Thus, meta-learning commentary is deferred until criterion CC11. Finally, if there is no centralised framework for storing project artefacts, it is understandably hard to automate shared access and collaboration (E6).

Table 20. Scores for open-source comprehensive AutoML systems (E7/E8). Evaluates the extent of automation for assisted data exploration (E7) and data preparation (E8). See Table 3 for rubric.

Name	E7	E8	Sum (out of 5)
Auto_ViML	0	2	2
AutoGluon	0	2	2
AutoKeras	0	2	2
AutoML Alex	0	2	2
Auto-PyTorch	0	2	2
auto-sklearn	0	2	2
carefree-learn	0	2	2
FLAML	0	2	2
GAMA	0	2	2
HyperGBM	0	2	2
Hyperopt-sklearn	0	2	2
Igel	0	2	2
Lightwood	0	2	2
Ludwig	0	2	2
Mljar	0	2	2
mlr3automl	0	2	2
OBOE	0	2	2
PyCaret	0	2	2
TPOT	0	2	2
Sum	0/57	38/38	

Moving on to discussions of effort minimisation during phases of an MLWF, this survey could not identify any significant assistance with data exploration (E7) among open-source packages. Such a result is likely due to their primary focus on model development, with general assumptions that input datasets are essentially predetermined. Accordingly, in the open-source domain, the onus for exploring/understanding data remains on the ML stakeholder. Supplementary tools explicitly dedicated to EDA must be sourced externally. In contrast, Table 20 indicates that all surveyed comprehensive AutoML systems do automate some form of data preprocessing (E8).

Of course, data preparation can involve many processes, so it is insufficiently informative to merely state that a comprehensive AutoML system assists or automates this MLWF phase. Therefore, we also assess the coverage of each tool across several common forms of data preprocessing:

- A. Categorical processing, e.g. dummy and one-hot encoding.
- B. Standardisation/normalisation, i.e. feature rescaling.
- C. Bucketing/binning, i.e. for continuous variables.
- D. Text preparation, e.g. tokenising or embedding.
- E. Time-period extraction, e.g. determining day of week from dates.
- F. Class-imbalance management, i.e. sampling techniques.

As Table 21 reveals, there appears to be plenty of variability in the scope of automated data preparation. The only consistency is that all comprehensive AutoML systems manipulate categorical

Table 21. Coverage of data-preparation processes for open-source comprehensive AutoML systems (E8A–E8F). Evaluates categorical processing (E8A), standardisation/normalisation (E8B), bucketing/binning (E8C), text processing (E8D), time-period extraction (E8E), and management of class imbalance (E8F). Scores: 0 for absent and 1 for present.

Name	E8-A	E8-B	E8-C	E8-D	E8-E	E8-F	Sum (out of 6)
Auto_ViML	1	1	1	1	1	1	6
Ludwig	1	1	0	1	1	0	4
Mljar	1	1	0	1	1	0	4
PyCaret	1	1	1	0	0	1	4
AutoGluon	1	0	0	1	1	0	3
auto-sklearn	1	1	0	0	0	1	3
FLAML	1	0	0	1	0	1	3
HyperGBM	1	1	0	0	0	1	3
Hyperopt-sklearn	1	1	0	1	0	0	3
Igel	1	1	0	0	1	0	3
Lightwood	1	1	0	0	1	0	3
AutoKeras	1	1	0	0	0	0	2
AutoML Alex	1	1	0	0	0	0	2
Auto-PyTorch	1	1	0	0	0	0	2
carefree-learn	1	1	0	0	0	0	2
mlr3automl	1	0	0	0	0	1	2
OBOE	1	1	0	0	0	0	2
TPOT	1	1	0	0	0	0	2
GAMA	1	0	0	0	0	0	1
Sum	19/19	15/19	2/19	6/19	6/19	6/19	

features (E8A). Most also simplify feature rescaling (E8B), although, as exemplified by the Ludwig package, this may sometimes need a flag to be set when employing AutoML techniques. We did not consider toggle buttons as invalidating the presence of automation, provided that, once set, users do not need to configure the procedure manually, i.e. the system takes care of it.

On the other end of the availability spectrum, only PyCaret and Auto_ViML assist with binning continuous variables (E8C). Granted, this process is not essential for many ML problems, but it does highlight the differentiation between tools. As for text (E8D) and temporal data (E8E), both formats provide their own challenges. It is clear from their correlated coverage in Table 21 – four out of six tools that can process one can also process the other – that a small subset of comprehensive AutoML systems is prioritising problem extensibility, e.g. supporting NLP and time-series forecasting. Finally, there is the matter of class imbalance (E8F). The statistics of observable data in real-world settings can be highly skewed, leading to flawed ML models and issues with fairness. Hence, the dearth of open-source AutoML packages that assist with or automate sampling-based corrections is somewhat surprising, suggesting a mismatch between academic and industrial expectations of ML challenges. Whatever the reason, class imbalance remains a risk for non-technical users, both in terms of detection and mitigation, that can stymie satisfactory outcomes and, consequently, AutoML democratisation.

Admittedly, some of the above processes blur with the phase of feature engineering. The nuance here is perhaps intent; data preparation aims to establish a baseline of information, while feature engineering seeks to improve the quality of this baseline via further transformation. With that

Table 22. Scores for open-source comprehensive AutoML systems (E9–E13). Evaluates the extent of automation for dataset feature generation (E9), reuse (E10), and selection (E11). Also evaluates whether HPO is completely specified by default (E12) and whether optimisation involves an entire ML pipeline (E13). See Table 3 for rubric.

Name	E9	E10	E11	E12	E13	Sum (out of 7)
TPOT	2	0	2	1	1	6
Auto_ViML	1	0	2	1	1	5
AutoML Alex	2	0	2	1	0	5
HyperGBM	2	0	2	1	0	5
auto-sklearn	0	0	2	1	1	4
mlr3automl	0	0	2	1	1	4
PyCaret	1	0	2	1	0	4
Mljar	0	0	2	1	0	3
GAMA	0	0	0	1	1	2
Hyperopt-sklearn	0	0	0	1	1	2
OBOE	0	0	0	1	1	2
AutoGluon	0	0	0	1	0	1
AutoKeras	0	0	0	1	0	1
Auto-PyTorch	0	0	0	1	0	1
carefree-learn	0	0	0	1	0	1
FLAML	0	0	0	1	0	1
Igel	0	0	0	1	0	1
Lightwood	0	0	0	1	0	1
Ludwig	0	0	0	1	0	1
Sum	8/38	0/19	16/38	19/19	7/19	

Table 23. Distribution of scores for open-source comprehensive AutoML systems (E9). Evaluates the extent of automation for dataset feature generation. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

E9	
Score	# Tools
0	14
1	2
2	3
U	0

perspective, Table 22 reveals which comprehensive open-source systems embrace the generative half of AutoFE. It turns out, as Table 23 attests, there are not many. Only three tools incorporate feature generation as part of their core automation: AutoML Alex, HyperGBM, and TPOT. Another two, namely Auto_ViML, and PyCaret, assist users in manually configuring/running feature generation. Naturally, certain transformations may be relevant/applicable to more than one ML application, so, for the sake of consistency and effort minimisation, e.g. ensuring that all departments in an organisation agree on a ‘sales’ calculation, it would also be valuable to retain such knowledge for reuse (E10). However, as already discussed for other criteria, no open-source system has implemented the level of persistence required to support such a mechanism.

Table 24. Distribution of scores for open-source comprehensive AutoML systems (E11). Evaluates the extent of automation for dataset feature selection. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

E11	
Score	# Tools
0	11
1	0
2	8
U	0

Now, every system that touches feature generation in Table 22 also has procedures for their sampling. In fact, Table 24 shows that feature selection (E11) is sometimes available even when generation is not. Moreover, all eight systems that support feature selection tend to do so at a substantially automated level. This outcome means that an interested stakeholder can potentially mix and match the surveyed comprehensive AutoML systems with the dedicated tools in Section 4.1.1. We also stress that, just because a system offers a specific capability, this review does not necessarily judge the quality and granular extent of that capability. So, for instance, TPOT can explore ‘polynomial’ features, but requirements for deeper feature synthesis are better satisfied elsewhere.

Table 25. Distribution of scores for open-source comprehensive AutoML systems (E13). Evaluates whether optimisation involves an entire ML pipeline. Scores: 0 for no, 1 for yes, and U for unclear.

E13	
Score	# Tools
0	12
1	7
U	0

Finally, although later criteria address the topic in greater depth, an operational-efficiency assessment also has to consider the model-development phase. Indeed, it can be daunting for a user to specify HPO search spaces or algorithms, so ensuring that a comprehensive system has a default procedure for undertaking CASH (E12) is paramount. Unsurprisingly, given that this review defines comprehensive systems to revolve around model development, exceptions that force a user to configure HPO are virtually nonexistent. In contrast, optimising an extended ML pipeline (E13), complete with data preprocessing and model postprocessing components, is still relatively rare. Table 25 shows only seven out of 19 open-source packages do so. Such a state of the ‘market’ makes sense, as CASH+ involves managing a much more expansive hyperparameter space; this requires implementing even more advanced theoretical techniques.

Switching from operational to technical efficiency, this monograph notes that the computational load of the model-development stage is mostly intrinsic to the algorithms implemented by an AutoML developer team. It is beyond the scope of this review to benchmark which packages take less time and memory than others to tackle any specific ML problem. However, we do note that none of the surveyed open-source comprehensive AutoML systems offers workload optimisation (E14). Such an absence may be noticeable to large-scale enterprise deployments, which wrestle with many non-negligible costs, e.g. those of cloud computing. Likewise, it is arguable that workload optimisation could also benefit ML applications on local hardware, where an experimental/production

Table 26. Scores for open-source comprehensive AutoML systems (E14–E16). Evaluates the availability of workload optimisation (E14), modelling time limits (E15), and modelling iteration limits (E16). See Table 5 for rubric.

Name	E14	E15	E16	Sum (out of 4)
FLAML	0	1	1	2
HyperGBM	0	1	1	2
Hyperopt-sklearn	0	1	1	2
AutoKeras	0	0	1	1
AutoML Alex	0	1	0	1
Auto-PyTorch	0	1	0	1
auto-sklearn	0	1	0	1
GAMA	0	1	0	1
Mljar	0	1	0	1
mlr3automl	0	1	0	1
OBOE	0	1	0	1
TPOT	0	1	0	1
Auto_ViML	0	0	0	0
AutoGluon	0	0	0	0
carefree-learn	0	0	0	0
Igel	0	0	0	0
Lightwood	0	0	0	0
Ludwig	0	0	0	0
PyCaret	0	0	0	0
Sum	0/38	11/19	4/19	

Table 27. Distribution of scores for open-source comprehensive AutoML systems (E15). Evaluates the availability of modelling time limits. Scores: 0 for absent, 1 for present, and U for unclear.

E15	
Score	# Tools
0	8
1	11
U	0

Table 28. Distribution of scores for open-source comprehensive AutoML systems (E16). Evaluates the availability of modelling iteration limits. Scores: 0 for absent, 1 for present, and U for unclear.

E16	
Score	# Tools
0	15
1	4
U	0

environment may be highly constrained. Managing costs and resources for ML work can be challenging even for a technician. Fortunately, Table 27 shows that more than half of the surveyed tools allow users to limit model development by time. Trial limits are also offered, although, as

Table 28 shows, not as broadly. This disparity is understandable, as iteration limits tend to be a step more technical than time limits, requiring a configurer to understand what is being iterated, i.e. they need to know the algorithms involved in model development.

Table 29. Scores for open-source comprehensive AutoML systems (DD1). Evaluates the extent of automation for cleaning dirty data. See Table 6 for rubric.

Name	DD1
Auto_ViML	2
AutoGluon	2
AutoKeras	2
AutoML Alex	2
Auto-PyTorch	2
auto-sklearn	2
carefree-learn	2
FLAML	2
GAMA	2
HyperGBM	2
Hyperopt-sklearn	2
Igel	2
Lightwood	2
Ludwig	2
Mljar	2
mlr3automl	2
OBOE	2
PyCaret	2
TPOT	2
Sum	38/38

4.2.2 *Dirty Data.* At this point, open-source comprehensive AutoML systems have been assessed primarily on how they enhance operational efficiency. Part of that analysis has examined the extent of automation during the preparation and subsequent engineering of dataset features. However, this review defines such tasks under the assumption that incoming data is essentially clean, if perhaps sampled awkwardly, e.g. with a class imbalance. In contrast, real-world data has to contend with many gaps and formatting issues. In fact, the ‘dirtiness’ of these inputs is a significant and ubiquitous challenge in industrial applications, relatively avoidable for many academic research projects, which is why its management warrants a distinct criterion. Given that context, it is perhaps surprising then that every surveyed tool in Table 29 automates, at least to some extent, the cleaning of dirty data (DD1).

Naturally, the cleaning capabilities of the surveyed open-source systems are varied, as indicated by Table 30. There is only one quasi-consistency that Table 31 shows, in that almost all tools can infer the types of data passed into the system (DD2), potentially enabling specialised cleaning/processing. The Ludwig package is a rare exception. In fact, Ludwig appears to serve as a platform that is almost fully controlled via configuration files. This reliance on manual specification means that the system barely scrapes into consideration as AutoML software. Granted, required configurability is not an immediate disqualification, provided that the finer details of a process are sufficiently mechanised.

Table 30. Scores for open-source comprehensive AutoML systems (DD2–DD5). Evaluates the extent of automation for data-type inference (DD2), missing-value imputation (DD3), and outlier management (DD4). Also evaluates the existence of domain-specific/advanced cleaning operations (DD5). See Table 6 for rubric.

Name	DD2	DD3	DD4	DD5	Sum (out of 6)
PyCaret	1	2	2	0	5
AutoML Alex	1	1	2	0	4
OBOE	1	1	2	0	4
auto-sklearn	1	2	0	0	3
Igel	1	2	0	0	3
TPOT	1	2	0	0	3
Auto_ViML	1	1	0	0	2
Auto-PyTorch	1	1	0	0	2
carefree-learn	1	1	0	0	2
HyperGBM	1	1	0	0	2
Ludwig	0	2	0	0	2
Mljar	1	1	0	0	2
mlr3automl	1	1	0	0	2
AutoGluon	1	0	0	0	1
AutoKeras	1	0	0	0	1
FLAML	1	0	0	0	1
GAMA	1	0	0	0	1
Hyperopt-sklearn	1	0	0	0	1
Lightwood	1	0	0	0	1
Sum	18/19	18/38	6/38	0/19	

Table 31. Distribution of scores for open-source comprehensive AutoML systems (DD2). Evaluates the extent of automation for data-type inference. Scores: 0 for absent, 1 for present, and U for unclear.

DD2	
Score	# Tools
0	1
1	18
U	0

In any case, the treatment of missing values (DD3) is an example of varied capability, with open-source packages scoring a summed 18 out of a maximal 38. The distribution in Table 32 reveals that this is simply a result of statistical spread, with AutoML developers almost uniformly choosing between ignorance, automated detection alone, and automated detection/resolution. Where applicable, packages often differ in imputation strategy, regardless of whether they employ convenience features or outright automation. For instance, TPOT, Mljar and AutoML Alex prefer taking a median of observed values, while Igel, PyCaret, Auto-PyTorch and mlr3automl replace missing values with a mean. Both auto-sklearn and OBOE are a little more sophisticated, as imputation methods are considered part of an ML pipeline with optimisable hyperparameters. As for the Ludwig package, it curiously defaults to replacement by zero values. Additionally, this survey could not easily determine the technical details employed by HyperGBM, carefree-learn, and Auto_ViML.

Table 32. Distribution of scores for open-source comprehensive AutoML systems (DD3). Evaluates the extent of automation for missing-value imputation. Scores: 0 for none, 1 for automatic detection, 2 for automatic detection/resolution, and U for unclear.

DD3	
Score	# Tools
0	6
1	8
2	5
U	0

Table 33. Distribution of scores for open-source comprehensive AutoML systems (DD4). Evaluates the extent of automation for outlier management. Scores: 0 for none, 1 for automatic detection, 2 for automatic detection/resolution, and U for unclear.

DD4	
Score	# Tools
0	16
1	0
2	3
U	0

Outlier management (DD4) is a far rarer capability among open-source systems, which makes sense when considering that a missing value is easier to identify than a statistical anomaly. For the same reason, Table 33 shows that if an AutoML developer bothers with detection, they will likely pursue full resolution. Admittedly, the solution is typically a simple excision of anomalous data instances. Regardless, minor notes for this sub-criterion include PyCaret requiring the user to set a flag when seeking to find/process anomalies, AutoML Alex employing the interquartile range (IQR) method, and OBOE exploring various techniques for outlier management. Beyond these mechanisms, this survey found nothing else among the assessed tools that might qualify as advanced or domain-specific cleaning (DD5).

4.2.3 Completeness and Currency. While the ability to polish data and enhance its information content is crucial to ML applications, end-to-end AutoML will probably always revolve most tightly around the model-development phase of an MLWF. Stakeholders deciding on which software to engage with will likely ask a fundamental question: what problems can these ML models solve? The answer, at least for the surveyed open-source systems, is provided in Table 34.

Of all the assessed capabilities, unsupervised learning (CC1) would appear to be the most unusual. While the process can still be expressed under the regular aims of finding a mathematical model that best approximates a desirable function, this desirable function has no prior supervisory hints as to what it may be. Indeed, what typically happens is that principled techniques, such as clustering methods that minimise intra-cluster distances or inertia, are thrown at a problem in an exploratory fashion with the hope of teasing out patterns/insights that may satisfy stakeholders. Accordingly, there is rarely any consistent way to apply CASH, as the objective function to maximise is human interest, with arcane dependencies on real-world context. The only real exception is if the outcomes of an unsupervised-learning process help train subsequent ML predictors/prescriptors, the performances of which can be tested and validated. In this case, automated unsupervised learning can be seen as a sophisticated form of feature generation. Given all these nuances, Table 35

Table 34. Scores for open-source comprehensive AutoML systems (CC1–CC9). Evaluates capabilities for unsupervised learning (CC1), regression on tabular data (CC2), standard classification on tabular data (CC3), multi-class classification on tabular data (CC4), time-series forecasting (CC5), image-based problem solving (CC6), text-based problem solving (CC7), multi-modal problem solving (CC8), and ensemble techniques (CC9). See Table 7 for rubric.

Name	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9	Sum (out of 17)
TPOT	2	2	2	2	2	2	2	0	2	16
AutoKeras	0	2	2	2	2	2	2	1	2	15
auto-sklearn	1	2	2	2	2	1	1	0	2	13
Ludwig	2	2	2	2	1	1	1	0	2	13
AutoGluon	0	2	2	2	2	2	2	0	0	12
Igel	1	2	2	2	1	1	2	0	1	12
PyCaret	2	2	2	2	2	2	0	0	0	12
FLAML	2	1	2	2	1	1	1	0	1	11
Mljar	1	2	2	2	2	0	2	0	0	11
Auto-PyTorch	1	2	2	2	1	1	1	0	0	10
Auto_ViML	0	2	2	2	2	0	0	0	0	8
AutoML Alex	0	2	2	2	2	0	0	0	0	8
carefree-learn	2	2	2	2	0	0	0	0	0	8
GAMA	1	1	2	2	1	0	1	0	0	8
HyperGBM	0	2	2	2	0	0	0	0	2	8
Hyperopt-sklearn	0	2	2	2	1	1	0	0	0	8
Lightwood	0	2	2	2	1	0	1	0	0	8
OBOE	1	1	1	1	1	1	1	0	1	8
mlr3automl	1	1	1	1	1	1	1	0	U	7
Sum	17/38	34/38	36/38	36/38	25/38	16/38	18/38	1/19	13/38	

Table 35. Distribution of scores for open-source comprehensive AutoML systems (CC1). Evaluates capabilities for unsupervised learning. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC1	
Score	# Tools
0	7
1	7
2	5
U	0

is therefore notable for showing that unsupervised learning still has a healthy representation among surveyed tools, both in terms of convenience features and automation.

Nonetheless, as expected, supervised learning is the predominant ML process that all comprehensive AutoML systems prioritise. Tabular data tends to be the most simple structure to deal with, so regression (CC2) and classification (CC3) capabilities are well represented, as shown by Table 36 and Table 37, respectively. Only two packages constitute the difference, not quite automating regression while still doing so with classification. As for multi-class classification (CC4), both the detailed Table 34 and the aggregated Table 38 indicate that it seems just as easy to offer as the standard binary version.

Table 36. Distribution of scores for open-source comprehensive AutoML systems (CC2). Evaluates capabilities for regression on tabular data. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC2	
Score	# Tools
0	0
1	4
2	15
U	0

Table 37. Distribution of scores for open-source comprehensive AutoML systems (CC3). Evaluates capabilities for standard classification on tabular data. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC3	
Score	# Tools
0	0
1	2
2	17
U	0

Table 38. Distribution of scores for open-source comprehensive AutoML systems (CC4). Evaluates capabilities for multi-class classification on tabular data. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC4	
Score	# Tools
0	0
1	2
2	17
U	0

Table 39. Distribution of scores for open-source comprehensive AutoML systems (CC5). Evaluates capabilities for time-series forecasting. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC5	
Score	# Tools
0	2
1	9
2	8
U	0

Problems based on time-series forecasting (CC5) are not too exotic, and Table 39 shows their coverage is moderate, with slightly fewer packages providing full automation over convenience features. However, once stakeholder aims move to text-based problems (CC6) and image-based

Table 40. Distribution of scores for open-source comprehensive AutoML systems (CC6). Evaluates capabilities for image-based problem solving. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC6	
Score	# Tools
0	7
1	8
2	4
U	0

Table 41. Distribution of scores for open-source comprehensive AutoML systems (CC7). Evaluates capabilities for text-based problem solving. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC7	
Score	# Tools
0	6
1	8
2	5
U	0

Table 42. Distribution of scores for open-source comprehensive AutoML systems (CC8). Evaluates capabilities for multi-modal problem solving. Scores: 0 for absent, 1 for present, and U for unclear.

CC8	
Score	# Tools
0	18
1	1
U	0

problems (CC7), there is a substantial drop-off in automated capability, as reflected in Table 40 and Table 41, respectively. Around a third of the surveyed systems do not support NLP tasks, image recognition, and the like. Understandably, the data structures involved tend to be more complex, usually warranting DL techniques that are more commonly found in AutoDL software. As for multi-modal problems (CC8), they are essentially a more challenging target of such capabilities, combining data sources in different formats, e.g. tabular information, free-form text, and images. It is not overly surprising that Table 42 depicts only one package operating in this space, i.e. AutoKeras.

Finally, although ensemble techniques (CC9) represent not an ML problem but a way to solve them, their inclusion in an ML application can dramatically change the relationship between an ML solution and a constituent ML model [308]. In short, depending on context/implementation, leveraging ensembles may well feel like tackling a different type of ML problem. Overall, Table 43 suggests that most open-source comprehensive systems avoid these techniques, with only five offering significant automation. Given that ensemble approaches are well known to be powerful in ML, e.g. reducing predictive variance without increasing bias, one reason for this may be the computational complexity of robustly tuning solutions with multiple predictors. How to do so effectively remains an open research question.

Table 43. Distribution of scores for open-source comprehensive AutoML systems (CC9). Evaluates capabilities for ensemble techniques. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC9	
Score	# Tools
0	10
1	3
2	5
U	1

Table 44. Coverage of model-selection processes for open-source comprehensive AutoML systems (CC10/CC11). Evaluates capabilities for custom evaluation metrics (CC10). Also evaluates the existence of HPO techniques (CC11), i.e. grid search (GR), random search (RA), Bayesian optimisation (BA), multi-armed bandit strategies (MAB), genetic/evolutionary algorithms (GE), and meta-learning (M). For convenience, additionally classifies whether HPO mechanisms beyond grid/random search (GR+) are provided. Scores: 0 for absent and 1 for present.

Name	CC10	GR	RA	BA	MAB	GE	M	SUM (out of 7)	GR+
AutoML Alex	1	1	1	0	1	1	0	5	1
carefree-learn	1	1	1	1	1	0	0	5	1
PyCaret	0	1	1	1	1	0	0	4	1
AutoGluon	1	0	1	1	1	0	0	4	1
auto-sklearn	1	0	0	1	0	0	1	3	1
FLAML	1	0	1	1	0	0	0	3	1
HyperGBM	1	0	1	0	0	1	0	3	1
mlr3automl	0	1	1	0	1	0	0	3	1
Auto_ViML	0	1	1	0	0	0	0	2	0
AutoKeras	1	0	0	1	0	0	0	2	1
GAMA	0	0	1	0	0	1	0	2	1
Hyperopt-sklearn	1	0	0	1	0	0	0	2	1
Igel	0	1	1	0	0	0	0	2	0
Lightwood	1	0	0	1	0	0	0	2	1
Mljar	0	1	1	0	0	0	0	2	0
TPOT	1	0	0	0	0	1	0	2	1
Auto-PyTorch	0	0	0	0	1	0	0	1	1
Ludwig	0	0	0	1	0	0	0	1	1
OBOE	0	0	0	0	0	0	1	1	1
Sum	10/19	7/19	11/19	9/19	6/19	4/19	2/19		16/19

Now, regardless of ML application, a core focus of comprehensive AutoML is finding the best ML solution for an ML problem; this involves optimising hyperparameters, which may or may not include the type of ML model itself. Optimisation, both at the low level of parameters – this is known as model training – and the high level of hyperparameters, is driven by objective functions. Customising such evaluation metrics (CC10) can prove very attractive to industry stakeholders, allowing them to define success for specific contexts and domains. For instance, recommendation engines may prefer to optimise accuracy for their top N suggestions rather than

Table 45. Distribution of scores for open-source comprehensive AutoML systems (CC10). Evaluates capabilities for custom evaluation metrics. Scores: 0 for absent, 1 for present, and U for unclear.

CC10	
Score	# Tools
0	9
1	10
U	0

deal with a thorough aggregate, even if the recommendation quality rapidly falls outside this top-N selection. After all, end-users are unlikely to scroll very far through a list of suggestions [241]. Accordingly, acknowledging the benefits, around half of the surveyed open-source systems offer metric customisability, as detailed by Table 44 and summed within Table 45.

As for HPO strategies (CC11), Table 44 elaborates on which types have currently found purchase amongst open-source software. The most basic form, grid search (GR), is provided by approximately a third of the surveyed tools. Of course, grid search is not renowned for its performance, partially because a regularly defined grid is somewhat arbitrary. A comical package named HungABunga [433] exemplifies this, offering users the ‘power’ of grid-search HPO at extreme granularity. Random search (RA) is then an alternative, although it is often lumped in with grid search as another unsophisticated approach. Such a perspective is why this review finds that 16 of 19 open-source systems go beyond grid/random search (GR+). In fairness, though, random search has proven to be remarkably competitive, especially for its simplicity and within highly dimensional spaces. Unsurprisingly, the technique is the most widely covered in this space, offered by ten packages.

Out of the more sophisticated methods, Bayesian approaches (BA) prove to be the most popular, offered by nine open-source systems. This result mimics their prevalence for dedicated HPO tools, discussed in Section 4.1.3. It remains challenging to determine whether Bayesian techniques are truly theoretically and implementationally optimal for the modern era of hardware or whether they simply profit from a first-mover advantage. Most likely, both factors play a part. The same goes for the popularity of multi-armed bandit strategies (MAB), represented by the Hyperband implementation, which has been embraced by the Bayesian optimisation community in the form of BOHB and associated methods. Finally, genetic algorithms and other evolutionary approaches (GE) have minor representation in four systems, although this relative sparsity may arguably be due to modern hardware still failing to leverage their relative robustness. Naturally, there are also other HPO techniques on offer that do not fit in the categories above, but they are typically confined to one AutoML framework each, e.g. fine-tuning based on hill-climbing for Mljar and BlendSearch for FLAML.

Optimisation strategies for AutoML are, of course, constantly evolving. Meta-learning (M), the ability to accelerate model development based on prior experience, is a noticeable gap in this space. Only two tools claim to provide such a capability, i.e. auto-sklearn and OBOE. Admittedly, effective meta-learning requires a challenging infrastructure to be in place. Local long-term storage is unlikely to be sufficiently informative, as a single user may not generate enough history to benefit their future ML applications. Instead, meta-learning is ideally fuelled by a hive repository that services many users, and developing/managing this may be out of reach for most open-source developers. Granted, the theory behind the concept is also continuously being explored and refined [64, 309, 344, 413]. Nonetheless, other innovations have surfaced around HPO implementations, primarily relating to configurability. For instance, AutoGluon allows users to toggle the ‘quality’ of an AutoML run, limiting fit/inference time. Likewise, FLAML has a ‘low_cost_partial_config’ parameter that

enables the quick generation of low-cost models. In the meantime, Mljar provides multiple modes of operation that tune the balance of expected technical performance and explainability from the desired ML solution, i.e. ‘explain’, ‘perform’, ‘compete’, and ‘Optuna’. This review finds that such customisability is more frequent among vendor-based AutoML products, as seen in Section 4.3, so the provision of such services among open-source systems suggests the acknowledgement that model validity is not everything; some users and scenarios prioritise speed or explainability.

Table 46. Coverage of popular libraries for open-source comprehensive AutoML systems (CC12).

Name	Sklearn	Keras	TF	XGBoost	LightGBM	CatBoost	PyTorch	Ax	R
Auto_ViML	1			1		1			
AutoGluon	1				1	1			
AutoKeras		1							
AutoML Alex	1			1	1	1			
Auto-PyTorch							1		
auto-sklearn	1								
carefree-learn								1	
FLAML	1			1	1				
GAMA	1								
HyperGBM				1	1	1			
Hyperopt-sklearn	1								
Igel	1								
Lightwood								1	
Ludwig		1							
Mljar	1			1	1	1			
mlr3automl									1
OBOE	1								
PyCaret	1			1	1	1			
TPOT	1								
Sum	12/19	1/19	1/19	6/19	6/19	6/19	2/19	1/19	1/19

The final major sub-criterion under completeness & currency notes which major algorithmic libraries each comprehensive system interfaces with (CC12). As Table 46 shows, Sklearn is dominant, which accords with the widespread modern uptake of Python for scientific programming. Such a concentration can be risky, as problems in an oversubscribed dependency will propagate to numerous AutoML packages. However, in practice, the community is large enough that any issues should be rapidly fixed. There is also a smaller but still notable presence of boosting libraries referenced within the open-source codebases, such as XGBoost, LightGBM, and CatBoost. Less well represented are the DL packages, i.e. Keras, TF, and Torch. Only one or two systems interface with each one. Of course, such a result would likely change if this review intended to dive deeper into NAS software. Finally, as outliers, one open-source system operates with the Facebook Ax library, while mlr3automl is linked with the R ecosystem. We use the programming language R to refer to this latter category, as, unlike with the Pythonic concentration of ML algorithms into a package like Sklearn, the dependencies of mlr3automl are splintered and hard to group.

As a side note, there is a sub-criterion dedicated to whether a piece of AutoML software is being actively maintained (CC13). However, as mentioned when discussing exclusions at the beginning of Section 4, every package in the main body of this review has seen significant activity in recent times. Faded systems of interest are listed in Appendix B.

4.2.4 Explainability. Despite research and debate stretching back decades, trust in AI has only really become of critical mainstream interest in recent years. This surge is partially driven by the

Table 47. Scores for open-source comprehensive AutoML systems (EX3–EX7). Evaluates capabilities for enhancing global interpretability (EX3), enhancing local interpretability (EX4), scenario analysis (EX5), bias/fairness assessment (EX6), and bias/fairness management (EX7). See Table 8 for rubric.

Name	EX3	EX4	EX5	EX6	EX7	Sum (out of 9)
Auto_ViML	2	0	0	0	0	2
AutoGluon	2	0	0	0	0	2
GAMA	2	0	0	0	0	2
Lightwood	2	0	0	0	0	2
Ludwig	2	0	0	0	0	2
Mljar	2	0	0	0	0	2
PyCaret	2	0	0	0	0	2
AutoKeras	0	0	0	0	0	0
AutoML Alex	0	0	0	0	0	0
Auto-PyTorch	0	0	0	0	0	0
auto-sklearn	0	0	0	0	0	0
carefree-learn	0	0	0	0	0	0
FLAML	0	0	0	0	0	0
HyperGBM	0	0	0	0	0	0
Hyperopt-sklearn	0	0	0	0	0	0
Igel	0	0	0	0	0	0
mlr3automl	0	0	0	0	0	0
OBOE	0	0	0	0	0	0
TPOT	0	0	0	0	0	0
Sum	14/38	0/38	0/19	0/38	0/38	

Table 48. Distribution of scores for open-source comprehensive AutoML systems (EX3). Evaluates capabilities for enhancing global interpretability. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

EX3	
Score	# Tools
0	12
1	0
2	7
U	0

accelerating uptake of ML technologies within broad sectors of industry and society, as well as the increasing friction between model outcomes and real-world nuances. Thus, with many open-source implementations of AutoML hailing from academic roots, it is not surprising that developers have been slow to prioritise the explainability requirements of non-technical users. Indeed, we found little evidence of facilities among the surveyed tools that help clarify data lineage (EX1) and modelling steps (EX2), although TPOT does produce pipeline visualisations and HyperGBM does provide experimental progress graphs via a dashboard. Overall, as indicated by Table 47, the only mechanisms related to explainability that have a significant presence among open-source systems are those related to global model interpretability (EX3). The provision of such services varies in quality and extensiveness, but Table 48 asserts that seven tools do so with sufficient automation.

At one end of the spectrum, AutoGluon provides a table of global feature importance, while, at the other, PyCaret supplies a dashboard and extensive graphs.

For now, no tool appears to enhance local interpretability (EX4), i.e. by allowing specific predictions to be analysed for the forces that drive them. Capabilities for scenario analysis (EX5) are similarly missing. This absence is likely to be keenly felt by stakeholders wishing to employ ML for prescriptive analytics, where the expected impact of an action is best evaluated against a counterfactual prediction, i.e. the outcome that occurs when the action is not taken. As for bias and fairness issues, the survey found no notable presence of mechanisms for either identification (EX6) or mitigation (EX7). Accordingly, the current open-source ‘market’ of AutoML limits itself in terms of appeal to industry stakeholders, at least from an explainability perspective. However, technically savvy users do have the option of integrating the dedicated tools listed in Section 4.1.2, once their own limitations have been acknowledged. Such a workaround may patch any relevant gaps within an MLWF. Furthermore, the dearth of explainability services in the open-source space needs to be seen in the context of ongoing trends; the next generation of AutoML software will likely have had more time to react to the current societal focus on trust in AI.

Table 49. Scores for open-source comprehensive AutoML systems (EU1–EU5). Evaluates the availability of interactions via coding (EU1), CLI (EU2), and GUI (EU3). Also evaluates software client type (EU4) and level of documentation (EU5). See Table 9 for rubric.

Name	EU1	EU2	EU3	EU4	EU5	Sum (out of 7)
GAMA	1	1	1	2	2	7
HyperGBM	1	1	1	2	1	6
Igel	1	1	1	2	1	6
Ludwig	1	1	0	0	2	4
TPOT	1	1	0	0	2	4
AutoGluon	1	0	0	0	2	3
AutoKeras	1	0	0	0	2	3
AutoML Alex	1	0	0	0	2	3
auto-sklearn	1	0	0	0	2	3
carefree-learn	1	0	0	0	2	3
Mljar	1	0	0	0	2	3
PyCaret	1	0	0	0	2	3
Auto_ViML	1	0	0	0	1	2
Auto-PyTorch	1	0	0	0	1	2
FLAML	1	0	0	0	1	2
Hyperopt-sklearn	1	0	0	0	1	2
Lightwood	1	0	0	0	1	2
mlr3automl	1	0	0	0	1	2
OBOE	1	0	0	0	1	2
Sum	19/19	5/19	3/19	6/38	29/38	

4.2.5 *Ease of Use.* Assessing open-source comprehensive AutoML systems for their problem-solving capabilities is all fine and well, but even the most mechanised frameworks will be avoided if stakeholders struggle to interface with them. Thus, the ease-of-use criterion aims to gauge accessibility to some degree, although deeper discussions of HCI are better reserved for another review [313]. Of course, the codebases of all open-source tools are, by definition, open to the public,

Table 50. Distribution of scores for open-source comprehensive AutoML systems (EU2). Evaluates the availability of interactions via CLI. Scores: 0 for absent, 1 for present, and U for unclear.

EU2	
Score	# Tools
0	14
1	5
U	0

Table 51. Distribution of scores for open-source comprehensive AutoML systems (EU3). Evaluates the availability of interactions via GUI. Scores: 0 for absent, 1 for present, and U for unclear.

EU3	
Score	# Tools
0	16
1	3
U	0

so Table 49 starts off by highlighting that all surveyed systems can be used via scripting. Most of the packages are written in Python, so the technical barrier is arguably lower than it could be. Nonetheless, for non-technical users, a simple CLI (EU2) is the next step in accessibility. Immediately, support for this kind of HCI drops from 19 packages to five, as Table 50 shows. Furthermore, only three of these five, as aggregated by Table 51, bother to develop a GUI (EU3).

Table 52. Distribution of scores for open-source comprehensive AutoML systems (EU4). Evaluates software client type. Scores: 0 for desktop only, 1 for browser only, 2 for desktop or browser, and U for unclear.

EU4	
Score	# Tools
0	16
1	0
2	3
U	0

Table 53. Distribution of scores for open-source comprehensive AutoML systems (EU5). Evaluates level of documentation. Scores: 0 for none, 1 for partial, 2 for extensive, and U for unclear.

EU5	
Score	# Tools
0	0
1	9
2	10
U	0

Given these results, it will become evident in Section 4.3 that commercial AutoML products are much more concerned with accessibility. Here, this contrast immediately suggests that the

notion of democratisation remains a buzzword in the academically rooted open-source sphere, at least compared to the goal of ‘data-science enhancement’ specified in Section 3.3. This reasoning also explains why few non-commercial packages are designed for accessible clients (EU4), e.g. as a browser-based application rather than desktop software. Only the rare GUI providers do so, i.e. GAMA, HyperGBM, and Igel. To their credit, as noted in Table 52, they support both browser *and* desktop access. In any case, Table 53 indicates that all surveyed systems are reasonably well documented on average, although the breadth and detail vary considerably. For instance, at the time of review, mlr3automl provided only a readme file on its GitHub page, admittedly including a vignette, whereas PyCaret was supported by comprehensive and sectioned documentation hosted on a separate website.

At this point, we highlight two open-source packages that did not make the cut for the core survey but still exemplify the diversity of developmental efforts in AutoML. In particular, they implement unique approaches to HCI. First is Libra [496], which contains an NLP engine intending to translate naturally stated queries into machine-understandable instructions. Such a framework marks a rare case of an AutoML tool tackling automation at the earliest stages of an MLWF, i.e. problem formulation & context understanding. Significant advances of this type would make it easier for stakeholders beyond data scientists to translate a business objective into an ML problem, e.g. by establishing whether an intended task can be reframed as unsupervised clustering or supervised classification/regression. The other noteworthy package is Otto [154], which operates as a chatbot. A user essentially answers a series of questions and is returned a scikit-learn script that can be run to generate desired ML models. While not technically a ‘comprehensive’ system, as an ML model is not explicitly created and managed, Otto thus demonstrates an alternative way to view ML applications. Ultimately, time will tell whether Libra and Otto are simply anomalies or whether they are early forerunners to a future wave of HCI innovation in the technological space of AutoML.

4.2.6 Remaining Criteria. By and large, this review found that open-source AutoML systems do not address the last criteria related to performant ML. Specifically, none of the surveyed tools has invested in the automation of any deployment & management effort (DM1–DM11), while mechanisms for governance & security (G1–G3) are likewise absent. This result may be partially due to all packages operating primarily offline. That stated, some of the software does offer assistance for deployment, which is worth noting. For instance, AutoGluon and PyCaret both provide tutorials on how to productionise their generated ML models. Additionally, research run by the AutoGluon team has also investigated model compression [211]. Given its close links to Amazon Web Services (AWS), the former unsurprisingly promotes SageMaker for deployment, while the latter suggests various options. As for Ludwig and Igel, both wrap FastAPI within assistive functions to build an API endpoint for a trained model. Then there is the carefree-learn package, which seemingly has a sister ‘deployment’ repository [259], albeit with little detail and documentation. Ultimately though, while governance & security can presently be excused as luxury features, the gap in automated coverage along the latter stages of an MLWF is glaring, at least for organisations that rely on running real-world ML applications regularly. Perhaps managing deployment and ongoing maintenance, i.e. MLOps, is too costly and out of scope for most open-source developers. However, as will be seen, once money gets involved, the status quo for AutoML services shifts.

4.3 Comprehensive AutoML Systems: Commercial

As reviewed at length, interested stakeholders have numerous open-source tools available that can automate parts of an ML application. A decent amount is ‘comprehensive’ in that the packages can manage many core tasks around generating a good ML model without any human involvement. However, there are apparent gaps in their coverage and priorities relating to performant ML. For

instance, there is very little in the way of deploying and maintaining an ML solution, i.e. the latter phases of an MLWF. Additionally, open-source UIs are not typically designed with general accessibility in mind. Thus, one may argue that, with its academic roots, open-source AutoML software is almost intended as an optional addendum to an ML application; its focus is to enhance the productivity of already technically capable data scientists. However, as Section 3.1 detailed, there are many non-technical stakeholders who are keen to leverage ML but are unwilling to spend time and money on data-science training. At the broader scale, it is not even feasible to satiate the surging demand for robust data-driven decision-making by educating more data scientists alone. Ideally, autonomously operating ML systems will be what meets these industry-wide needs. Nevertheless, the point here is that lay users require a far greater degree of hand-holding through the processes of an ML application. Accordingly, there is a profit to be made for any vendors willing to provide services in this market, but, for that coin to be earned, some product differentiation must necessarily exist. In light of this, we review the commercial comprehensive AutoML systems that have arisen to meet the broader demand of industries and organisations.

Now, to some extent, this survey collated a list of commercial packages in the same manner as for open-source software, i.e. via popular search engines, blogs, and other relevant websites. Examples of common information sources include AnalyticsVidhya [75], KD Nuggets [26], Gartner [16], and Wikipedia [557]. However, the commercial sphere does introduce novel nuances. For instance, with vendors competing for clientele, some summary software details could be inspected on comparison sites, such as G2 [46] and AlternativeTo [3].

Another consequence of competition is that developers in this space have seemingly sought to find and occupy niches with greater intensity than in the open-source market. Sometimes, this differentiation appears at the fringes of the provided AutoML services, given that most vendors still aim to supply ‘comprehensive’ coverage of an MLWF. However, there are ‘dedicated’ commercial tools as well, even if they are a mix of too few and too opaque to warrant the extended discussion their open-source counterparts received in Section 4.1. As examples of such dedicated commercial tools, Hazy [24] and MostlyAI [30] both focus on data preparation, fleshing out model-training inputs with statistically similar synthetic data. There is an added business benefit: using fake data maintains privacy around the real instances. Elsewhere, provided that a user does all the preliminary coding and modelling, there are platforms for logging, sorting, viewing and assessing experiments. These include Neptune [32], Comet [11], and DeterminedAI [19]. Then there is MLOps, another area of AutoML tool dedication. The associated packages that *only* focus on this facet are limited and do not qualify under a review of comprehensive systems, but they highlight that the deployment and maintenance phases of an MLWF are of surging interest.

Table 54. The list of surveyed comprehensive AutoML systems that are commercial and active.

Name	Ref.
Alteryx	[4]
Auger	[5]
SageMaker (AWS)	[40]
B2Metric	[6]
Big Squid	[7]
BigML	[8]
cnvrg.io	[10]
Compellon	[12]
D2iQ	[13]

Databricks	[14]
Dataiku	[15]
DataRobot	[17]
Deep Cognition	[18]
Einblick	[23]
Cloud AutoML (Google)	[9]
H2O	[242]
Watson Studio (IBM)	[48]
Knime	[27]
Azure AutoML (Microsoft)	[29]
MyDataModels	[31]
Number Theory	[34]
RapidMiner	[39]
Viya (SAS)	[47]
Spell	[43]
TIMi	[45]

Of course, to keep the survey feasible and maximally relevant to the aims of this review, a number of packages were further excluded. Some products appear to wrap around other AutoML software and thus do not contribute to unique commentary, e.g. Iguazio [25], Domino [21], and Qubole [38]. However, we also cannot overstate the challenge – completely expected – of analysing closed-source software, i.e. its opacity. Specifically, while a shortlist of 37 commercial products was compiled, only the 25 comprehensive AutoML systems in Table 54 supplied enough details for meaningful assessment. For completeness, the remaining 12 are listed in Appendix C. Even then, information on the selected 25 packages is replete with gaps, at least to the public eye, and many more ‘unclear’ scores are to be expected from their evaluation. Incidentally, Table 54 notes, in places, both the AutoML product and the overseeing organisation. We will often refer to the vendor names in the incoming analysis rather than a product name, as some software effectively operates within a grander ecosystem, implemented with deep connectivity to features drawn from a suite of other applications.

Before proceeding, it is worth making a few prefacing comments. First, it is interesting to consider the incoming assessment in the context of a 2020 Kaggle State of Machine Learning and Data Science report [295], which asked about the use of enterprise ML tools. The question sampled data scientists who use the dominant AWS, Google Cloud and Microsoft Azure platforms, so it is no surprise that the flagships SageMaker (16.5%), Google Cloud ML (14.8%) and Azure ML Studio (12.9%) received strong responses. However, 55.2% of the respondents claimed to use no ML tool on the cloud. The report also found that 33% of data scientists, presumably unbound by the platform-related subsampling, do not use AutoML tools. The remaining usage statistics were listed as 13.9% for Google Cloud AutoML, 9.5% for H2O Driverless AI, 8.4% for DataRobot AutoML, and 6.5% for Databricks AutoML. The takeaway is that, while AutoML has made definite inroads, the technology is still far from indispensable to ML practitioners. At the same time, its potential, alongside data science as a whole, is becoming increasingly recognised by industry every year. There has been a noteworthy spate of acquisitions in recent times, e.g. Sigopt by Intel in October 2020 [365], BigSquid and Compellon by Qlik and Clearsense, respectively, in September 2021 [454, 558], Ople.AI by Aktana in October 2021 [109], and Spell by Reddit in 2022 [253]. In

essence, large BI and data-processing companies appear to be chasing the competitive advantage that such capabilities can provide, which aligns very well with the democratising goal of AutoML.

With those perspectives in mind, the following findings are presented in a virtually identical fashion to the open-source software analysis in Section 4.2. Accordingly, we now proceed to discuss the criteria, organised as follows: efficiency in Section 4.3.1, dirty data in Section 4.3.2, completeness & currency in Section 4.3.3, explainability in Section 4.3.4, ease of use in Section 4.3.5, deployment & management effort in Section 4.3.6, and governance & security in Section 4.3.7. A final overarching commentary on commercial comprehensive AutoML systems is provided in Section 4.3.8.

Table 55. Scores for commercial comprehensive AutoML systems (E1–E3). Evaluates the existence of a model repository (E1), a model VCS (E2), and experiment tracking (E3). See Table 1 for rubric.

Name	E1	E2	E3	Sum (out of 4)
cnvrg.io	1	1	2	4
Databricks	1	1	2	4
Dataiku	1	1	2	4
DataRobot	1	1	2	4
Microsoft	1	1	2	4
SageMaker	1	1	2	4
SAS	1	1	2	4
Spell	1	1	2	4
Alteryx	1	0	2	3
Auger	1	0	2	3
Big Squid	1	0	2	3
BigML	1	1	1	3
Deep Cognition	0	1	2	3
Google	1	0	2	3
H2O	1	0	2	3
MyDataModels	1	0	2	3
RapidMiner	1	1	1	3
D2iQ	0	0	2	2
Einblick	U	U	2	2
IBM	1	0	1	2
Number Theory	1	0	1	2
KNIME	0	0	1	1
B2Metric	U	U	U	0
Compellon	0	0	0	0
TIMi	0	0	0	0
Sum	18/25	11/25	39/50	

4.3.1 *Efficiency.* Differentiation between open-source and commercial AutoML products is immediately evident from the moment effort in tracking and management during experimentation is assessed. The details are provided in Table 55. Indeed, vendors are geared for real-world applications, where the generation and refinement of ML models, before and after deployment, are expected to be iterative and exploratory. Thus, as summed by Table 56, 18 out of 25 commercial development teams have identifiably elected to incorporate some form of persistence via an ML model repository (E1). Such a mechanism typically serves as a basis for a model VCS (E2); sure enough, this direct

Table 56. Distribution of scores for commercial comprehensive AutoML systems (E1). Evaluates the existence of a model repository. Scores: 0 for absent, 1 for present, and U for unclear.

E1	
Score	# Vendors
0	5
1	18
U	2

Table 57. Distribution of scores for commercial comprehensive AutoML systems (E2). Evaluates the existence of a model VCS. Scores: 0 for absent, 1 for present, and U for unclear.

E2	
Score	# Vendors
0	12
1	11
U	2

Table 58. Distribution of scores for commercial comprehensive AutoML systems (E3). Evaluates the existence of experiment tracking. Scores: 0 for none, 1 for storage/access with limited automation/visuals, 2 for storage/access with automatic log visualisation, and U for unclear.

E3	
Score	# Vendors
0	2
1	5
2	17
U	1

upgrade is the case for ten packages. Overall, Table 57 highlights that 11 out of 25 surveyed systems provide the capability for stakeholders to roll back model iterations automatically. Although far from ubiquitous, this integration of robust software-development practices shows growing market maturity. At the very least, almost all AutoML products provide some form of experiment tracking (E3), and, as Table 58 shows, 17 out of these 22 do so at sophisticated levels, enabling users to inspect the history of an ML application conveniently.

Naturally, the 25 surveyed commercial systems have already attained a spectrum of scores at this point. Those ranked highest tend to have more of an MLOps focus, e.g. Databricks and cnvrg.io. A couple positioned lower in the rankings are trickier to assess, such as Einblink, previously Northstar, which has a unique modelling approach. Nonetheless, on average, the commercial AutoML sector cares more about this space than open-source developers. One possible interpretation is that academically sourced implementations are focussed on pushing the limits of full black-box automation, while vendors, maybe sensitive to business/legal ramifications, are much keener on allowing stakeholders to be as involved in an ML application, within reason, as they want.

The next sub-criteria of assessment revolve around the acceleration of an ML application via the leveraging of prior work and collaboration. In contrast to open-source systems, Table 59 indicates that commercial products have some decent coverage here. Perhaps appreciating that lay users do not have a background in ML to lean on, more than half of the surveyed systems provide

Table 59. Scores for commercial comprehensive AutoML systems (E4–E6). Evaluates the existence of a template/code repository (E4). Also evaluates capabilities for prior-work recommendation (E5) and project collaboration (E6). See Table 2 for rubric.

Name	E4	E5	E6	Sum (out of 5)
Databricks	2	0	2	4
Dataiku	2	0	2	4
cnvrg.io	2	0	1	3
DataRobot	2	0	1	3
H2O	2	0	1	3
Microsoft	1	0	2	3
SageMaker	2	0	1	3
SAS	2	0	1	3
Deep Cognition	2	0	0	2
IBM	1	0	1	2
KNIME	2	0	0	2
Number Theory	1	0	1	2
Alteryx	0	0	1	1
Einblick	U	0	1	1
Google	1	0	0	1
RapidMiner	1	0	0	1
Spell	0	0	1	1
Auger	U	U	U	0
B2Metric	U	0	U	0
Big Squid	0	0	0	0
BigML	0	0	0	0
Compellon	0	0	0	0
D2iQ	0	0	0	0
MyDataModels	0	0	0	0
TIMi	0	0	0	0
Sum	23/50	0/25	16/50	

Table 60. Distribution of scores for commercial comprehensive AutoML systems (E4). Evaluates the existence of a template/code repository. Scores: 0 for none, 1 for manual generation, 2 for automatic leveraging, and U for unclear.

E4	
Score	# Vendors
0	7
1	5
2	9
U	4

repositories for templates or coded scripts (E4), as aggregated by Table 60. In many of these cases, the frameworks are set up to support users in generating and sharing these documents. However, some packages, exemplified by SageMaker and its Jumpstart mechanism [96], allow projects to be automatically kickstarted by templates for common use cases. That stated, none of the 25

surveyed vendors seems to have a more sophisticated means of suggesting prior art (E5). This result is interesting to note, given the long-running academic focus on meta-learning and the era of intelligent discovery assistants (IDAs) that preceded the current wave of AutoML [64, 308, 344]. Admittedly, a deeper examination of other survey-excluded systems may be warranted ahead of any strident conclusions. For instance, TAZI [44], Iguazio [25] and Kortical [28] all claim to support off-the-shelf use cases, and Domino [21] does so similarly via its Domino Knowledge Center. However, within the survey sample, only Auger hints at possible meta-learning; a blog post mentions that its optimiser is selected appropriately for a given dataset, perhaps leveraging a history of similar ML problems [87].

Table 61. Distribution of scores for commercial comprehensive AutoML systems (E6). Evaluates capabilities for project collaboration. Scores: 0 for none, 1 for basic features, 2 for advanced features, and U for unclear.

E6	
Score	# Vendors
0	10
1	10
2	3
U	2

Mechanisms to support collaboration on a project (E6) are present in around half of the commercial AutoML systems. Most have been assessed by Table 61 to be fairly rudimentary, e.g. simply allowing project folders to be shared. For some instances, e.g. SageMaker [95] and IBM [48], a score of 1 was granted because, while notebooks themselves could be shared, each user had to work with a unique copy of the code, thus risking desynchronisation issues. In contrast, Microsoft and Databricks [180] exemplify a score of 2 being granted for the presence of live co-editing. That stated, there is a discussion to be had around the utility of such a feature, as it is unlikely to become a common software practice when such a robust industry-standard VCS such as Git exists. In the meantime, Einblick, the commercialisation of the academic product Northstar [324], supports a unique collaboration model via a whiteboard-style ‘visual analytics’ interface upon which users drag and drop connectable nodes. This form of non-standard HCI has obvious benefits but also several drawbacks. For one thing, the unusual approach means it is more difficult to assess the system for certain functions, e.g. around deployment and management. The jury is also out on whether such a hands-on UI, advocated in promotional material, will actually live up to its aims of enhancing productivity for data-science teams at scale. Granted, this uncertainty is to be expected for a nascent technology; the long-term success of radical innovations in AutoML will probably not be evident for many years.

Moving on to matters of effort reduction across an MLWF, we begin by assessing how a dataset ready for ML arises. Immediately evident, there is a stark contrast between Table 62 and the previous analysis for open-source tools: EDA (E7). Again, most vendors are geared for real-world applications run by relatively non-academic stakeholders, meaning that data sources are likely to be messy and, initially, poorly understood. Thus, most surveyed commercial systems facilitate some form of exploration, allowing issues, insights, structures and patterns in data to be sought out. However, Table 63 indicates varying levels of support, with nine packages still requiring users to click about with convenience features. Another nine go a step further into actual automation, generating visualisations to guide user interpretations, e.g. histograms, charts, and graphs. Only two, DataRobot and Compellon, advise a user how best to sift through these visuals, thus earning the highest score of 3. Indeed, it is a reoccurring theme in this review that automation is not binary,

Table 62. Scores for commercial comprehensive AutoML systems (E7/E8). Evaluates the extent of automation for assisted data exploration (E7) and data preparation (E8). See Table 3 for rubric.

Name	E7	E8	Sum (out of 5)
DataRobot	3	2	5
Auger	2	2	4
Big Squid	2	2	4
Dataiku	2	2	4
H2O	2	2	4
Microsoft	2	2	4
Compellon	3	0	3
IBM	2	1	3
Number Theory	2	1	3
Alteryx	1	1	2
cnvrg.io	1	1	2
D2iQ	2	0	2
Databricks	1	1	2
Einblick	1	1	2
Google	0	2	2
KNIME	1	1	2
MyDataModels	2	U	2
RapidMiner	1	1	2
SageMaker	1	1	2
SAS	1	1	2
B2Metric	0	1	1
BigML	0	1	1
Deep Cognition	0	1	1
TIMi	1	0	1
Spell	0	0	0
Sum	33/75	27/50	

Table 63. Distribution of scores for commercial comprehensive AutoML systems (E7). Evaluates the extent of automation for assisted data exploration. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, 3 for substantial automation with useful alerts, and U for unclear.

E7	
Score	# Vendors
0	5
1	9
2	9
3	2
U	0

and the mechanised generation of information alone is not ideal for effort reduction without being accompanied by some form of mechanised selection/emphasis. Nonetheless, given that contextual

insights are usually the domain of BI, with ML more focussed on the actual models, it is noteworthy that EDA has already been integrated so substantially into commercial AutoML.

Table 64. Distribution of scores for commercial comprehensive AutoML systems (E8). Evaluates the extent of automation for data preparation. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

E8	
Score	# Vendors
0	4
1	13
2	7
U	1

In contrast, there is less excuse for a comprehensive AutoML system to neglect the automation of data preparation (E8), given the substantial burden it places on a data scientist. Thus, the fact that Table 64 indicates more than half of the surveyed services provide only convenience features, with four vendors appearing to ignore the topic outright, suggests an oversight in the market. Stakeholders essentially need to be mindful of this hidden ‘upstream’ cost that an otherwise appealing product may neglect. That stated, while this review has systematised an MLWF that suggests an organic evolution of AutonoML technology, i.e. steadily outwards from the model-development phase, it is worth remembering that vendors pursuing business benefits need not subscribe to an academic view of AutoML. Indeed, Compellon and TIMi exemplify frameworks leaning towards exploratory analytics over the deployment of ML models, seemingly targeting business users with their messaging. Ignoring data preparation may be intentional on their part, and Section 4.3.8 provides some extended commentary on this fragmentation of the ‘commercial AutoML agenda’. Of course, it is also possible that the tools do undertake data-preparation work but simply do not promote it, possibly because it is not a priority.

Table 65. Coverage of data-preparation processes for commercial comprehensive AutoML systems (E8A–E8F). Evaluates categorical processing (E8A), standardisation/normalisation (E8B), bucketing/binning (E8C), text processing (E8D), time-period extraction (E8E), and management of class imbalance (E8F). Scores: 0 for absent and 1 for present.

Name	E8-A	E8-B	E8-C	E8-D	E8-E	E8-F	Sum (out of 6)
Dataiku	1	1	1	1	1	1	6
DataRobot	1	1	1	1	1	1	6
Google	1	1	1	1	1	0	5
Auger	1	1	0	0	1	1	4
H2O	1	0	1	1	1	0	4
Microsoft	1	1	0	1	1	0	4
Big Squid	1	1	0	0	0	0	2
Sum	7/7	6/7	4/7	5/7	6/7	3/7	

This closed-source opacity becomes murkier as analysis becomes steadily more granular, and the Table 65 breakdown of what is included under data preparation (E8A–E8F) can only be applied to a handful of packages. It is also no surprise that the table only lists the seven commercial systems that boast an automated form of this capability; the details effectively become part of their promotion.

In any case, it is notable that industry-heavyweight SageMaker Autopilot does not provide any automatic assistance with data preparation. Another minor comment is that the data-preparation mechanism employed by Google does work with weighted arrays, which no other vendors consider.

Table 66. Scores for commercial comprehensive AutoML systems (E9–E11). Evaluates the extent of automation for dataset feature generation (E9), reuse (E10), and selection (E11). See Table 3 for rubric.

Name	E9	E10	E11	Sum (out of 5)
DataRobot	2	1	2	5
Dataiku	1.5	1	2	4.5
B2Metric	2	0	2	4
H2O	2	0	2	4
RapidMiner	2	0	2	4
IBM	1	0	2	3
Microsoft	2	0	1	3
Alteryx	1.5	0	1	2.5
Compellon	0	0	2	2
Databricks	1	1	0	2
Einblick	1	0	1	2
KNIME	1	0	1	2
Number Theory	1	0	1	2
SageMaker	1	1	0	2
Auger	0	0	1	1
Big Squid	0	0	1	1
BigML	0	1	0	1
Deep Cognition	0	0	1	1
Google	0	0	1	1
MyDataModels	0	0	1	1
SAS	0	0	1	1
cnvrg.io	0	0	U	0
D2iQ	0	0	0	0
Spell	0	0	0	0
TIMi	0	0	0	0
Sum	19/50	4/25	25/50	

Table 67. Distribution of scores for commercial comprehensive AutoML systems (E9). Evaluates the extent of automation for dataset feature generation. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear. The 1.5 denotes existence of convenience plugins enabling automation.

E9	
Score	# Vendors
0	12
1	6
1.5	2
2	5
U	0

Turning to AutoFE, Table 66 shows that the surveyed vendors have a moderate coverage of this space. In fact, commercial systems are granted a higher average score for dataset feature generation (E9) than open-source systems in Section 4.2.1, i.e. 19/25 versus 8/19. However, Table 67 indicates that half of the surveyed systems still do not bother with this aspect, and half again of those that do choose to stick with convenience features, e.g. manually triggered functions that sum/average dataset columns. Notably, two AutoML services, Alteryx and Dataiku, are unique in that they support the automated generation of dataset features via an optional plugin. For this reason, they are marked with scores of 1.5, reflecting the blur between a convenience feature and an automated mechanism. As for the operator ‘nodes’ of KNIME, they are considered feature transformations that require a degree of user manipulation, so the associated system scores a 1.

Table 68. Distribution of scores for commercial comprehensive AutoML systems (E10). Evaluates the extent of automation for dataset feature reuse. Scores: 0 for absent, 1 for present, and U for unclear.

E10	
Score	# Vendors
0	20
1	5
U	0

Now, generating useful features can be a complex affair, so it would make sense to store them for future reuse (E10). This reasoning is especially the case where an organisation may use the same datasets for numerous ML projects run by different teams. Such a notion is similar to constructing/managing data marts in a data warehouse with corporately controlled metric registers, thus ensuring consistency and accelerating preparatory processes. Admittedly, one could argue that this is a luxury compared to other priorities, likely only to become more valuable as the AutoML market further matures. Nonetheless, Table 68 shows that four reviewed vendors have already considered the innovation.

Table 69. Distribution of scores for commercial comprehensive AutoML systems (E11). Evaluates the extent of automation for dataset feature selection. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

E11	
Score	# Vendors
0	6
1	11
2	7
U	1

Of course, whether they are generated or recalled, dataset features must be chosen carefully from the pool of available options when used in modelling. Overall, commercial systems assist strongly with dataset feature selection (E11), much more than generation, which mimics the trend seen for open-source tools. However, this varies substantially between packages, and Table 69 reveals that a majority of 11 offer convenience features, e.g. the ability to manually drop features via a checkbox/widget, sometimes assisted by an informed view of what to select. Another seven packages then go further and provide substantial automation. Interestingly, BigML, Auger, Microsoft and RapidMiner exemplify services that approach this process from a perspective of exclusion

rather than inclusion, advising which dataset features are ‘bad’. For instance, Microsoft avoids high cardinality, while RapidMiner targets those of ‘low quality’ and redundancies revealed by high correlations. Crucially, most of these selection processes, whether automated or not, are applied prior to the existence of any ML model, i.e. they are filter-style, not wrapper-style.

As a side note, this review could not make determinations for two sub-criteria regarding commercial AutoML, namely whether HPO search spaces and algorithms are already specified by default (E12) and whether HPO operates on an ML pipeline of components (E13). It is very likely that, for the former sub-criterion, all packages should be marked ‘yes’ with a 1, as model selection processes are generally obscured from users in a comprehensive system; they should not lock up in the absence of manual specification. Such a result would subsequently mean that the sub-criterion does not differentiate between any of the open-source and closed-source software presented in this review. However, there are open-source tools excluded from this survey that do force users to make choices prior to HPO, so the question on specification remains valid to ask. As for the latter sub-criterion, it is difficult to guess the structure of ML solutions under the hood of a commercial system. However, optimising extended pipelines is challenging from a theoretical/implementational standpoint. It is already rare among the more technically driven open-source offerings, even being elevated as a core innovation within a package named AutoWEKA4MCPS [162, 414, 481, 482]. Thus, most probably, few commercial systems, if any, will presently lump preprocessors and postprocessors into an HPO space.

Finally, on matters of technical efficiency, Table 70 shows that the commercial space of workload management is not too dissimilar from that of open-source AutoML systems. However, as aggregated by Table 71, there are uniquely a few packages that assist with workload optimisation (E14). Most of these provide only convenience features and notably have a promotional focus on MLOps, e.g. Spell, cnvrg.io and D2iQ. Granted, via the AWS ecosystem, SageMaker also has the option of plugging into Auto Scaling to manage relevant EC2 server instances. A similar capability may be possible for Google and Microsoft as well, but it has not been explicitly promoted in their cases.

Naturally, time limits (E15) and trial limits (E16) are available among AutoML systems, highlighted by Table 72 and Table 73, respectively. However, support for such modelling control is surprisingly rare, and providers of one are often those offering the other. It is also notable that, barring DataRobot, any packages supporting workload management facilitate either infrastructural optimisation or algorithmic constraints, but never both. This hints that there is a divide in the approach to resource management between vendors focussed on MLOps and those leaning into user-driven model development. As for DataRobot, it is the only commercial system that appears to substantially automate workload optimisation, iteratively increasing data size and computational resources as it reduces the number of candidate models during a search. This process, of course, is very similar to successive halving, so other products may do so, too, without explicitly promoting it. Whatever the case, the takeaway is that resource management has not yet received significant attention from either open-source or closed-source AutoML tools. Perhaps fine-tuning the technical efficiency of an ML application is not yet a priority compared with ensuring that training and deploying ML models are robust functions. Indeed, given reasonable cloud-computing prices and supplied operational efficiencies elsewhere, the sub-criteria above may not become vital differentiators until the market matures further.

4.3.2 Dirty Data. As already assessed for open-source tools, this criterion examines whether commercial AutoML systems can aid stakeholders with the initial cleaning of data (DD1) prior even to its preparation for ingestion by an ML model. Accordingly, the evaluations are displayed in Table 74. Further elaborated in Table 75, we find that most vendors only supply convenience features for this task; users still need to make decisions and take explicit action. However, eight

Table 70. Scores for commercial comprehensive AutoML systems (E14–E16). Evaluates the availability of workload optimisation (E14), modelling time limits (E15), and modelling iteration limits (E16). See Table 5 for rubric.

Name	E14	E15	E16	Sum (out of 4)
DataRobot	2	1	0	3
Auger	0	1	1	2
BigML	0	1	1	2
Databricks	0	1	1	2
Dataiku	0	1	1	2
Alteryx	0	0	1	1
cnvrg.io	1	0	0	1
D2iQ	1	0	0	1
H2O	0	1	0	1
IBM	0	0	1	1
Microsoft	0	1	0	1
SageMaker	1	0	0	1
Spell	1	0	0	1
B2Metric	0	0	0	0
Big Squid	0	0	0	0
Compellon	0	U	U	0
Deep Cognition	0	0	0	0
Einblick	0	0	0	0
Google	0	0	0	0
KNIME	0	0	0	0
MyDataModels	0	U	U	0
Number Theory	0	0	0	0
RapidMiner	0	0	0	0
SAS	0	0	0	0
TIMi	0	0	0	0
Sum	6/50	7/25	6/25	

Table 71. Distribution of scores for commercial comprehensive AutoML systems (E14). Evaluates the availability of workload optimisation. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

E14	
Score	# Vendors
0	20
1	4
2	1
U	0

packages do provide substantial automation, and seven of these were the same that also scored highest for the automation of data preparation in Table 62. This outcome is not surprising. If development effort is to be expended on the data-engineering phase of an MLWF, it makes sense to support cleaning and preparation simultaneously.

Table 72. Distribution of scores for commercial comprehensive AutoML systems (E15). Evaluates the availability of modelling time limits. Scores: 0 for absent, 1 for present, and U for unclear.

E15	
Score	# Vendors
0	16
1	7
U	2

Table 73. Distribution of scores for commercial comprehensive AutoML systems (E16). Evaluates the availability of modelling iteration limits. Scores: 0 for absent, 1 for present, and U for unclear.

E16	
Score	# Vendors
0	17
1	6
U	2

Of course, opacity remains an issue here as it did for data preparation, so Table 76 only breaks down capabilities for commercial systems that claim significant automation. As summarised by Table 77, almost all packages can identifiably infer data types (DD2). Likewise, Table 78 indicates that almost all packages can detect and resolve missing values (DD3), even though the preferred processes differ. For instance, Microsoft, DataRobot, Big Squid and Auger all appear solely to impute. Dataiku and H2O additionally provide an option to drop missing values. Embedding approaches are also viable, being supported by both Dataiku and Google. The latter vendor even seems capable of assessing the severity of a null value, which may suggest leaving it alone. Essentially, handling missing values appears to be a standard offering for automated data cleaning.

In contrast, outlier management (DD4) is a rarity. The reasons are the same as discussed in Section 4.2.2, and, sure enough, Table 79 finds only three services handling automatic detection of anomalies. Notably, the one that claims to automatically resolve outliers, i.e. DataRobot, is also the only comprehensive system within this review that supplies advanced cleaning operations (DD5). Scoring 1 in Table 80, the software is seemingly able to identify ‘inliers’, i.e. erroneous data points within the interior of a statistical distribution. Naturally, such an analysis is more complex than simply highlighting distant outliers.

4.3.3 Completeness and Currency. When assessing the types of ML problems that commercial comprehensive AutoML systems are designed for, Table 81 reveals a spectrum of coverage. The survey results are not entirely dissimilar to the evaluations in Section 4.2.3, although, for the summed scores, commercial software does have a less generalist/automated tail that goes below the open-source range of 7 to 16. Admittedly, opacity makes it harder to judge a couple of these outlying vendor products.

In any case, Table 82 suggests that it is almost a coin flip whether a package will assist/automate unsupervised learning (CC1) or ignore it entirely. Understandably, a lack of a clearly defined target can make it harder to provide case studies or business pitches that objectively highlight superior ML model validity. On the other hand, unsupervised learning meshes well with EDA for the systems that lean more in that direction. So, its provision ultimately depends on the agenda of a vendor. Supervised learning, however, is almost compulsory to offer. Noted by Table 83, which considers regression on tabular data (CC2), the vast majority of packages provide significant automation in

Table 74. Scores for commercial comprehensive AutoML systems (DD1). Evaluates the extent of automation for cleaning dirty data. See Table 6 for rubric.

Name	DD1
Auger	2
Big Squid	2
Dataiku	2
DataRobot	2
Google	2
H2O	2
Microsoft	2
MyDataModels	2
Alteryx	1
B2Metric	1
BigML	1
cnvrg.io	1
Databricks	1
Deep Cognition	1
Einblick	1
IBM	1
KNIME	1
Number Theory	1
RapidMiner	1
SageMaker	1
SAS	1
Compellon	0
D2iQ	0
Spell	0
TIMi	0
Sum	29/50

Table 75. Distribution of scores for commercial comprehensive AutoML systems (DD1). Evaluates the extent of automation for cleaning dirty data. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

DD1	
Score	# Vendors
0	4
1	13
2	8
U	0

this space. Given that supervised learning is the fundamental task that innovations in ML model selection originally targeted, comprehensive systems struggle to be labelled as AutoML if they do not even automate this core process. Scores for classification capabilities on tabular data, both standard (CC3) and multi-class (CC4), affirm this argument. They are summarised in Table 84 and

Table 76. Scores for commercial comprehensive AutoML systems (DD2–DD5). Evaluates the extent of automation for data-type inference (DD2), missing-value imputation (DD3), and outlier management (DD4). Also evaluates the existence of domain-specific/advanced cleaning operations (DD5). See Table 6 for rubric.

Name	DD2	DD3	DD4	DD5	Sum (out of 6)
DataRobot	1	2	2	1	6
H2O	1	2	1	0	4
Auger	1	2	0	0	3
Big Squid	1	2	0	0	3
Dataiku	1	2	0	0	3
Google	1	2	0	0	3
Microsoft	1	2	0	0	3
MyDataModels	U	1	1	0	2
Sum	7/8	15/16	4/16	1/8	

Table 77. Distribution of scores for commercial comprehensive AutoML systems (DD2). Evaluates the extent of automation for data-type inference. Scores: 0 for absent, 1 for present, and U for unclear.

DD2	
Score	# Vendors
0	0
1	7
U	1

Table 78. Distribution of scores for commercial comprehensive AutoML systems (DD3). Evaluates the extent of automation for missing-value imputation. Scores: 0 for none, 1 for automatic detection, 2 for automatic detection/resolution, and U for unclear.

DD3	
Score	# Vendors
0	0
1	1
2	7
U	0

Table 79. Distribution of scores for commercial comprehensive AutoML systems (DD4). Evaluates the extent of automation for outlier management. Scores: 0 for none, 1 for automatic detection, 2 for automatic detection/resolution, and U for unclear.

DD4	
Score	# Vendors
0	5
1	2
2	1
U	0

Table 80. Distribution of scores for commercial comprehensive AutoML systems (DD5). Evaluates the existence of domain-specific/advanced cleaning operations. Scores: 0 for absent, 1 for present, and U for unclear.

DD5	
Score	# Vendors
0	7
1	1
U	0

Table 81. Scores for commercial comprehensive AutoML systems (CC1–CC9). Evaluates capabilities for unsupervised learning (CC1), regression on tabular data (CC2), standard classification on tabular data (CC3), multi-class classification on tabular data (CC4), time-series forecasting (CC5), image-based problem solving (CC6), text-based problem solving (CC7), multi-modal problem solving (CC8), and ensemble techniques (CC9). See Table 7 for rubric.

Name	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9	Sum (out of 17)
DataRobot	2	2	2	2	2	2	2	1	2	17
H2O	2	2	2	2	2	2	2	0	2	16
Dataiku	2	2	2	2	2	2	1	0	2	15
Microsoft	1	2	2	2	2	1	1	0	2	13
BigML	2	2	2	2	2	2	0	0	0	12
Google	0	2	2	2	2	2	2	0	0	12
Number Theory	1	2	2	2	1	1	2	0	1	12
Einblick	1	2	2	2	2	0	2	0	0	11
RapidMiner	2	1	2	2	1	1	1	0	1	11
SageMaker	1	2	2	2	1	1	1	0	0	10
Alteryx	0	2	2	2	1	1	0	0	0	8
Auger	0	2	2	2	0	0	0	0	2	8
B2Metric	2	2	2	2	0	0	0	0	0	8
Big Squid	0	2	2	2	2	0	0	0	0	8
Databricks	0	2	2	2	2	0	0	0	0	8
IBM	0	2	2	2	1	0	1	0	0	8
KNIME	1	1	1	1	1	1	1	0	1	8
SAS	1	1	2	2	1	0	1	0	0	8
cnvrg.io	1	1	1	1	1	1	1	0	U	7
Deep Cognition	0	0	2	2	0	2	0	0	U	6
MyDataModels	0	2	2	2	0	0	0	0	U	6
Spell	0	1	1	1	1	1	1	0	0	6
TIMi	0	2	2	0	0	0	0	0	U	4
Compellon	U	U	2	U	U	U	U	U	U	2
D2iQ	U	U	U	U	U	U	U	U	0	0
Sum	19/50	39/50	45/50	41/50	27/50	20/50	19/50	1/25	13/50	

Table 85, respectively. Arguably, classification is slightly easier than regression, so it has even better coverage in terms of automation.

Again, as with open-source tools, Table 86 reveals a drop in coverage for time-series forecasting (CC5). Similarly, by the time discussion moves to image-based tasks (CC6) and NLP (CC7), Table 87 and Table 88 suggest, respectively, that these capabilities are now relatively niche. This result is

Table 82. Distribution of scores for commercial comprehensive AutoML systems (CC1). Evaluates capabilities for unsupervised learning. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC1	
Score	# Vendors
0	10
1	7
2	6
U	2

Table 83. Distribution of scores for commercial comprehensive AutoML systems (CC2). Evaluates capabilities for regression on tabular data. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC2	
Score	# Vendors
0	1
1	5
2	17
U	2

Table 84. Distribution of scores for commercial comprehensive AutoML systems (CC3). Evaluates capabilities for standard classification on tabular data. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC3	
Score	# Vendors
0	0
1	3
2	21
U	1

Table 85. Distribution of scores for commercial comprehensive AutoML systems (CC4). Evaluates capabilities for multi-class classification on tabular data. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC4	
Score	# Vendors
0	1
1	3
2	19
U	2

not surprising, as the more complex data inputs are typically reserved for AutoDL approaches, which, beyond being out of scope for this review, are currently nowhere near the technological maturity of standard AutoML. Consequently, if handling a single source of non-tabular data is a rare

Table 86. Distribution of scores for commercial comprehensive AutoML systems (CC5). Evaluates capabilities for time-series forecasting. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC5	
Score	# Vendors
0	5
1	9
2	9
U	2

Table 87. Distribution of scores for commercial comprehensive AutoML systems (CC6). Evaluates capabilities for image-based problem solving. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC6	
Score	# Vendors
0	9
1	8
2	6
U	2

Table 88. Distribution of scores for commercial comprehensive AutoML systems (CC7). Evaluates capabilities for text-based problem solving. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC7	
Score	# Vendors
0	9
1	9
2	5
U	2

Table 89. Distribution of scores for commercial comprehensive AutoML systems (CC8). Evaluates capabilities for multi-modal problem solving. Scores: 0 for absent, 1 for present, and U for unclear.

CC8	
Score	# Vendors
0	22
1	1
U	2

capability, the result shown in Table 89 is also not unexpected; only one commercial system claims to handle multi-modal tasks (CC8), i.e. DataRobot. Granted, AutoML frameworks that optimise ML pipelines could theoretically merge multiple preprocessing components dedicated to different data modalities, but it is unclear whether any existing implementation can genuinely support such a process. Finally, Table 90 shows that only eight vendors identifiably confront the nuances of tackling

Table 90. Distribution of scores for commercial comprehensive AutoML systems (CC9). Evaluates capabilities for ensemble techniques. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

CC9	
Score	# Vendors
0	12
1	3
2	5
U	5

ML problems with ensemble techniques (CC9), although the quantity of U scores is uniquely high, as methodological details tend to be obfuscated for closed-source software. It is possible that, in the early days of ML adoption by the business community, basic models are sufficient to provide substantial benefits already. The power of ensembles may not yet be warranted, especially given their typical complexity and associated adverse impact on explainability.

Now, as previously emphasised, an increasingly granular review of commercial AutoML confronts decreasing transparency around system details. Thus, while it is reasonably clear which ML problems can be addressed by the ML solutions provided by a package, it is far less obvious how these solutions are derived. Accordingly, over a third of the systems surveyed in Table 91 have information gaps regarding their mechanical processes around CASH. However, what is clear from Table 92, is that vendors are evenly split around whether they support customisable metrics (CC10). Recalling Section 4.3.1, an interpretation was suggested that many commercial AutoML features appear to favour the primacy of user control when compared to open-source tools, possibly to minimise potential liability, but the result here implies that there is still a broad range of how much freedom is genuinely supplied. For instance, in this particular space, DataRobot and Microsoft exemplify vendors making the decisions, while Google and SageMaker support user-driven customisation. Of course, a cynical perspective might be that it benefits commercial systems to sell the *impression* of user control while still making the choices where they actually matter. However, in fairness, there is a diversity of agendas and intended audiences within the AutoML marketplace, as discussed in Section 4.3.8; these can determine how much technical control can be safely granted to a stakeholder.

Turning to the HPO methods employed (CC11), a couple of commercial systems involve users pipelining ML processes by connecting operator ‘nodes’, e.g. RapidMiner and KNIME. It is difficult to determine whether the manual installation of associated HPO nodes aligns with the spirit of full AutoML, so they have been represented here with values of 0.5. Regardless, wherever details can be extracted, most AutoML products supply the basic algorithms of grid search (GR) and random search (RA). In fact, beyond these methods (GR+), heavyweight Microsoft goes no further. However, the remaining packages that can be commented about almost ubiquitously support Bayesian optimisation (BA).

That stated, the sparsity of multi-armed bandit techniques (MAB) – 3/16 identifiable commercial providers versus 6/19 open-source offerings – hints that the technical depth of the commercial market may be somewhat limited. This outcome may be due to a lag in how quickly theoretical advances disseminate into the commercial sphere. Alternatively, if this is a genuine indicator of the marketplace, vendors may simply find no need to implement more sophisticated methods if strong operational performance is available elsewhere, e.g. via experiment tracking, a decent UI, convenience features, and so on. The H2O package, for instance, is one of the earlier entrants into the

Table 91. Coverage of model-selection processes for commercial comprehensive AutoML systems (CC10/CC11). Evaluates capabilities for custom evaluation metrics (CC10). Also evaluates the existence of HPO techniques (CC11), i.e. grid search (GR), random search (RA), Bayesian optimisation (BA), multi-armed bandit strategies (MAB), genetic/evolutionary algorithms (GE), and meta-learning (M). For convenience, additionally classifies whether HPO mechanisms beyond grid/random search (GR+) are provided. Scores: 0 for absent, 1 for present, and U for unclear. The 0.5 denotes the HPO technique can be manually included as an operator ‘node’.

Name	CC10	GR	RA	BA	MAB	GE	M	SUM (out of 7)	GR+
D2iQ	1	1	1	1	1	0	0	5	1
Dataiku	1	1	1	1	0	0	0	4	1
H2O	1	1	1	0	0	1	0	4	1
SAS	U	1	1	1	0	1	0	4	1
Spell	1	1	1	1	0	0	0	4	1
B2Metric	0	1	1	1	0	0	0	3	1
Databricks	1	0	0	1	0	1	0	3	1
KNIME	1	0.5	0.5	0.5	0.5	0	0	3	1
SageMaker	1	0	1	1	0	0	0	3	1
RapidMiner	U	0.5	0.5	0.5	0.5	0.5	0	2.5	1
BigML	1	0	0	1	0	0	0	2	1
cnvrg.io	1	1	0	0	0	0	0	2	0
Microsoft	0	1	1	0	0	0	0	2	0
Auger	0	0	0	1	0	0	0	1	1
Einblick	0	0	0	1	0	0	0	1	1
Google	1	U	U	U	U	U	0	1	U
Number Theory	1	U	U	U	U	U	0	1	U
Alteryx	0	U	U	U	U	U	0	0	U
Big Squid	0	U	U	U	U	U	0	0	U
Compellon	U	U	U	U	U	U	0	0	U
DataRobot	0	U	U	U	U	U	0	0	U
Deep Cognition	0	U	U	U	U	U	0	0	U
IBM	0	0	0	0	0	0	0	0*	0
MyDataModels	0	U	U	U	U	U	0	0	U
TIMi	0	U	U	U	U	U	0	0	U
Sum	11/25	9/25	9/25	11/25	2/25	3.5/25	0/25		13/25

Table 92. Distribution of scores for commercial comprehensive AutoML systems (CC10). Evaluates capabilities for custom evaluation metrics. Scores: 0 for absent, 1 for present, and U for unclear.

CC10	
Score	# Vendors
0	11
1	11
U	3

AutoML sphere that eschewed popular Bayesian techniques for more straightforward grid/random search methods; evidently, the implementation was sufficiently performant to maintain its existence.

Notably, it has since become one of the few commercial adopters of genetic/evolutionary algorithms (GE). Finally, other than Auger perhaps hinting at meta-learning, there is no evidence that any vendor explores this functionality, so its mention is excluded from Table 91. Additionally, as a side note, while IBM scores zero on the standard forms of HPO, it appears unique in employing derivative-free optimisation via the RBFOpt package.

Moving on to the last sub-criteria, this review finds that closed-source opacity prevents any clear overview of which popular ML libraries are interfaced with (CC12). Proprietary algorithms may have been developed in-house, which further limits their exposure. We do note though that Google uses TF, D2iQ employs TF and Torch, and Dataiku interfaces with Sklearn, XGBoost, and H2O. As for the sub-criterion on currency (CC13), we re-emphasise that all the surveyed tools in this analysis were found to be active at the time of survey.

4.3.4 Explainability. As has been reiterated many times, evaluating the detailed mechanics of a commercial AutoML system is challenging. For one thing, unlike open-source software, there are financial costs to fully exploring the routine processes of a tool; a product trial is unlikely to be sufficiently informative. Moreover, even with full access, the detailed code underlying these services remains obscured. Thus, this review, primarily a collation of published material about current AutoML software, cannot assess every sub-criterion. For instance, we could not adequately gauge whether vendors provide facilities to clarify data lineage (EX1) and modelling steps (EX2). Naturally, the value of such an assessment for judging performant ML does remain, and deeper investigations are recommended in the future.

Nonetheless, once considerations of explainability move from system mechanics to system inputs/outputs, evaluations are easier to make, and these are detailed in Table 93. Indeed, if existent, the interpretability features discussed here are often well integrated into a UI and strongly promoted. Accordingly, Table 94 reveals that it is almost standard among vendors to explain global aspects of a model (EX3) automatically. Most of the time, this is identifying the importance of each input dataset feature to the potentially complex input-output mapping that a user receives. In contrast, Table 95 shows that local interpretability at the level of individual predictions (EX4) is mostly overlooked, though its enhancement is often well automated when addressed. Given that individual predictions/prescriptions are often associated with unique entities, e.g. choosing whether a loan is approved, businesses typically cannot afford to make complacent decisions even at this granular level. Thus, a lack of improvement in this area could stifle industrial uptake.

The distribution of scores only continues to worsen for each subsequent sub-criterion. Table 96 indicates that less than a third of surveyed vendors identifiably support scenario capabilities (EX5), which would allow a user to simulate the impact of changing variables on model outputs at a global/local level. Granted, similar to sensitivity analysis in other fields, this functionality leans more towards leveraging ML for exploration and insights rather than model production. Its provision may well depend on how a vendor markets its AutoML product. For a churn-related example of how scenario analysis may function, BigSquid claims its simulation tool allows stakeholders to see (1) the probability of a person churning, (2) the features that locally drive this churn, and (3) how this probability changes when certain variables are tuned. Compellon has a similar offering, but it goes a step further in allowing a user to specify a desired reduction in churn, which the tool responds to by highlighting both impactful features to target and the expected drop in churn associated with their variation. Such a comparison suggests a distinction between manual exploration and automatic optimisation, but judging the effectiveness of these scenario-based offerings is out of scope for this review. Elsewhere, while not included within the 25 comprehensive systems here, Ople.AI notably promotes a scenario tool with a unique ‘ROI’ tab that converts model metrics into financial values, aiming to translate technical outputs into a business understanding efficiently.

Table 93. Scores for commercial comprehensive AutoML systems (EX3–EX7). Evaluates capabilities for enhancing global interpretability (EX3), enhancing local interpretability (EX4), scenario analysis (EX5), bias/fairness assessment (EX6), and bias/fairness management (EX7). See Table 8 for rubric.

Name	EX3	EX4	EX5	EX6	EX7	Sum (out of 9)
Dataiku	2	2	1	2	0	7
IBM	2	0	1	2	2	7
SageMaker	2	2	0	2	1	7
DataRobot	2	2	0	2	0	6
Google	2	2	1	1	0	6
Big Squid	2	2	1	0	0	5
H2O	2	2	0	1	0	5
Microsoft	2	2	0	1	0	5
RapidMiner	2	0	1	1	0	4
SAS	2	2	0	0	0	4
MyDataModels	2	0	1	0	0	3
BigML	2	0	0	0	0	2
Compellon	1	0	1	0	0	2
Databricks	1	1	0	0	0	2
Deep Cognition	2	0	0	0	0	2
Einblick	1	1	0	0	0	2
Number Theory	2	0	0	0	0	2
TIMi	2	0	0	0	0	2
cnvrg.io	1	0	0	0	0	1
D2iQ	1	0	0	0	0	1
Spell	1	0	0	0	0	1
Alteryx	0	0	0	0	0	0
Auger	U	U	U	0	0	0
B2Metric	U	U	U	0	0	0
KNIME	0	0	0	0	0	0
Sum	36/50	18/50	7/25	12/50	3/50	

Table 94. Distribution of scores for commercial comprehensive AutoML systems (EX3). Evaluates capabilities for enhancing global interpretability. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

EX3	
Score	# Vendors
0	2
1	6
2	15
U	2

Regarding bias and fairness issues, their identification (EX6) and management (EX7) have not yet convincingly found their way into AutoML implementations within the commercial sphere. This state of affairs starkly contrasts the volume of discussion held within the media, business community, and academic literature. Indeed, only a third of surveyed vendors, half of these via

Table 95. Distribution of scores for commercial comprehensive AutoML systems (EX4). Evaluates capabilities for enhancing local interpretability. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

EX4	
Score	# Vendors
0	13
1	2
2	8
U	2

Table 96. Distribution of scores for commercial comprehensive AutoML systems (EX5). Evaluates capabilities for scenario analysis. Scores: 0 for absent, 1 for present, and U for unclear.

EX5	
Score	# Vendors
0	16
1	7
U	2

Table 97. Distribution of scores for commercial comprehensive AutoML systems (EX6). Evaluates capabilities for bias/fairness assessment. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

EX6	
Score	# Vendors
0	17
1	4
2	4
U	0

Table 98. Distribution of scores for commercial comprehensive AutoML systems (EX7). Evaluates capabilities for bias/fairness management. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

EX7	
Score	# Vendors
0	23
1	1
2	1
U	0

convenience features, are shown by Table 97 to surface associated metrics. Mitigation and remediation, either assisted or automated, are shown by Table 98 to be virtually nonexistent. This absence of functionality provides an ongoing risk for current AutoML providers, especially as regulations and policies for compliance continue to build, all aiming to safeguard trust in AI. However, even if the liability for unfair outcomes rests solely with the client, this feature gap is not ideal. For most

options, stakeholders presently need to undertake their own bias/fairness analysis, yet understanding how and why relevant metrics are calculated is an advanced skill in its own right [342]. Only IBM thus far seems to assess and handle such issues automatically, and the processes involved are not well detailed. That all stated, some lag is always to be expected between the time a concern rises into mainstream consciousness and the time it is pragmatically addressed. There is no current indication that the AutoML marketplace will not adapt to meet these evolving requirements held by industrial organisations and broader society.

Table 99. Scores for commercial comprehensive AutoML systems (EU1–EU5). Evaluates the availability of interactions via coding (EU1), CLI (EU2), and GUI (EU3). Also evaluates software client type (EU4) and level of documentation (EU5). See Table 9 for rubric.

Name	EU1	EU2	EU3	EU4	EU5	Sum (out of 7)
BigML	1	0	1	2	2	6
cnvrg.io	1	1	1	1	2	6
Dataiku	1	1	1	1	2	6
Google	1	1	1	1	2	6
H2O	1	0	1	2	2	6
KNIME	1	0	1	2	2	6
Microsoft	1	1	1	1	2	6
RapidMiner	1	0	1	2	2	6
SageMaker	1	1	1	1	2	6
Alteryx	0	0	1	2	2	5
Auger	1	1	1	1	1	5
D2iQ	1	1	0	1	2	5
DataRobot	1	0	1	1	2	5
Deep Cognition	1	0	1	1	2	5
IBM	1	0	1	1	2	5
SAS	1	1	1	1	1	5
Spell	1	1	0	1	2	5
Big Squid	0	0	1	1	2	4
Databricks	1	0	0	1	2	4
Einblick	0	0	1	1	2	4
MyDataModels	0	0	1	1	1	3
B2Metric	0	0	1	1	0	2
Compellon	0	0	1	1	0	2
Number Theory	0	0	1	1	0	2
TIMi	0	0	1	0	0	1
Sum	17/25	9/25	22/25	29/50	39/50	

4.3.5 Ease of Use. This survey does not assess the effectiveness of HCI options employed by comprehensive AutoML systems, leaving broader commentary to another review [313]. However, it can highlight how users are expected to interface with these tools, which correlates with intended audiences and accessibility. Unsurprisingly, reflecting the diverging slants of free and commercial software towards technicians and lay users, respectively, the open-source evaluations in Section 4.2.5 end up markedly different from those in Table 99. This assertion about divergence is, of course, a generalisation, as all AutoML implementations strive to serve as many stakeholders as possible,

Table 100. Distribution of scores for commercial comprehensive AutoML systems (EU1). Evaluates the availability of interactions via coding. Scores: 0 for absent, 1 for present, and U for unclear.

EU1	
Score	# Vendors
0	8
1	17
U	0

but the trends are nonetheless apparent. For instance, Table 100 shows that eight surveyed vendors do not even support text-based coding (EU1). This outcome contrasts with the overwhelming expectation that users, presumably data scientists, are meant to script with the functions provided in open-source software. After all, few free tools even provide alternative forms of HCI.

Table 101. Distribution of scores for commercial comprehensive AutoML systems (EU2). Evaluates the availability of interactions via CLI. Scores: 0 for absent, 1 for present, and U for unclear.

EU2	
Score	# Vendors
0	16
1	9
U	0

Table 102. Distribution of scores for commercial comprehensive AutoML systems (EU3). Evaluates the availability of interactions via GUI. Scores: 0 for absent, 1 for present, and U for unclear.

EU3	
Score	# Vendors
0	3
1	22
U	0

Now, while some vendors provide a CLI (EU2), as indicated by Table 101, such a form of UI does not seem fashionable for either open-source or closed-source software. Instead, with three exceptions, Table 102 confirms that GUIs (EU3) are the go-to within the commercial sphere of AutoML. Of course, designing an effective graphical interface takes extra effort beyond what is needed for the core AutoML engine, so developing such a UI is more likely to be spurred on when financial incentives are involved. However, the resulting benefits are clear: convenient controls, aesthetic appeal, and broader accessibility. The downside is that GUIs form a layer of abstraction, constraining user freedom to whatever functions have been encapsulated by graphical widgets. Broad accessibility and deep access are rarely both supported by the same UI, meaning that data scientists, in the absence of low-level coding, may struggle to fully/flexibly leverage the technical power a closed-source framework could provide.

Another indicator that commercial AutoML and open-source AutoML diverge in their intended audience is the client type of its software (EU4). Specifically, while all open-source comprehensive systems tend to be hosted on a desktop, potentially requiring technical know-how around their installation, Table 103 shows vendors prefer browser-based access. This trend aligns with the

Table 103. Distribution of scores for commercial comprehensive AutoML systems (EU4). Evaluates software client type. Scores: 0 for desktop only, 1 for browser only, 2 for desktop or browser, and U for unclear.

EU4	
Score	# Vendors
0	1
1	19
2	5
U	0

Table 104. Distribution of scores for commercial comprehensive AutoML systems (EU5). Evaluates level of documentation. Scores: 0 for none, 1 for partial, 2 for extensive, and U for unclear.

EU5	
Score	# Vendors
0	4
1	3
2	18
U	0

software-as-a-service (SaaS) movement that has recently become popular. In fact, commercial packages that support both browsers and desktops are often older entrants to the market, e.g. KNIME, RapidMiner, and Alteryx. As for the level of documentation (EU5) available, Table 104 suggests that vendors are more diligent on the whole than open-source developers, potentially needing to robustly support users with minimal ML expertise. That stated, there are a few services with low scores here, although the associated vendors may simply reserve shielded documentation for paid customers.

Whatever the case, ease of use is rarely considered when evaluating AutoML technology, at least compared with technical performance metrics. This criterion is currently associated with few research papers, and each only examines a small sample size of stakeholders [198, 550]. Accordingly, this review welcomes future systematic treatments of this topic, as they may shed further light on how much usability influences AutoML adoption by industrial organisations.

4.3.6 Deployment and Management Effort. Open-source comprehensive AutoML systems appear to focus heavily on developing the best ML models possible. Although they are still intended for general use, including by enterprises, such software is often closely linked to academic endeavours and thus prioritises exploiting technical innovations in ML. Commercial software, on the other hand, appears to care less about enhancing an ML solution and more about how to make it pragmatically useful to the paying client. The degree of this differentiation is debatable, but the concrete outcome is clear: commercial comprehensive AutoML systems distinguish themselves from free tools by investing heavily in MLOps capabilities.

However, as with all the preceding stages of the MLWF displayed in Fig. 2, the deployment phase and subsequent monitoring and maintenance are clusterings of various possible tasks. For instance, one process of interest in the academic literature is model compression (DM1), where a predictor with a complicated structure is simplified into a much smaller ML model that faithfully approximates the original function. This procedure ideally trades a negligible amount of solution validity for massive savings in computational space and inference time. Nonetheless, despite occasional associated research [211], we could not find this feature promoted by any AutoML

implementation. Seemingly, general concerns about model size remain negligible. For one thing, businesses may still be leveraging enough value from classic ML models to leave compression techniques unwarranted. The cheapening costs of cloud usage may also be limiting the pressure of computational constraints on all but the highest-volume vendors. That all said, as businesses begin to leverage more complicated ML solutions for their possible added value, perhaps even with AutoDL maturing as a technology, model compression may become particularly appealing.

Table 105. Scores for commercial comprehensive AutoML systems (DM2–DM4). Evaluates capabilities for deployment in specific environments (DM2). Also evaluates the existence of advanced mechanisms for deployment tests (DM3) and deployment updates (DM4). See Table 10 for rubric.

Name	DM2	DM3	DM4	Sum (out of 4)
cnvrg.io	2	1	1	4
Dataiku	2	1	0	3
DataRobot	2	1	0	3
H2O	2	1	0	3
Number Theory	2	1	0	3
RapidMiner	2	1	0	3
SAS	2	1	0	3
Alteryx	2	U	0	2
Big Squid	2	0	0	2
BigML	2	0	0	2
D2iQ	2	0	0	2
Deep Cognition	2	0	0	2
IBM	2	0	0	2
KNIME	2	0	0	2
Microsoft	2	0	0	2
Spell	2	0	0	2
B2Metric	U	1	0	1
Databricks	1	U	0	1
Google	1	0	0	1
SageMaker	1	0	0	1
Auger	U	0	0	0
Compellon	0	0	0	0
Einblick	0	0	0	0
MyDataModels	0	0	0	0
TIMi	0	0	0	0
Sum	35/50	8/25	1/25	

For now, some standard questions can be asked of commercial AutoML systems, especially around deployment. They and their evaluated answers are detailed in Table 105. First, a commercial service must be assessed for whether it even supports the productionisation of an ML solution. If yes, it is essential to know where that product is deployed (DM2). It turns out, as Table 106 shows, that the vast majority of vendors are flexible with their production environments; only four surveyed packages identifiably ignore this element of an MLWF. However, stakeholders should still be mindful that a few commercial heavyweights are restricted solely to the cloud, e.g. Google

Table 106. Distribution of scores for commercial comprehensive AutoML systems (DM2). Evaluates capabilities for deployment in specific environments. Scores: 0 for none, 1 for cloud only, 2 for cloud or on-premise, and U for unclear.

DM2	
Score	# Vendors
0	4
1	3
2	16
U	2

and SageMaker. This constraint may be too limiting for organisations that, for security reasons, desire on-premise deployments for their ML applications.

Table 107. Distribution of scores for commercial comprehensive AutoML systems (DM3). Evaluates the existence of advanced mechanisms for deployment tests. Scores: 0 for absent, 1 for present, and U for unclear.

DM3	
Score	# Vendors
0	15
1	8
U	2

Table 108. Distribution of scores for commercial comprehensive AutoML systems (DM4). Evaluates the existence of advanced mechanisms for deployment updates. Scores: 0 for absent, 1 for present, and U for unclear.

DM4	
Score	# Vendors
0	24
1	1
U	0

Next up, once automated deployment support is identified, the most pertinent question is how this deployment works in practice. For a real-world ML application, an organisation may iterate through numerous ML solutions, so Table 107 summarises whether commercial systems have advanced methods of testing deployments (DM3), particularly in relation to models that are presently online. Such procedures include A/B testing and champion-challenger setups; they also tend to form the basis of monitoring approaches, which will be discussed shortly. The takeaway here is that, while vendors do overwhelmingly assist in deploying ML solutions, only a small number automatically check how a prospective deployment compares against a previous one. As for sophisticated means of shifting to that new deployment (DM4), which can be critical if end-users are actively employing predictions/prescriptions from an existing ML model, Table 108 notes that only one service explicitly provides. Specifically, cnvrg.io supports a canary deployment strategy, where a new ML solution can be rolled out incrementally to subsets of users. Evidently, while enthusiastic, the embrace of MLOps by the commercial AutoML marketplace is still very nascent.

Table 109. Scores for commercial comprehensive AutoML systems (DM5–DM11). Evaluates the setup procedure for solution monitoring (DM5). Also evaluates capabilities for monitoring hardware performance (DM6), model performance (DM7), data/concept drift (DM8), and bias/fairness metrics (DM9). Additionally evaluates the existence of reactive retraining (DM10) and proactive retraining (DM11). See Table 11 for rubric.

Name	DM5	DM6	DM7	DM8	DM9	DM10	DM11	Sum (out of 10)
DataRobot	2	1	1	1	1	2	0	8
Microsoft	1	1	1	1	1	2	0	7
SageMaker	2	1	1	1	1	1	0	7
cnavrg.io	2	1	1	0	0	2	0	6
Dataiku	2	0	1	1	0	2	0	6
H2O	2	0	1	1	0	2	0	6
IBM	2	0	1	1	1	1	0	6
Auger	2	0	1	0	0	2	0	5
B2Metric	2	U	U	U	U	2	1	5
Number Theory	2	0	1	0	0	2	0	5
Alteryx	2	1	1	0	0	U	0	4
Google	1	1	1	1	0	0	0	4
KNIME	2	1	0	0	0	1	0	4
RapidMiner	1	0	1	1	0	1	0	4
SAS	1	0	1	0	0	2	0	4
D2iQ	2	1	0	0	0	0	0	3
Spell	1	1	1	0	0	0	0	3
Databricks	1	1	0	0	0	U	0	2
BigML	1	U	U	U	U	0	0	1
Big Squid	U	U	U	U	U	0	0	0
Compellon	0	0	0	0	0	0	0	0
Deep Cognition	0	0	0	0	0	0	0	0
Einblick	0	0	0	0	0	0	0	0
MyDataModels	0	0	0	0	0	0	0	0
TIMi	0	0	0	0	0	0	0	0
Sum	31/50	10/25	14/25	8/25	4/25	22/75	1/25	

Deployment capabilities, of course, are broadly applicable, i.e. equally relevant to one-and-done ML projects. Iterations can also be warranted simply for implementational reasons, e.g. fine-tuning end-user access to ML solution outputs. Thus, the notion of deployment alone does not capture the sense of continuous learning, which is a crucial prerequisite to AutonoML and next-generation frameworks [308]. This concept is not just a topic of intellectual curiosity either; many businesses during the COVID pandemic have wised up to the need for adaptation. Hence, the remaining sub-criteria here assess commercial comprehensive systems for the other important facets of MLOps: monitoring and maintenance. The evaluations are detailed in Table 109.

First, this review finds that a moderate number of vendors do support some form of monitoring. As summarised by Table 110, half of the commercial tools automate its setup (DM5), with many others supplying convenience features. Curiously, although the field of MLOps tends to care about how best to manage infrastructure around an ML model object, Table 111 reveals that less than half of the sampled systems use this monitoring to keep a close eye on hardware usage and performance (DM6). In contrast, it is somewhat more common, as shown by Table 112, to track

Table 110. Distribution of scores for commercial comprehensive AutoML systems (DM5). Evaluates the setup procedure for solution monitoring. Scores: 0 for none, 1 for convenience features, 2 for substantial automation, and U for unclear.

DM5	
Score	# Vendors
0	5
1	7
2	12
U	1

Table 111. Distribution of scores for commercial comprehensive AutoML systems (DM6). Evaluates capabilities for monitoring hardware performance. Scores: 0 for absent, 1 for present, and U for unclear.

DM6	
Score	# Vendors
0	12
1	10
U	3

Table 112. Distribution of scores for commercial comprehensive AutoML systems (DM7). Evaluates capabilities for monitoring model performance. Scores: 0 for absent, 1 for present, and U for unclear.

DM7	
Score	# Vendors
0	8
1	14
U	3

Table 113. Distribution of scores for commercial comprehensive AutoML systems (DM8). Evaluates capabilities for monitoring data/concept drift. Scores: 0 for absent, 1 for present, and U for unclear.

DM8	
Score	# Vendors
0	14
1	8
U	3

technical metrics for model performance (DM7). However, even if 14 out of 25 packages do watch for model deterioration, such a practice is relatively basic. Indeed, indicative that few commercial systems genuinely exploit more advanced theoretical innovations, Table 113 notes that less than a third identifiably bother to inspect the data itself for deleterious dynamics. Preemptively catching data/concept drift (DM8) is understandably more complicated than simply noticing that the validity of an ML model is decreasing over time, but it is also much more informative. As for monitoring bias/fairness metrics (DM9), Table 114 supports previous discussion in Section 4.3.4. Specifically, if so few AutoML tools even calculate these variables to begin with, it is no surprise that examining their time dependence is very rare.

Table 114. Distribution of scores for commercial comprehensive AutoML systems (DM9). Evaluates capabilities for monitoring bias/fairness metrics. Scores: 0 for absent, 1 for present, and U for unclear.

DM9	
Score	# Vendors
0	18
1	4
U	3

Table 115. Distribution of scores for commercial comprehensive AutoML systems (DM10). Evaluates the existence of reactive retraining. Scores: 0 for none, 1 for manual with convenience features, 2 for automatic with user-defined triggers, 3 for automatic with developer-defined triggers, and U for unclear.

DM10	
Score	# Vendors
0	10
1	4
2	9
3	0
U	2

Table 116. Distribution of scores for commercial comprehensive AutoML systems (DM11). Evaluates the existence of proactive retraining. Scores: 0 for absent, 1 for present, and U for unclear.

DM11	
Score	# Vendors
0	24
1	1
U	0

Finally, maintenance is the follow-up to monitoring, a process during which AutoML systems ideally attend to the accumulated flaws they have identified. In theory, any framework that employs mechanisms to compare/update deployed ML solutions is already geared to be adaptive. Thus, it is no surprise that Table 115 finds 13 surveyed commercial services identifiably support stakeholders with reactive model retraining (DM10). Sometimes the user is only provided with a convenient pathway to enact this procedure, but most of these AutoML systems take care of updates automatically based on user-specified triggers. As for the proactive form of retraining (DM11), where an ML solution is constantly being improved after deployment rather than simply being preserved against detrimental data dynamics, it is shown by Table 116 to be much rarer. Admittedly, there are AutoML tools not included within this analysis that claim to support such capabilities. For instance, TAZI [44] states it has a mechanism, i.e. ‘TAZI Hunt’, that is continuously looking for better models. Elsewhere, Pecan notes on its website that “our platform continuously monitors and optimizes your models”. This review cannot verify the effectiveness of either due to a lack of publicly available detail.

As a concluding comment, while reactive retraining is reasonably common in the commercial AutoML sphere, the technical sophistication of these mechanisms should not be presumed. Indeed, not one surveyed comprehensive system counted in Table 115 takes full responsibility for retraining,

i.e. via developer-provided triggers, although, granted, this may again be a consequence of vendors limiting their liability for poor outcomes. Furthermore, the standard champion-challenger setup may be interpreted as comparing various ML ‘experts’, but it does not ensemble them in any meaningful way to preserve local information. Indeed, model retraining tends to be all-or-nothing, facing risks such as catastrophic interference/forgetting. There are definitely more potent algorithms and frameworks to be found in the academic literature, and these may be required for next-generation AutonoML systems [102, 287, 308, 313, 472, 538]. However, it is also an open question regarding when adaptation should be triggered, and cost-benefit analyses are required [583]. Server prices in cloud infrastructures are usage-based, so model updates have a direct and quantifiable monetary cost; proactive retraining may not be worthwhile. Ultimately, the best practices for persistent AutoML and adaptation to data dynamics will continue to be hashed out on both the academic and technological fronts.

Table 117. Scores for commercial comprehensive AutoML systems (G1–G3). Evaluates capabilities for auditing (G1). Also evaluates the existence of access controls for model/data artefacts (G2) and ML application processes (G3). See Table 12 for rubric.

Name	G1	G2	G3	Sum (out of 3)
BigML	1	1	1	3
Databricks	1	1	1	3
Dataiku	1	1	1	3
DataRobot	1	1	1	3
Google	1	1	1	3
IBM	1	1	1	3
KNIME	1	1	1	3
Microsoft	1	1	1	3
RapidMiner	1	1	1	3
SageMaker	1	1	1	3
SAS	1	1	1	3
Alteryx	1	1	0	2
B2Metric	1	1	U	2
D2iQ	0	1	1	2
Einblick	0	1	1	2
Spell	0	1	1	2
cnvrg.io	0	1	0	1
Auger	U	U	U	0
Big Squid	0	0	0	0
Compellon	U	U	U	0
Deep Cognition	0	0	0	0
H2O	0	0	0	0
MyDataModels	U	U	U	0
Number Theory	U	U	U	0
TIMi	0	0	0	0
Sum	13/25	17/25	14/25	

4.3.7 Governance and Security. Enterprise application of ML exists within a different environment of concerns to that of a hobbyist seeking low-stakes ML models for personal use. Businesses caring

Table 118. Distribution of scores for commercial comprehensive AutoML systems (G1). Evaluates capabilities for auditing. Scores: 0 for absent, 1 for present, and U for unclear.

G1	
Score	# Vendors
0	8
1	13
U	4

about regulatory compliance and privacy matters may therefore be reluctant to employ free AutoML tools, which do not seem to support such safeguards. Accordingly, as for the case of MLOps, Table 117 shows that this criterion is a significant differentiator between the commercial and open-source markets. There is, however, a range of aggregate scores among the vendors. Large corporations are already very familiar with the operational requirements of businesses, so the AutoML services of certain heavyweights score well in this area, e.g. Google, Microsoft, and SageMaker. Furthermore, much infrastructure supporting performant ML with governance and security does not need to be overly specialised, so vendors with existing mature frameworks in place for other applications can benefit here. For instance, when considering auditability (G1), corporations that supply general cloud access already tend to have suitable governance mechanisms to leverage. In combination with other packages that specifically emphasise suitability for regulatory requirements, such as BigML, it turns out that over half of the surveyed commercial systems identifiably support auditability. This result is shown in Table 118.

Table 119. Distribution of scores for commercial comprehensive AutoML systems (G2). Evaluates the existence of access controls for model/data artefacts. Scores: 0 for absent, 1 for present, and U for unclear.

G2	
Score	# Vendors
0	4
1	17
U	4

Table 120. Distribution of scores for commercial comprehensive AutoML systems (G3). Evaluates the existence of access controls for ML application processes. Scores: 0 for absent, 1 for present, and U for unclear.

G3	
Score	# Vendors
0	6
1	14
U	5

As for security involving the data and model (G2), Table 119 highlights that at least 17 vendors have installed some artefact access controls in their AutoML services. Furthermore, if there are protection mechanisms for objects, then functions are also typically safeguarded (G3), e.g. to prevent unauthorised users from overwriting existing ML deployments. Only a few exceptions make up the differing distribution of scores in Table 120, e.g. Alteryx and cnvrg.io. Of course, as policies around responsible AI continue to evolve internationally, this criterion may warrant expansion into other

detailed questions in the future. However, for now, these evaluations should help emphasise that there are many facets to the smooth operation of a performant ML application; not all of them are technical.

4.3.8 Discussion. In hindsight [308] and with foresight [313], many aspects regarding the progression of AutoML may appear expected. It can thus be tempting to ascribe a singular pathway to the field along which developments tread, especially given an academic community that is largely collaborative. This sense of a convergent roadmap is further prompted by the modern realities of scientific funding and popular fervour; it is usually not a good academic career move to be *too* radical. However, when it comes to the competitive commercial sphere, this review finds that, by and large, all bets are off. The only guaranteed driving force shared across all AutoML vendors is money, which requires engagement from stakeholder clients. Accordingly, in this unusual swarm optimisation of performant ML, it is contrastingly desirable for vendor entities to diverge and differentiate, seeking their own profitable niches. They, of course, remain grouped by the survey requirement that they supply a ‘comprehensive’ AutoML system. Nonetheless, such an environment supports the fragmentation of agendas, each with its own power to influence the ongoing evolution of AutoML technology. Therefore, we highlight specific trends and clusters that have surfaced as part of this survey.

A licence to explore. Relevant systems:

- Einblick
- Compellon
- MyDataModels
- TIMi

Some AutoML systems are developed and promoted to dive deep into data. They tend to score lower for deployment-related criteria while solidly supporting EDA and related auxiliary functions. So, for example, they will likely not productionise an ML model that automatically identifies users at risk of churn. Instead, an intended use case may be a marketing manager (1) uploading a spreadsheet of customers, (2) seeking to examine the driving features of churn, and (3) experimenting with various scenarios. Moreover, as the listed tools focus on harvesting and understanding insights from data, they tend to prioritise clear visualisations and explainability.

The primacy of MLOps. Relevant systems:

- Google
- SageMaker
- cnvrg.io
- D2iQ
- Databricks
- SPELL

On the other end of the MLWF to EDA, it is often apparent which AutoML services prioritise MLOps. These commercial systems score well on deployment and management metrics, but they often support other core functions of an ML application in a more rudimentary manner, if at all. Amidst this list, Google and SageMaker are particularly interesting. The associated corporations already had a business in providing cloud services and DevOps before touching AutoML, so their entrance into the marketplace may be a logical extension of supplying MLOps support to developers. Essentially, their services have grown from the opposite direction to teams that started automating model selection and now focus on productionisation. As a side note, heavyweight Microsoft has not been included here, given that the AutoML framework seems more suited to less-technical users, often providing a GUI rather than relying on users to code around convenience features. These

factors suggest a ‘from-scratch’ approach to AutoML instead of a gradual expansion originating in DevOps.

Some assembly required. Relevant systems:

- KNIME (open marketplace)
- RapidMiner (open marketplace)
- cnvrg.io (semi-open marketplace)
- Alteryx (closed marketplace)
- Deep Cognition (closed marketplace)
- Einblick (closed marketplace)
- Number Theory (closed marketplace)

Several commercial AutoML frameworks are designed under a ‘building block’ paradigm, where users typically fashion a network of nodes to represent and process an ML application. Depending on the implementation, these nodes may be ML pipeline components, i.e. specific data transformations, or higher-level ML operations. A mix of the two may even be possible. Whatever the case, it has occasionally been challenging to assess associated systems for their automated capabilities. Indeed, the flexibility of free assembly has a drawback that, without appropriate defaults/constraints, users need to manually find and vet these building blocks before integrating them into modelling work. Additionally, under this paradigm, the supply of possible nodes can vary. Closed marketplaces, noted in the list of relevant systems, tend to have restricted offerings with assurances of functionality by the developer. In contrast, the open marketplaces of KNIME and RapidMiner, where community members submit their own building blocks, can stimulate high scores for completeness and currency. This boost, however, comes with the cost of further complicating trust in ML. Both approaches have pros and cons.

The legacy experience. Relevant systems:

- BigML
- MyDataModels
- TIMi

The topics of design and aesthetics are somewhat subjective; individual user responses to various developer decisions cannot easily be predicted. However, some trends drive the software industry, and many vendors have embraced the paradigm of SaaS. Present-day products are often easily accessible to general stakeholders via a browser, exhibiting UI artefacts generated by Bootstrap and other modern front-end web development frameworks. The tools listed here are thus notable for not existing within the same design sphere as many others. Nonetheless, it remains an open question whether a legacy user experience (UX) influences their uptake and, if so, by how much.

Degrees of transparency. Relevant systems:

- Dataiku (highly performant + highly transparent)
- DataRobot (highly performant + moderately transparent)
- H2O (moderately performant + highly transparent)
- IBM (moderately performant + moderately transparent)
- Auger (moderately performant + minimally transparent)
- B2Metric (moderately performant + minimally transparent)
- BigSquid (moderately performant + minimally transparent)

All commercial AutoML tools analysed within this section have been ‘comprehensive’, automating key model-development processes central to traditional AutoML while also, in many cases, supporting other tasks along an MLWF. Under this definition, the tools listed here are those within the analysis that have scored reasonably well across a breadth of criteria for performant ML. They

are not necessarily ‘better’ than their competitors, but, at the same time, it does not hurt to be so thorough. Indeed, some sub-criteria can be ‘make or break’ for potential clients, including whether on-premise deployment is supported, whether a system is auditable, whether business-appropriate data can be easily prepared, and so on.

That all said, transparency also matters when stakeholders consider their options. Such an assertion needs to be emphasised as, zero-scoring unknowns aside, the assessment framework for performant ML did not penalise commercial systems for every obscured detail. In fact, several tools at the bottom of the list above were often given the benefit of the doubt regarding promoted claims. Admittedly, this lack of transparency is common among smaller organisations. A cynical perspective may be that a commercial fog allows gaps in services to be veiled. More charitably, one could argue that smaller vendors must be especially defensive of their competitive advantages within a business environment. Whatever the case, it is likely that stakeholders will increasingly lean towards more transparent offerings as the commercial sphere becomes progressively more populated. Ultimately, the future will tell whether the AutoML marketplace will remain as diverse and fragmented as it currently is or whether an effective oligopoly driven by consensus consumer desires will emerge.

5 AUTOML APPLICATIONS

At this point, it is worth recalling that a core purpose of this review is to comment on the translation of AutoML theory/design into AutoML implementation/technology as of the early 2020s. Crucially, Section 4.3.8 acknowledged that both academia and the commercial sphere have varying levels of freedom to explore/generate new innovations in the field. On the academic side, public funding – we refer to fundamental-research grants – necessarily hews to some degree of conservatism, so that taxpayer money is not ‘wasted’. However, once endowed, these resources can often allow researchers to pursue threads of interest without immediate applicability. In contrast, a privately funded startup can be as radical as it wants, subject to the approval of its investors, but its survival depends on the ability to maintain a healthy income, i.e. those radical choices need to pay off. Thus, it is the consumers that primarily determine whether a provider stays in business and continues to influence the marketplace. The implication here is that, while innovation and experimentation will always influence AutoML products, the assertion in Section 3 continues to hold merit within industry settings and broader society: “demand creates supply”.

Unfortunately, the problem with evaluating a nascent technology is that it is often unclear what the demand profile for its use is, let alone how it will eventually look once it stabilises. For instance, recent years have witnessed debates around internet access becoming enshrined as a human right, whereas, a century ago, virtually no one could predict that this non-existent concept would become a need. So, to get a handle on the demand that drives AutoML technology, present and expected, this review has pursued a combination of several approaches. In Section 3.1, we reasoned at a granular level what target stakeholders want from an ML application to claim that it is ‘performant’. In Section 4, we surveyed extant AutoML software, following the logic that its supply is an oblique reflection of its demand.

Here, we examine the final piece of the puzzle: ML applications that have been processed via AutoML practices and systems. After all, stakeholder desires and the tools developed to satiate them are both symbols of hypothesis; they suggest AutoML will be helpful to industry without ever *proving* this to be true. Of course, such forecasting is still valuable in these early years of the technology, especially while the evidence of industrial use is still building up. Nonetheless, a survey of applications finds that there is already enough content to, in a preliminary fashion, directly analyse the ‘fulfilled’ demand for AutoML. Accordingly, with appropriate constraints, Section 5.1 investigates what academic literature reveals about the real-world use of this technology.

Then, acknowledging that some AutoML software has been preemptively designed for specific applications, Section 5.2 discusses the phenomenon of domain specialisation. Finally, although the associated narrative must be approached carefully, Section 5.3 examines commentary on applications according to the vendors of AutoML.

5.1 Academic Reports

The most reliable evidence of successfully completed AutoML applications is likely to be found within the academic literature, thanks to an enforced degree of peer review by independent experts. Naturally, the downside is that this perspective is also very limited, likely overlooking the more mundane usage of AutoML by commercial clients. Academic publications need to be of sufficient scientific importance. A further possible complication is that AutoML is meant to be an instrument for stakeholders to achieve other aims. Ideally, it should not be the promotional point of any research, making it harder to identify relevant applications. Nonetheless, this is not presently an obstacle. Given that much AutoML literature within the past decade is on theory and design, its practical use is relatively novel and expected to be highlighted. After all, the field is still looking to build an authoritative portfolio of archetypal examples to demonstrate rigorous proof of principle.

So, having acknowledged the nuances of an academic perspective, this section presents results from a literature survey. However, because this monograph is focussed on the broader use of AutoML technology, stringent constraints on the scope were necessary. First of all, the use of AutoML software by its own development team, without any significant industrial collaboration, did not make the cut any more than a self-citation would be considered a sign of broader research impact. Also, general benchmark papers, of which there are many [144, 358], were likewise excluded. For instance, testing experiments on UCI datasets [357] is not necessarily representative of real-world uptake. Publications deemed suitable for this survey need to target a specific industrial problem and associated dataset, proposing a solution obtained via AutoML. High-level reviews were thus similarly ignored, even though one such article is notable for considering adaptivity in real-world contexts [515]; this is both rare and increasingly recognised as important.

Naturally, such a tightly constrained survey of academic literature is not trivial to undertake. Citations of a publication that marks the release of an AutoML package cannot be blindly collated. For instance, the seminal Auto-WEKA [532] has been referenced numerous times to justify employing Bayesian optimisation [169, 559] without any concurrent use of the implementation. Similarly, perusing the literature with keywords related to ML and automation throws up many irrelevant results. As an example, in doing so, one may encounter a manuscript titled “automated staff assignment for building maintenance using natural language processing” [403]. This publication and others [263] employ ML to mechanise a manual operation; they do not automate the higher-level processes of ML. Of course, conversely to these false positives, such keywords can also miss valid applications that are not explicitly cognisant of the AutoML aspect but still use associated software. As previously mentioned, these examples are presently rare, but they do exist.

Another challenge with an application survey is deciding how to treat the history of AutoML. Particularly in earlier years, several proposed methods and advancements were not given names. Others that were named cannot presently be linked to an open-source repository or extant software product, e.g. SABLE [103], FLASH [573], Learn-O-Matic [492], Predict-ML [367], Net-Net [110], and HyperSPACE [248]. While this means exclusion from the earlier analyses of tools, many such AutoML approaches have a legitimate presence in an industry use case. They may not be employed again with ease, but they still prototypically demonstrate how organisations may leverage AutoML in pursuing real-world business objectives. So, in general, applications based on faded techniques and frameworks have been included in this survey. At the same time, to provide a current lens for this review, only a ten-year window for academic reports has been considered, ending on the 3rd

of August 2022. This restriction should not be seen as diminishing the importance of earlier works. For instance, a series of publications involving the chemical process industry is noteworthy for its prescient and sophisticated focus on automated adaptation [287–293], i.e. a crucial element of monitoring and maintenance within an MLWF.

Table 121. References to surveyed academic publications detailing AutoML applications. The references are grouped by major industries, subfields, and sub-subfields.

Industry	Subfield	Sub-Subfield	Ref.
Health & Biomedical (63)	Diagnosis (32)	Breast Cancer (6)	[383, 394, 456, 462, 504, 505]
		Mental Health (6)	[172, 173, 269, 338, 473, 543]
		Other (20)	[62, 99, 107, 119, 134, 194, 212, 275, 276, 280, 389, 425, 466, 468, 488, 507, 517, 523, 567, 573]
	Condition Management (19)		[108, 131, 150, 171, 247, 266, 271, 319, 337, 352, 354, 370, 372, 434, 463, 493, 503, 533, 552]
	Genetics & Biomarker Research (8)		[115, 164, 339, 374, 376, 407, 423, 539]
	Processing Medical Data (4)		[248, 367, 369, 422]
	Transport Demand (6)		[78, 79, 136, 207, 440, 537]
	Transport Infrastructure Management (5)		[260, 318, 341, 453, 455]
	Optimising Logistics (2)		[186, 264]
	Road Accidents (2)		[77, 498]
Chemical & Material Science (15)	Chemical Process (7)		[103–105, 479, 480, 482, 581]
	Electrical and Photovoltaic Compounds (3)		[366, 415, 431]
	Other - Material Science (3)		[192, 392, 426]
	Other - Chemistry (2)		[243, 548]
Information Technology (13)	Cybersecurity (10)		[149, 251, 278, 334–336, 447, 476, 500, 542]
	Software Development (2)		[226, 404]
	System Robustness (1)		[576]
	Water (5)		[312, 408, 524, 547, 569]
Energy & Utilities (12)	Power Generation (4)	Solar (1)	[174]
		Hydrocarbons (1)	[564]
		Wind (1)	[446]
		Other (1)	[477]
		Power Grid Optimisation (2)	[373, 549]
	Waste Management (1)		[368]
	Agriculture (8)		[80, 296, 314, 320, 351, 418, 430, 570]
Education (7)			[202, 228, 286, 322, 417, 484, 540]
Meteorological (5)			[377, 391, 492, 521, 577]
Radiography & Physics (5)			[359, 384, 427, 442, 571]
Finance (5)			[52, 53, 270, 285, 381]
Manufacturing & Machinery (5)			[132, 233, 333, 382, 390]
Robotics (3)			[67, 138, 201]
Aviation (3)			[58, 163, 347]
Telecom (2)			[215, 572]
Ecology (2)			[110, 217]
Retail (2)			[143, 279]
Advertising (1)			[148]
Professional Services (1)			[508]
Sport (1)			[229]
Media (1)			[50]

In total, 169 papers have been included in this survey. They are cited within Table 121; discussion around the listed groupings is deferred to later. For each report, the following elements were extracted:

- The tools used
- The application domain
- The metrics used in the experimental work
- The justifications given for the use of AutoML technology
- The stages of an MLWF involved in the study
- The year of publication

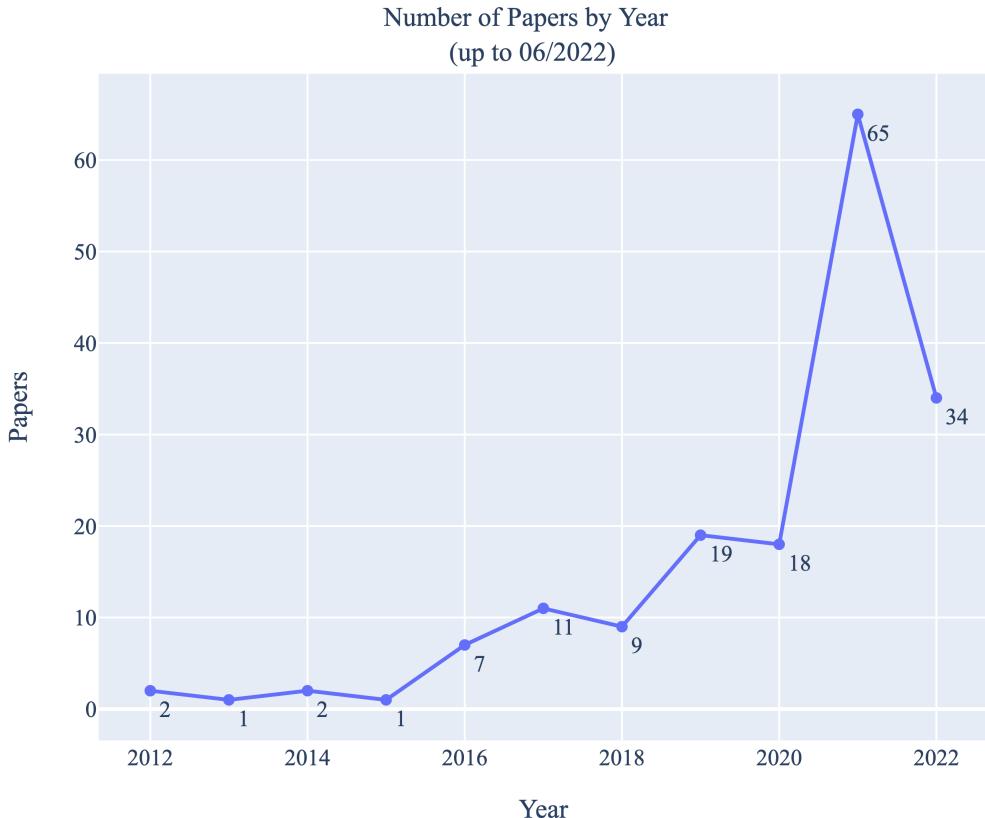


Fig. 3. The yearly numbers of academic publications reporting on AutoML applications.

To begin the analysis based on this curated information, Fig. 3 charts how many publications have reported on AutoML applications each year. Given the early August survey cut-off, the projection for 2022 should at least be on par with 2021. So, it is evident that the practical use of AutoML is surging, even in the eyes of academia. Granted, the overall volume of work in this space is increasing, but this result also highlights that the AutoML field is no longer purely in the fundamental research stage. Efforts to leverage its technological benefits within various domains are accumulating.

Now, while the surveyed applications have employed numerous AutoML frameworks, only a few have found repeat use. These are listed in Fig. 4. Accordingly, the chart of yearly reporting can

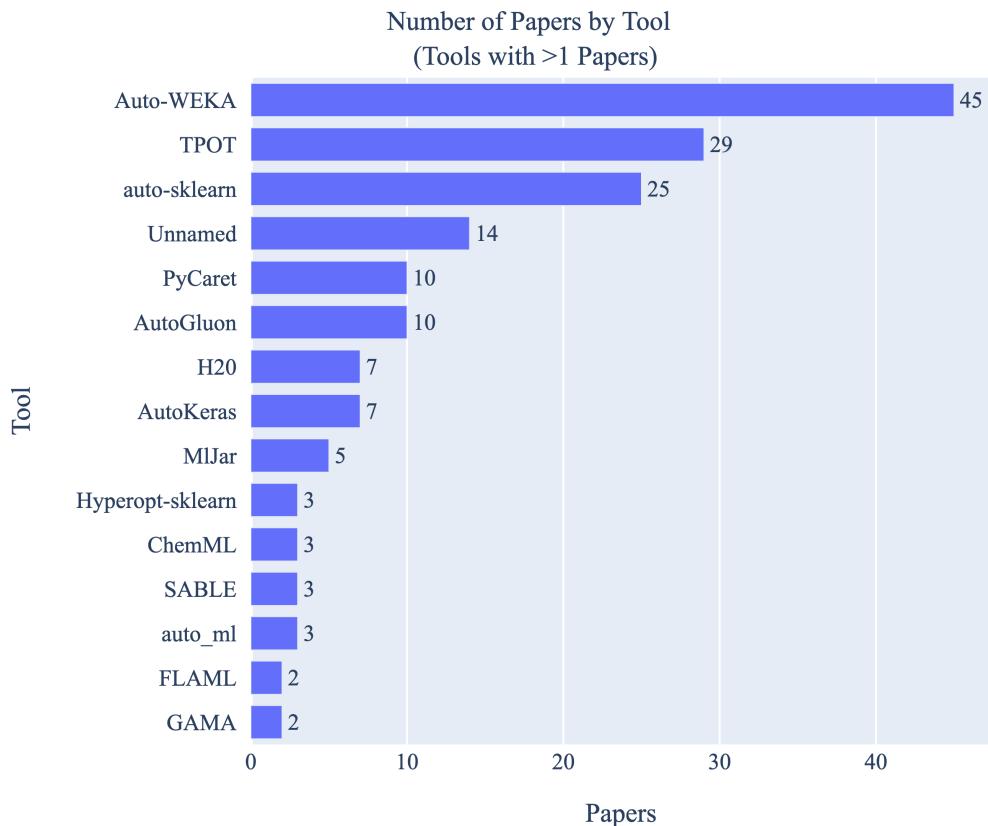


Fig. 4. The AutoML tools associated with at least two application papers.

be broken down further to investigate the most popular tools, at least by publication volume. It can immediately be seen from Fig. 5 that there is a ‘first mover’ effect in play. Despite the increasing amount and diversity of AutoML tools, many of which are likely to be more technically advanced and generally performant, Auto-WEKA and TPOT are still dominant in usage as of 2021. Both open-source packages were very early entrants into the AutoML ‘marketplace’. Accordingly, there may be a couple of factors involved. Firstly, a system without competitors has the luxury of cultivating a solid reputation, assuming it meets a minimum threshold of functionality. By the time there are competitors, the snowball effect is complete; prospective users are drawn to industry standards. Secondly, whether or not a stakeholder buys into the reputation of an established system, it does serve as a consistent baseline for comparative performance, especially where ML applications might try a few AutoML frameworks.

Of course, there are many more reasons why stakeholders may have turned to some of the listed AutoML frameworks. For instance, the rise of TPOT is likely supported by its relatively extensive coverage of an MLWF, e.g. in terms of data preparation and feature generation/selection. Other AutoML systems are also gradually developing a solid presence as of 2021, including auto-sklearn, AutoGluon, and PyCaret. These tools benefit from strong communities, good documentation, and high levels of active development/maintenance.



Fig. 5. The yearly numbers of academic publications reporting on AutoML applications associated with popular AutoML tools.

Turning to broader trends, this review notes that many of the popular AutoML systems are considered generalist, e.g. Auto-WEKA, auto-sklearn, TPOT, and AutoGluon. Within the limits of their technical coverage, they are designed to be agnostic in terms of domain and dataset. However, most application contexts are highly specialised. Some examples are as follows:

- Using AutoML to predict medical outcomes for diabetes patients [319].
- Applying AutoML to the domain of traffic forecasting [78, 79].
- Testing the ability of AutoML to anticipate crash severity in Colombia [77].

Admittedly, the core ML processes remain the same for many of these specialised applications, e.g. model selection for supervised learning. Thus, with the technological evolution of AutoML, it is clear why there is a gathering proliferation of publications testing AutoML in different domains. The technical hurdles of ML have been lowered, providing a prime opportunity for enterprising stakeholders to exploit associated techniques in numerous fields.

Nonetheless, applying generalist AutoML packages to varying domains is not trivial. There is undoubtedly differentiation that needs to be accounted for, the majority involving the preparation of data and the translation of model outcomes. The former, in particular, is an intense time sink for a typical data scientist. Perhaps this will eventually be ameliorated by the continuing evolution of general AutoML frameworks [308], but, in the meantime, developers of some AutoML software have vaulted ahead by specialising in a particular domain. Arguably, this might indeed be an appropriate way forward. Fine-tuning AutoML technology for the realities and demands of a specific context could avoid any software bloat that generality requires. However, ancillary features aside, specialised tools are no more revolutionary than generalist AutoML. Further discussion on this topic is reserved for Section 5.2.

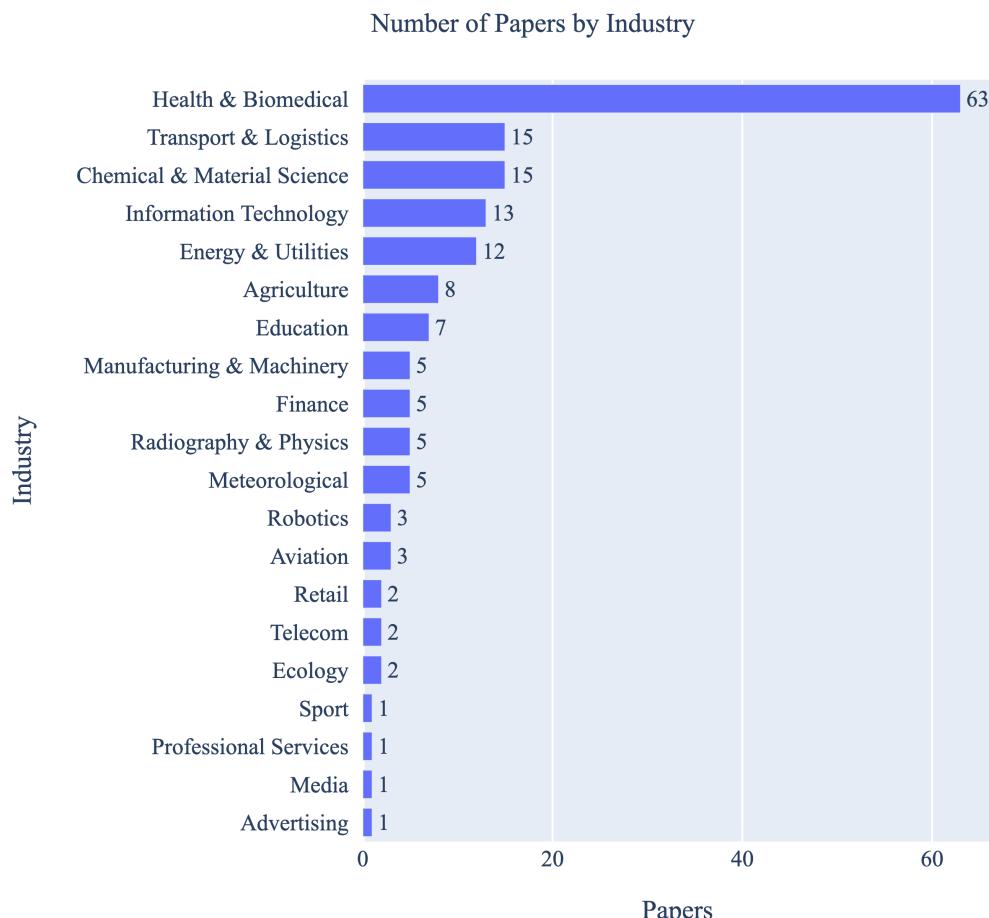


Fig. 6. The number of academic publications reporting on AutoML applications associated with each major industry.

Next, examining the major industries associated with AutoML applications renders the results in Fig. 6. They are essentially aggregations of the finer subfield breakdown, which was applied where possible, in Table 121. Overwhelmingly, if 63 out of 169 academic publications are to be seen as a reflection of broader use, AutoML appears to be predominantly employed in the Health & Biomedical sectors. The second place is shared between Transport & Logistics and Chemical & Material Science, with 15 reports each. Then comes Information Technology at 13 references, Energy & Utilities at 12 references, and many other domains in the long tail of the distribution.

Here, it is tricky to identify why AutoML use has been concentrated in the way it has. Chance likely plays some part, as there is little fundamental reason why the health sector was chosen as a backdrop for some of the earliest practical use of the technology. Of course, once proof of principle is established, other interested stakeholders can better see how to use AutoML in such a context. Perhaps the medical domain has simply had one of the most extended times since preliminary successes to cultivate a bandwagon effect. Nonetheless, in practice, some factors probably do

weight the probability of a domain receiving attention from AutoML entrepreneurs. For instance, there is the availability and quality of data, as well as the legal authorisation to use it and publish work around it. It is also arguable that several of the best-represented sectors in a survey of *academic* literature are themselves proximal to academia, i.e. Health & Biomedical, Chemical & Material Science, and Information Technology. The research, trials, investments and advancements in these fields are all better suited for scientific dissemination than ML problems in advertising or professional services, e.g. law, accounting, consulting, etc. Indeed, if a view of AutoML uptake is instead based on vendor reports, as discussed in Section 5.3, the top domains for applications are expected to be driven by commercial interests, e.g. Finance, Retail, Marketing, and Insurance. Simply put, perspective matters.

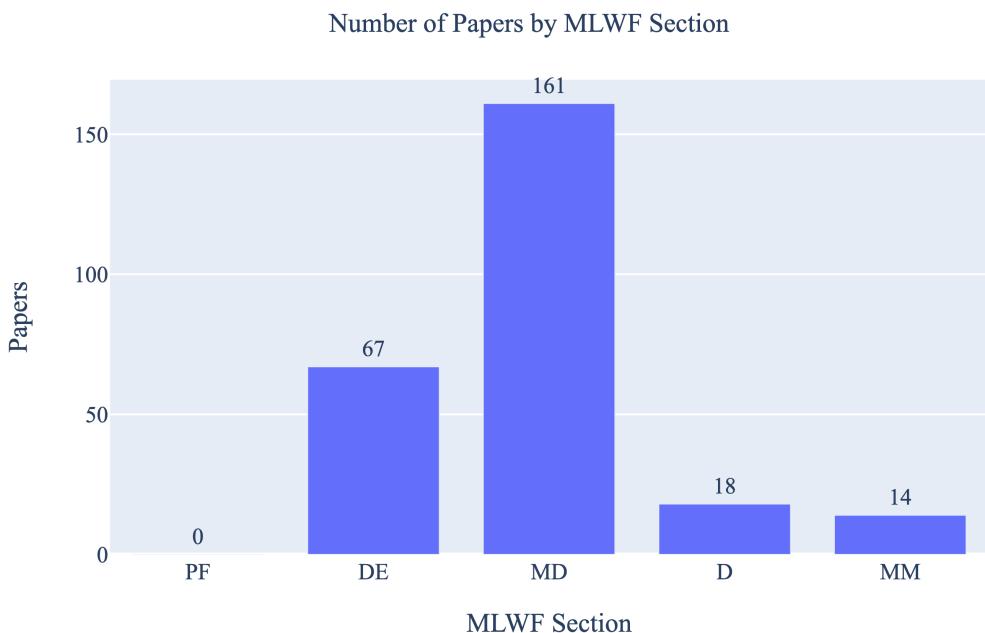


Fig. 7. The number of academic publications reporting on AutoML applications associated with each stage of an MLWF. Abbreviations: PF for problem formulation & context understanding, DE for data engineering, MD for model development, D for deployment, and MM for monitoring & maintenance.

Beyond domains, academic publications can also be categorised by the stages of an MLWF that are involved in their reported AutoML applications. Accordingly, acknowledging that one article can be associated with multiple phases, the resulting distribution is depicted in Fig. 7. Two expected details immediately stand out. Firstly, no academic report focusses on the preliminary stage of an MLWF; all applications begin with a well-defined and well-scoped problem. Secondly, almost all applications involve model development, primarily HPO, as this is widely seen as the core offering of AutoML.

More informatively, 67 papers, around 40% of the survey, discuss data engineering, i.e. pre-processing and feature generation/selection. Sometimes, this element appears as a direct by-product of an AutoML package, e.g. TPOT. In one particular case [127], the data-engineering discussion is even the central focus, as the associated application involves assessing AutoFE. Specifically, the

report examines the comparative utility of features generated by human experts, Featuretools, and tsfresh. In contrast, while data engineering receives a reasonable amount of attention, the last two stages of an MLWF are relatively ignored, with no more than about 10% coverage. This result may be because, as mentioned earlier, many applications reported in academia are seminal proof-of-principle works. They are typically one-and-done projects, often used to harvest contextual insights rather than establish long-term productionised models. That said, the automation of MLOps is also nascent, especially in terms of continuous learning, so time will tell whether the progression of these capabilities, as applied to real-world contexts, will receive commensurate academic interest.

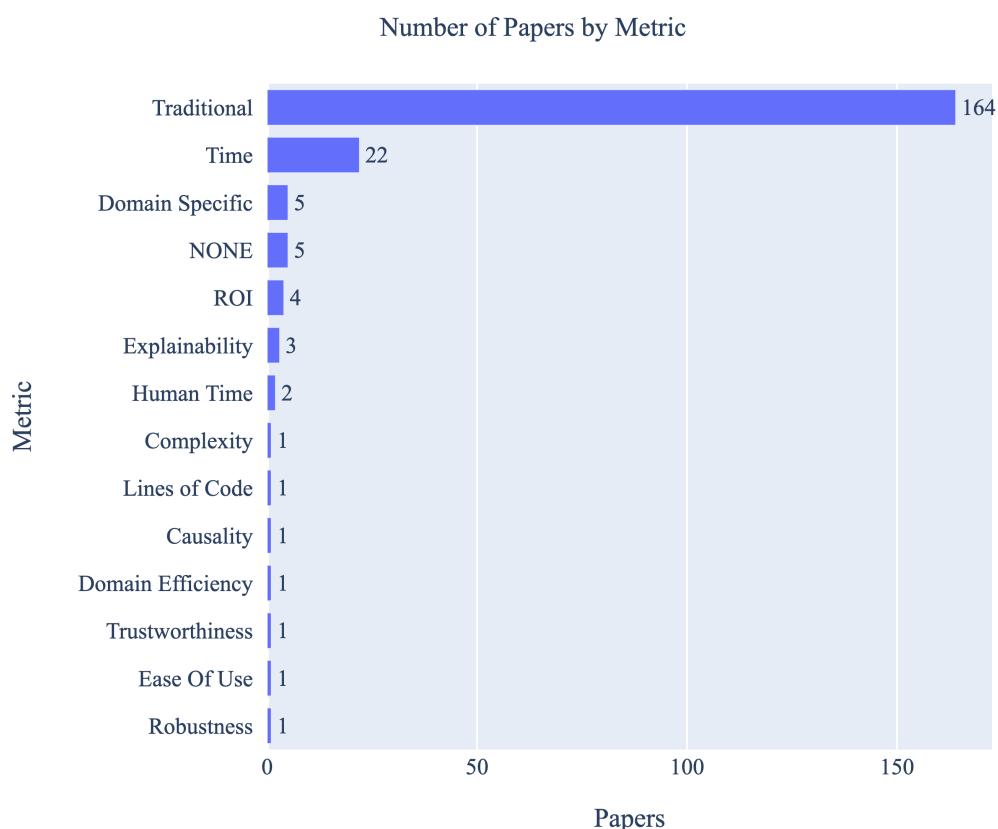


Fig. 8. The number of academic publications reporting on AutoML applications associated with evaluation metrics of interest.

At this point, it is worth recalling a central tenet to this monograph: performant ML, from the perspective of industry and broader society, relies on much more than model validity. So, as part of the current application survey, academic publications have been dissected for the metrics they care about. The resulting breakdown is shown in Fig. 8, where the ‘traditional’ term represents evaluations based on a confusion matrix, e.g. accuracy, sensitivity, specificity, AUC, the F_1 score, etc. Evidently, the academic lens is still firmly fixated on model error. In fact, it is remarkable that computation ‘time’, a metric for technical efficiency related to model training, is only considered by 22 papers, i.e. 13% of surveyed works. This result suggests deficiencies in scientific thoroughness

and rigour, even in the absence of loftier contemplations on the nature of ML performance. In fairness, a few works do focus more on data preparation or deployment; many reports that do not name specific evaluation metrics fall in this category. Nonetheless, the preoccupation with model validity in academia remains clear.

Now, even though only 23 reports, 13.6% of the surveyed academic publications, consider non-traditional metrics beyond computation time, it is still instructive to see where the attention falls. For instance, as a consequence of domain specialisation, several ML applications are judged by custom criteria related to context [143, 148, 164, 372, 500]. These projects include one that uses metrics associated with protein folding [164] and another that evaluates advertising relevance [148]. Then there are performance measures that lean towards ROI. These are not always concretely quantified, such as the vague ‘business benefits’ discussed by one publication [569]. Nonetheless, as exemplified by the expected benefit ratio (EBR) defined within another report [233], commercial metrics have a definite presence. The only oddity is that this presence is so small, despite the volume of papers dedicated to exploring AutoML in applied industrial settings. Again, perhaps discussions around business benefits are not considered of academic interest; any financial exploitation of uncovered contextual insights may occur subsequently to their reporting.

The remaining tail of the distribution in Fig. 8 gives tiny glimpses that suggest the criteria proposed in this monograph for performant ML are well founded. Granted, they do also highlight the challenge of quantification, especially where subjectivity is involved. For instance, the one application that explores ease of use does so by working with stakeholders to develop an associated score [217]. Regardless, finding some form of representation for these facets of performant ML is still worthwhile. Operational efficiency, for example, is captured by the notion of ‘human time’, i.e. how many hours it takes for a human to generate an ML solution with an AutoML tool. Two of the surveyed academic publications record and tabulate this metric [372, 425]. One of the applications even goes further by assessing efficiencies based on lines of code [425].

Table 122. Possible justifications for using AutoML for an ML application.

Reason	Explanation
Computational Cost	AutoML reduces the computational cost of running/optimising ML models.
Data Prep	AutoML can assist with data preparation, including preprocessing and feature engineering.
Deployment Considerations	AutoML can assist with deployment and/or monitoring & maintenance.
Domain (or Existing Process) Utility	AutoML might be useful in a new domain or in comparison with an existing process.
Expertise (Difficulty)	AutoML compensates for ML being difficult and requiring special expertise.
Explainability of AutoML	AutoML can assist with enhancing the explainability of ML work.
Guesswork (Human Error)	AutoML eliminates human error and guesswork involved in manual ML.
Model Simplicity	AutoML creates simple models.
N - Justifying ML	None. AutoML is used without justification, but the use of ML is justified.
N - No Justification	None. AutoML is used without justification.

Table 122. Possible justifications for using AutoML for an ML application.

Reason	Explanation
Open Source	A specific AutoML tool is open-source.
Package is SOTA (Popular)	A specific AutoML tool is the current state of the art (SOTA) or popular.
Reference Layperson (Democratisation)	AutoML can be used by a layperson, helping to democratise ML.
Reproducibility	AutoML can ensure ML work is reproducible.
Running Many Models	A specific AutoML tool can run many models at once.
Technical Power (Accuracy) of AutoML	AutoML provides superior technical power/performance, e.g. in terms of model validity.
Time (Effort)	AutoML significantly reduces the time/effort a human must invest in ML.

Admittedly, many of the exotic forms of evaluation metrics would perhaps be considered irrelevant if the primary motivation for using AutoML was enhanced predictive/prescriptive accuracy. To better gauge how true this is, a list of possible reasons for why one might employ AutoML was curated, detailed in Table 122. The 169 academic reports were then dissected for any mention of the listed justifications, with the resulting counts of associations charted in Fig. 9. It is immediately apparent that, according to academia, the technical proficiency of AutoML is neither first nor second in the mind of an average stakeholder when considering its use. Granted, the fact that it ranks third is still an interesting outcome, given that there is an ongoing debate about whether the outputs of AutoML, in terms of accuracy, significantly/consistently outperform ML models handcrafted by experts [249, 362, 582]. Nonetheless, nothing here supports the seeming obsession with model-validity metrics revealed by Fig. 8. One may thus contemplate the discrepancy: why are the benefits espoused about AutoML seemingly left unsubstantiated? An optimist may infer that the community already holds these truths to be self-evident. A cynic may question the sincerity of the assertions, perhaps seeing them exploited to elevate the importance of an otherwise rote ML publication. Ultimately, the pragmatic truth may lie somewhere in the middle. Of course, in fairness, every published AutoML application need not grapple with the challenge of quantifying the utility of the technology. However, this monograph hopes to encourage a greater degree of concrete analysis to support/refute broader performance claims about AutoML.

Returning to the chart of justifications in Fig. 9, the two most common assertions among academic publications posit that AutoML reduces the effort and expertise needed to employ ML. Obviously, there is a caveat here. Stakeholders who publish AutoML applications in academic journals are typically technically experienced individuals, and the tools involved often require code to be written. Therefore, academically published judgements on the utility of AutoML may not be representative for lay users. Indeed, one report talks explicitly in terms of “the researcher” [136] and another promotes AutoML as a time-saving technology for data scientists [119]. Nonetheless, various publications are more inclusive in how they perceive AutoML [547], and 12 specifically espouse the democratisation angle. A handful of the surveyed ML applications even discuss closely working with non-technical stakeholders [150, 217]. For instance, one acknowledges that selecting evaluation metrics is a technical process; it proposes personalising AutoML better by iteratively assisting users

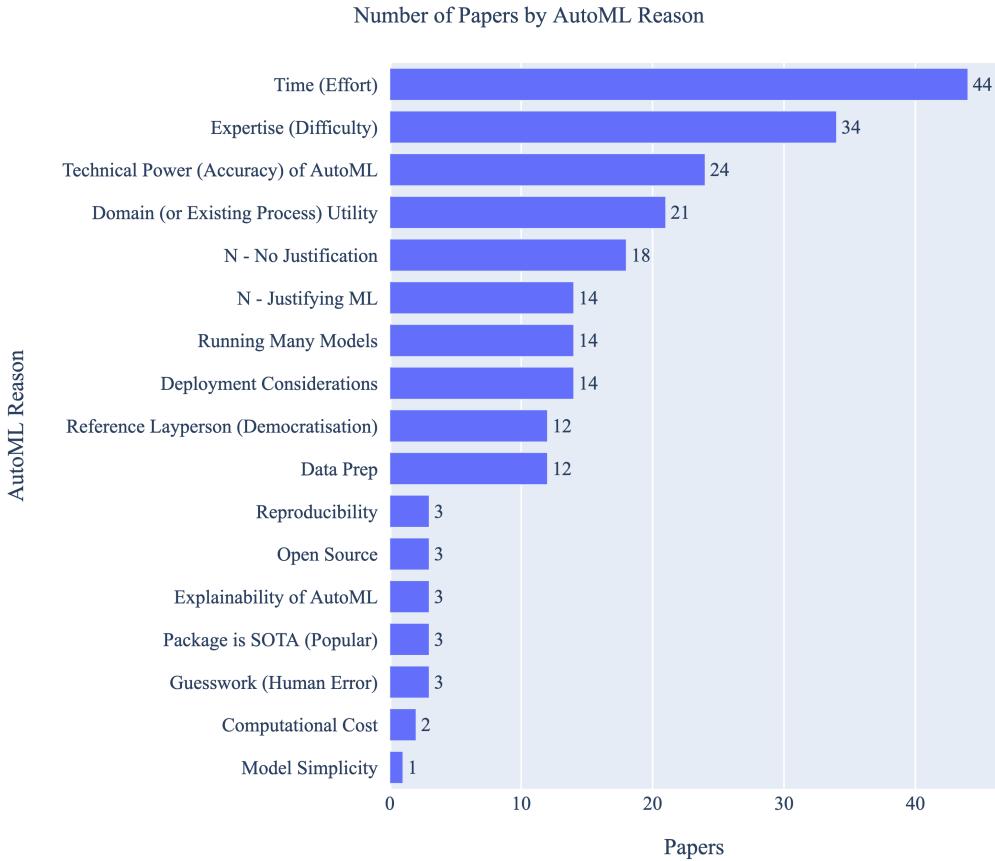


Fig. 9. The number of academic publications reporting on AutoML applications associated with each justification for AutoML use.

in making this choice [326]. Another example contemplates a similar stakeholder engagement process called ‘exploratory model analysis’ [137].

Next up, the fourth most popular reason for AutoML use, following faith in its technical prowess, is that the technology may be able to crack unsolved problems in new domains and improve upon existing processes that have not previously involved ML. Essentially, stakeholders that embrace this view are not primarily looking to push the limits of ML accuracy. They instead seek to pursue interdisciplinary endeavours, forging fruitful connections between methodology and context that provide ‘good enough’ preliminary outcomes. Certainly, when the barrier-to-use for ML techniques and approaches is reduced, interested researchers in other fields are freed up to focus on domain insights and otherwise accelerate their paper-publishing workflow.

As for the remaining justifications, several skew more towards explicit functionality rather than high-level motivations. For instance, one AutoML application is connected with the claim that automated processes can produce simpler ML models than a human [446]. This argument is somewhat contentious and depends very much on both implementation and problem context. What is much more uncontroversially appreciated is that AutoML software allows many possible models to be sampled, e.g. Auto-WEKA [52, 131, 164, 171, 186, 415, 455, 505, 533] and PyCaret [260, 322, 336].

In fact, for some AutoML applications, even the baseline offerings are seemingly insufficient for CASH; one project supplements Auto-WEKA with additional ML algorithms [67]. Finally, it is worth noting that a moderate amount of papers do not justify the use of AutoML, with a large portion of these not even bothering to rationalise employing ML for a particular domain. This result is interesting because it suggests AutoML may have already generated a reasonably authoritative reputation among prospective users. Unfortunately, assessing whether this reputation is warranted would require much deeper post hoc analyses of ML application outcomes as perceived by their stakeholders.

To some extent, a more balanced appraisal of AutoML usage is at least possible by noting any challenges reported while employing the technology. For instance, a couple of publications identify issues arising from small datasets [425, 569]. Of course, these complications are less likely to be problematic within larger enterprise environments, where the difficulty lies in extracting value from big data, not accumulating it. However, the more pertinent point here is that sparse data is a challenge for ML overall, not just AutoML. Presumably, the takeaway is that automation is not a magic solution for limitations *intrinsic* to an ML problem. Elsewhere, a couple of ML applications dwell on the inability to customise certain tools for domain-specific metrics and nuances [269, 423]. This matter seems to be a common concern, which is, again, why there has been a movement by some AutoML developers to cater for domain specialisation; see Section 5.2.

Occasionally, aligning closely with the interests of this review, some academic reports on AutoML applications have provided deeper commentary about potential obstacles to AutoML uptake. For example, a few papers consider a lack of explainability as a possible hindrance to broader user engagement [389, 571]. This recognition supplements current industry concerns and stakeholder requirements, legitimising the inclusion of the associated criteria for performant ML within Section 3.2. However, one publication stands out in particular for its fine-grained focus on small-to-medium enterprise (SME) and its uptake of AutoML technology [113]. The report cites a lack of expertise and funding to hire data-science talent – this implicitly rejects AutoML being sufficiently democratised – as obvious obstacles, but it also asserts that identifying business use cases for the technology may also be challenging. Such a concern is a rare and thus notable reference to the initial stage of an MLWF: problem formulation and context understanding. Until AutonoML becomes exceptionally advanced [313], supporting the translation of business objectives into ML tasks will need to be addressed in other ways.

Ultimately, a survey of AutoML applications published within academic literature reveals that the utilisation phase of the technology has begun in earnest. Many industries are proving fruitful for automated exploration, although, given this academic perspective, there is an obvious selection bias for observability that leans towards the research-heavy fields. In any case, this surge has only occurred within the last few years, so, while the proliferation is noteworthy, AutoML technology is far from being exploited to the limits of its capability. Partially, these constraints are governed by the software available at any particular time, which explains why few applications go far beyond the model-selection stage of an MLWF. However, among potential users, there is also perhaps a limited understanding of what AutoML truly promises regarding its roles and benefits. Indeed, this suggestion is most starkly supported by a disproportionate obsession with model-validity metrics that does not reflect what the application runners *themselves* claim to be the advantages of AutoML. Thus, the literature survey results end up serving as the perfect justification for bringing communal attention to this monograph and its more comprehensive definition of ‘performant ML’.

5.2 Domain Specialisation

In the course of reviewing AutoML software and applications for Section 4 and Section 5.1, respectively, it became apparent that several tools have been designed for highly specialised use cases.

This statement does not refer to the packages discussed in Section 4.1, which are specialised only in that they are ‘dedicated’ to particular tasks and stages within an MLWF. For instance, software that focusses on HPO in Section 4.1.3 is still general, developed to optimise – ideally – arbitrary functions, so long as they are both parameterisable and evaluable. This section, instead, examines the specialisation of AutoML tools/applications for particular technical problems and industrial domains.

Table 123. Specialised AutoML Tools. Notes the technical/industrial domain and whether the following mechanisms are included: Exp. for explainability, Viz. for visualisation, DP for data preparation, FG for feature generation, and FS for feature selection. Also details UI modes, HPO mechanisms, and library dependencies.

Name	Domain	Exp.	Viz.	UI	DP	FG	FS	HPO	Wraps	Ref.
Auto_TS	Time Series	N	N	Code	Y	Y	N	NONE	[Algorithms] Sklearn, FB Prophet, XGBoost, pmdarima	[494]
Luminaire	Outlier (Time Series)	N	Y	Code	Y	N	N	Bayesian	[HPO] Hyperopt	[142, 579]
TODS	Outlier (Time Series)	N	N	Code	Y	Y	N	NONE	[Algorithms] Sklearn, Tensorflow (TF), Keras, PyOD	[178, 329]
AlphaPy	Finance, Sport	N	Y	Code/CLI	Y	N	Y	Grid, Random	[Algorithms] Sklearn, Keras, XGBoost, LightGBM, CatBoost	[490]
Cardea	Medical Documents	N	N	Code	Y	Y	N	NONE	[Other] Compose, Featuretools	[66, 85]
ChemML	Chemistry	N	N	Code/GUI	N	N	N	Active Learning, Genetic	NONE	[86, 244]
EvalML	Fraud, Lead Scoring	N	Y	Code	Y	N	N	Bayesian, Grid, Random	[Algorithms] CatBoost, Sklearn, LightGBM, XGBoost, [HPO] skopt	[69]

Accordingly, the first lens through which to examine specialisation considers niche forms of data analysis. For instance, there is the objective of outlier detection. Certain AutoML tools, some discussed in Section 4.2.2, identify/manage obvious anomalies on the path to constructing better-fitting ML solutions, but this is not what is being referred to here. Instead, we highlight that some ML applications and their models are entirely dedicated to predicting unusual patterns of behaviour within select data environments. Academic examples of these applications have examined financial fraud [412], medical information about patients [254], and issues with machine maintenance [300]. Of course, there are many techniques that one may leverage in such a use case, with the Python library PyOD [575] implementing 33 algorithms from various sources at the time of this review. However, the tool itself is not considered AutoML, as substantial manual effort is needed to employ its offerings. In contrast, Luminaire [579] and TODS [178], both dissected in Table 123, offer outlier-detection capabilities for time series data with significant automation, e.g. in terms of feature

extraction and running/optimising algorithms. Both packages use scikit-learn, with the former depending on Hyperopt and the latter leveraging Keras, TF, and PyOD.

Incidentally, from a broader perspective, working exclusively with time-series data is itself a form of analytical specialisation. The Auto_ts package [494], also in the table, is one example that does so, automating many related processes. It supports users in generating ML solutions based on a limited but diverse pool of techniques, e.g. a random forest supplied by scikit-learn, an autoregressive integrated moving average (ARIMA) [128], or the Facebook Prophet [525].

Continuing onwards, the second lens through which to examine specialisation considers developments for particular industrial domains. Now, crucially, configuring for a specific problem *setting* typically has less of an impact on core ML processes than configuring for a specific problem *type*. For instance, in the AutoDL case, planning for time-series forecasting or image recognition tasks may suggest limiting CASH, specifically NAS, to recurrent or convolutional neural networks. In contrast, CASH designed for supervised ML on tabular data does not significantly change whether the data is financial, medical, and so on. Admittedly, if a problem type is virtually intrinsic to a problem setting, e.g. image recognition and the field of radiography, then the model-development phase of an MLWF can be configured appropriately in advance. However, in most cases, specialisation for an industrial domain predominantly impacts the input-output edges of an ML application, i.e. how data is prepared/engineered and how results are used/interpreted. In other words, convenience features are developed for the transition boundary between domain insights and computational data.

One of the earliest examples of AutoML software specialised in such a way is PredicT-ML [367], designed to perform ML on meaningful features that it extracts from big clinical data. Unfortunately, it cannot be robustly assessed beyond its associated publication, lacking either a public repository or any extensive follow-up. However, its existence does align well with the trend identified in Section 5.1, i.e. the heavy use of AutoML in the health & biomedical domains. Another more recent example also adhering to this trend is GenoML [374], designed explicitly for genomics, although it is too early to determine whether it will find significant purchase in academia.

Arguably exhibiting more uptake is ChemML [548], an open-source tool that has been cited in several chemistry investigations and applications [243, 431]. It can manipulate chemical data conveniently, e.g. by encoding and visualising molecules, and supports automated pathways to generating ML models based on relevant features. A similar package designed to work with RNA/DNA data is PyFeat [406], which has likewise seen decent use [57, 73, 371, 578]. It first constructs features from RNA/DNA sequences according to domain-specific options that a user can select, then optionally runs ML classifiers on the associated feature-enriched dataset. That said, the package is not particularly current in developer activity and only offers a CLI as an advanced UI, with which arguments are specified.

Here, the definition of AutoML already starts to blur, as the scope of ChemML and PyFeat is dictated more by the domain than by the practices of automating high-level ML operations. They can still solidly be classified as AutoFE packages, at the very least, but the priorities involved further colour the discussion in Section 4.3.8 on the ‘AutoML agenda’. An interesting question thus arises: in an age of closely integrated software, how rigid will the bounds on AutoML remain? For instance, Cardea [85] is an open-source tool designed to work with electronic health records. It does not intrinsically provide AutoML capability but does so holistically by wrapping other packages such as Featuretools, Compose, and MLBlocks. This software contrasts with another implementation in a similar space, i.e. AutoPrognosis [62, 63], which, although it has not been tended to in a while, was explicitly developed to construct ML pipelines for clinical prognoses automatically. The point here is that some packages broadly dedicated to data analytics for a specific domain have begun to integrate, either optionally or immediately, automated high-level ML processes.

The inverse, then, to a domain-specific package that includes AutoML functions is an AutoML package that includes domain-specific functions. Both are valid manifestations of specialisation, even if each prioritises one characteristic over the other. Accordingly, AlphaPy [490] is an example of this inverse, offering generic AutoML functionality at its base and wrapping around popular ML libraries such as XGBoost, LightGBM, CatBoost, and scikit-learn. What makes the software stand out is its supplementary ‘MarketFlow’ and ‘SportFlow’ pipelines, which are geared for financial analysis and sporting-event predictions, respectively. Interestingly, the AlphaPy framework seemingly sticks to grid/random search for its optimisations, which suggests the industry-proximal project is either (1) unaware of theoretical advances in AutoML, (2) incapable of implementing them, or (3) dismissive of their utility. Given finite developmental resources, the last option could be well justified if leveraging domain insights has more impact than employing SOTA techniques. Elsewhere, EvalML [69] similarly provides general AutoML functions, this time including Bayesian HPO via wrapping up the Scikit-Optimize library [489]. However, it too offers specially designed capabilities for industrial applications, i.e. fraud detection and lead scoring. These two use cases happen to be among the most recurrent according to vendor perspectives of AutoML; see Section 5.3. So, for the many stakeholders solely interested in these particular domains, EvalML likely holds a unique appeal, allowing users to avoid the brunt of system configuration and data formulation.

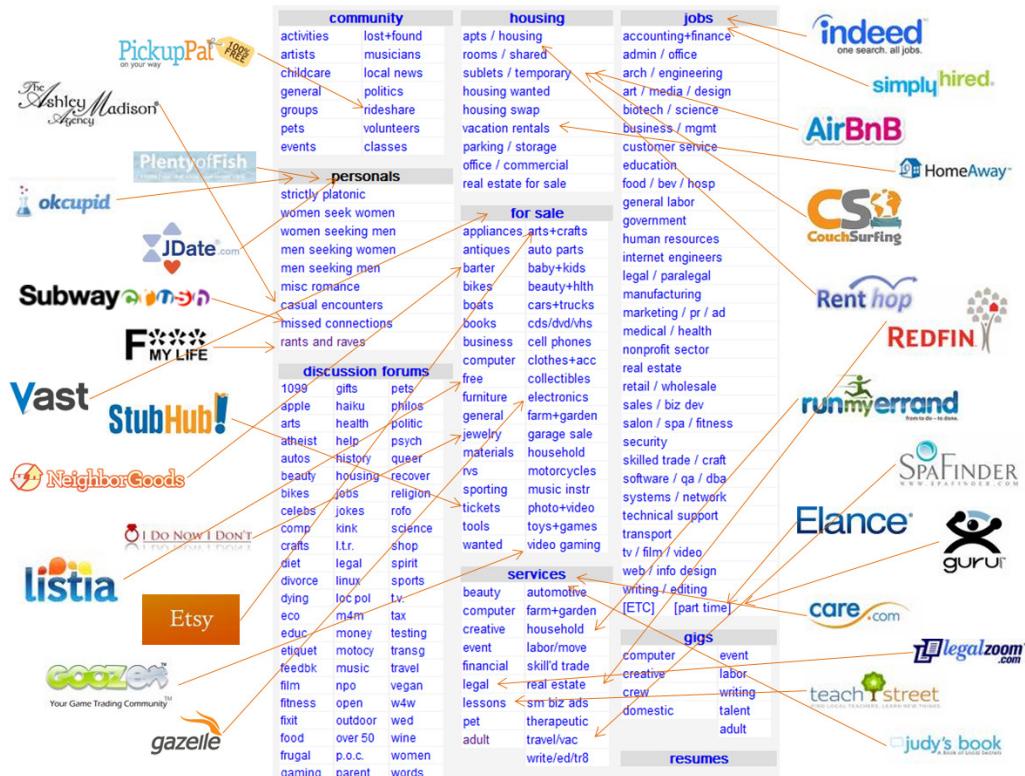


Fig. 10. A well-known diagram [428] linking startups to popular niches on Craigslist, which acts as an analogy to how AutoML developers may similarly pursue vertical growth strategies.

Ultimately, domain specialisation is still a nascent trend, even given the technological evolution of AutoML and its overall recency. After all, there is no point in investing substantial resources

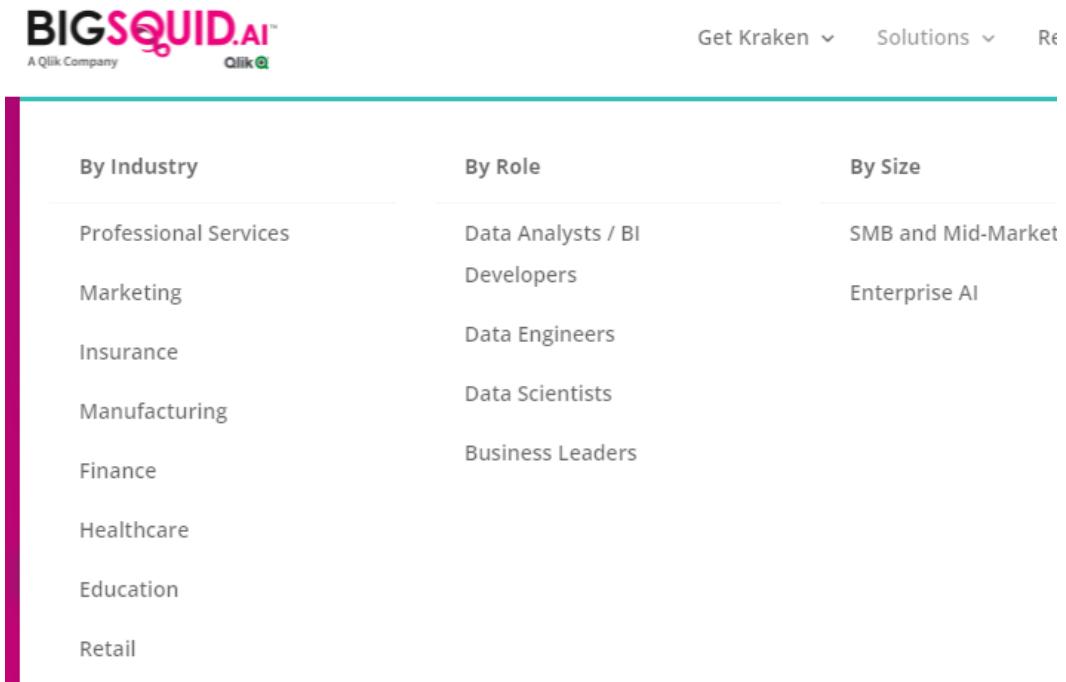
into fine-tuning software for a particular purpose/industry until the associated demand profile is evident. Accordingly, this phenomenon is closely tied with, if not reactive to, the body of AutoML applications that arises. Nonetheless, considering the speculation in Section 4.3.8 that, fragmented agendas aside, the general AutoML marketplace might eventually become an oligopoly, domain specialisation is one possible pathway to accommodate ongoing competition and differentiation. Indeed, analogised by Fig. 10 to how many startups grow vertically from existing niches, the future of AutoML technology may heavily involve prospective developers catering to specific tasks and industries. Such a process would undoubtedly contribute to the overall aims of the AutoML endeavour. For instance, consider a sports analyst lacking an understanding of general ML, let alone the skills to run an application without falling into common traps, e.g. data leakage [483]. While generic AutoML software might suffice for advancing their analytical ambitions, a closely assistive tool like AlphaPy could prove much less daunting, accelerating overall industrial uptake of the technology. Essentially, domain specialisation promises to further weaken the barriers to ML democratisation. Of course, only time will tell whether enterprising developers have *both* the domain expertise and academic skills to capably satisfy any emerging niche demand.

5.3 Vendor Reports

While academic literature tends to be transparent and peer-reviewed, it biases observability around AutoML applications towards those with apparent scientific benefits. However, a core tenet of the AutoML endeavour is that ML should ideally be made accessible wherever it can be found useful, including as part of far more mundane decision-making processes. So, short of exhaustively surveying ML stakeholders, the best way to assess the broader uptake of AutoML technology is to see what vendors have to say about it. Of course, all the usual caveats of taking commercial commentary at face value apply. There is no reliable window into sales volume or industrial activity; case studies and portfolios of services may be aspirational rather than representative. Nonetheless, this section collates promotional material from numerous vendors to gain at least a surface-level insight into the industries/problems the commercial AutoML sector targets.

Now, many AutoML vendors tend to use websites to list the domains in which they prioritise the targeting of their services. One example is shown in Fig. 11. Such summaries are often the result of two calculations: where a particular implementation is likely to have the most significant impact and, simply put, where the money is. Granted, while commercial income is a decent proxy for demand, other nuances can drive commercial agendas, e.g. maintaining a few well-paying clients in one industry may be more attractive than cheaply servicing numerous end-users in another. Regardless, setting aside contemplations of business strategy, we review 23 vendor websites and categorise the domains they target. Listed terms are occasionally split, combined or rephrased to sensibly fit a common framework. For instance, ‘Banking & Insurance’ targeted by Einblick translates to one tick for two categories, i.e. ‘Finance’ and ‘Insurance’.

The resulting distribution of the commercial focus on industries is depicted in Fig. 12. A long tail highlights the general utility of AutoML, but some domains receive higher levels of attention than others. There are numerous reasons why this may be the case, including a prevalence of big data and client funds. It is also possible that the types of ML problems arising in these targeted domains conform to well-defined archetypes, thus requiring less effort in the first MLWF phase, i.e. problem formulation and context understanding. Accordingly, the following sectors are well represented: finance, retail, marketing, and insurance. The ML model outcomes in these industries also tend to be proxies for profit, such as the volume of sales, so the value of AutoML can be quantified relatively quickly and objectively. Healthcare and energy also receive serious attention, as these are typically considered social goods and thus are often backed by government resources, either directly or via policies and less tangible benefits. For the same reason, more than a third of surveyed vendors



The screenshot shows a website header with the logo 'BIGSQUID.AI' (A Qlik Company) and the Qlik logo. Below the header, there are three columns: 'By Industry', 'By Role', and 'By Size'. The 'By Industry' column lists: Professional Services, Marketing, Insurance, Manufacturing, Finance, Healthcare, Education, and Retail. The 'By Role' column lists: Data Analysts / BI, Developers, Data Engineers, Data Scientists, and Business Leaders. The 'By Size' column lists: SMB and Mid-Market and Enterprise AI.

By Industry	By Role	By Size
Professional Services	Data Analysts / BI	SMB and Mid-Market
Marketing	Developers	Enterprise AI
Insurance	Data Engineers	
Manufacturing	Data Scientists	
Finance	Business Leaders	
Healthcare		
Education		
Retail		

Fig. 11. A snapshot of a commercial AutoML website, which lists industries targeted by its services. The vendor associated with this example is BigSquid.

seek to offer AutoML services to the public sector. In the meantime, privacy concerns and national defence considerations may stifle activity within some of the lower-ranked sectors, e.g. military, aviation, maritime, and cybersecurity. However, associated contracts may simply be less publicised.

The final analysis in this monograph involves looking into case studies promoted by AutoML vendors. These reports are typically accessible via associated websites, e.g. within sections named ‘blog’ or ‘resources’. Of course, due to the lack of peer review and often both depth and detail, only a glancing commentary can be made. Moreover, not all vendor publications are relevant or useful to this survey. For instance, partnership announcements are occasionally lumped in with the case studies. Alternatively, reports may be written from the perspective of someone ‘experimenting’ with the commercial AutoML product. Even among the remainder, case studies that do not clearly detail a business problem serve no purpose for a rigorous review.

Thus, the end result is a survey of 20 vendors and 130 documents of associated content. While attempts were made to cover the same ground as in Section 5.1, a lack of published information hampers an identical analysis. This outcome is not unexpected. For instance, vendors are unlikely to discuss the challenges of using AutoML while spruiking their products. Ultimately, the only significant insights that can be extracted from this study relate to the problem types targeted by AutoML vendors, displayed in Fig. 13. As a side note, given that some vendors dedicate multiple reports to a single use case, the charted counts are normalised. Accordingly, while product recommendation would otherwise top the rankings with 16 case studies, only seven vendors advertise this use case.

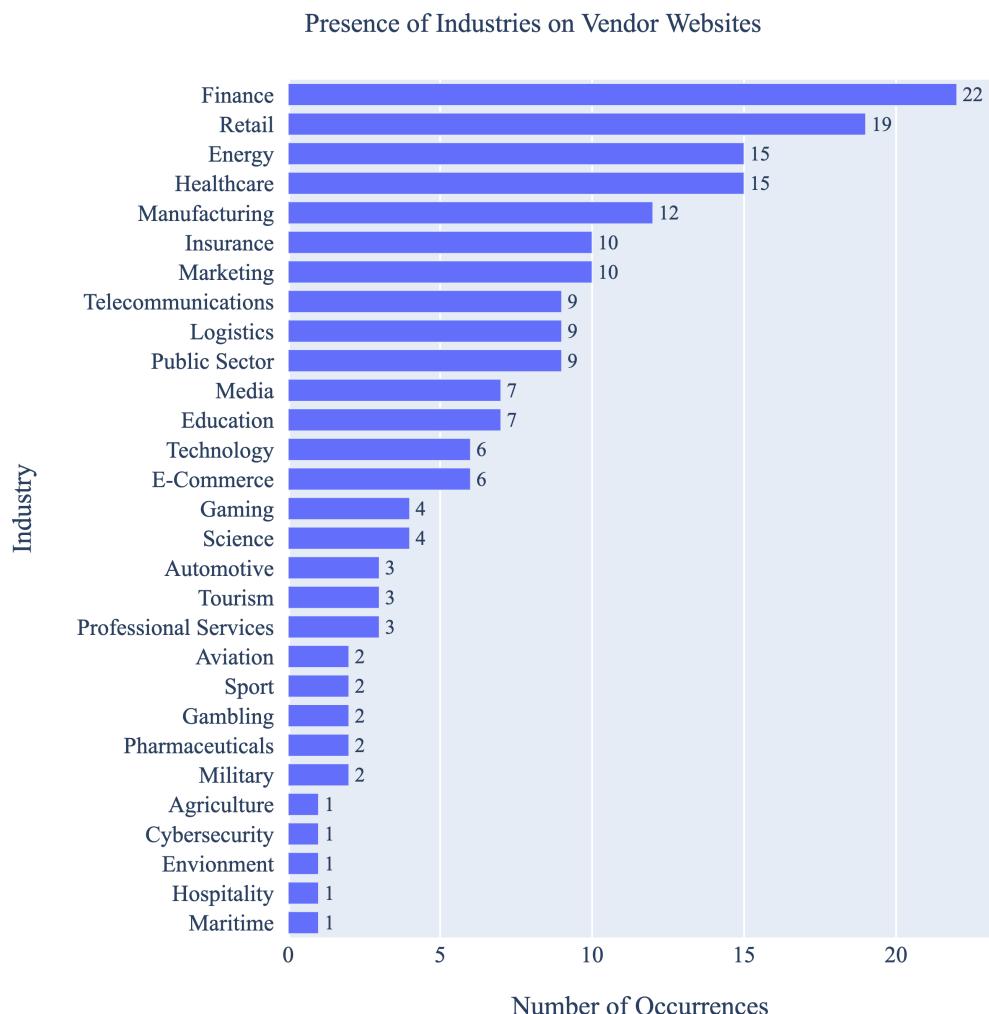


Fig. 12. Major industries suitable for ML. Each one is associated with the number of commercial AutoML websites that target the industry. This survey covers 23 vendors.

Unsurprisingly, the long tail on the chart indicates that AutoML is deemed useful for a diverse array of insight-gathering and decision-making problems. Certainly, with supervised learning as the default form of ML, classification/regression tasks can be formulated within numerous industrial contexts. Of course, as with the targeted industries, particular problems see the most *apparent* use of AutoML. These ML applications include demand forecasting for products/services, predictive maintenance, and product recommendation to new and existing customers, i.e. cross-selling and up-selling. Such problem types may be well-suited to AutoML, as they are often accompanied by plenty of labelled data and a clear target/objective.

However, it is notable that the term ‘influential factors’ also ranks very high; associated ML tasks involve not just making predictions but also understanding what is *driving* those predictions.

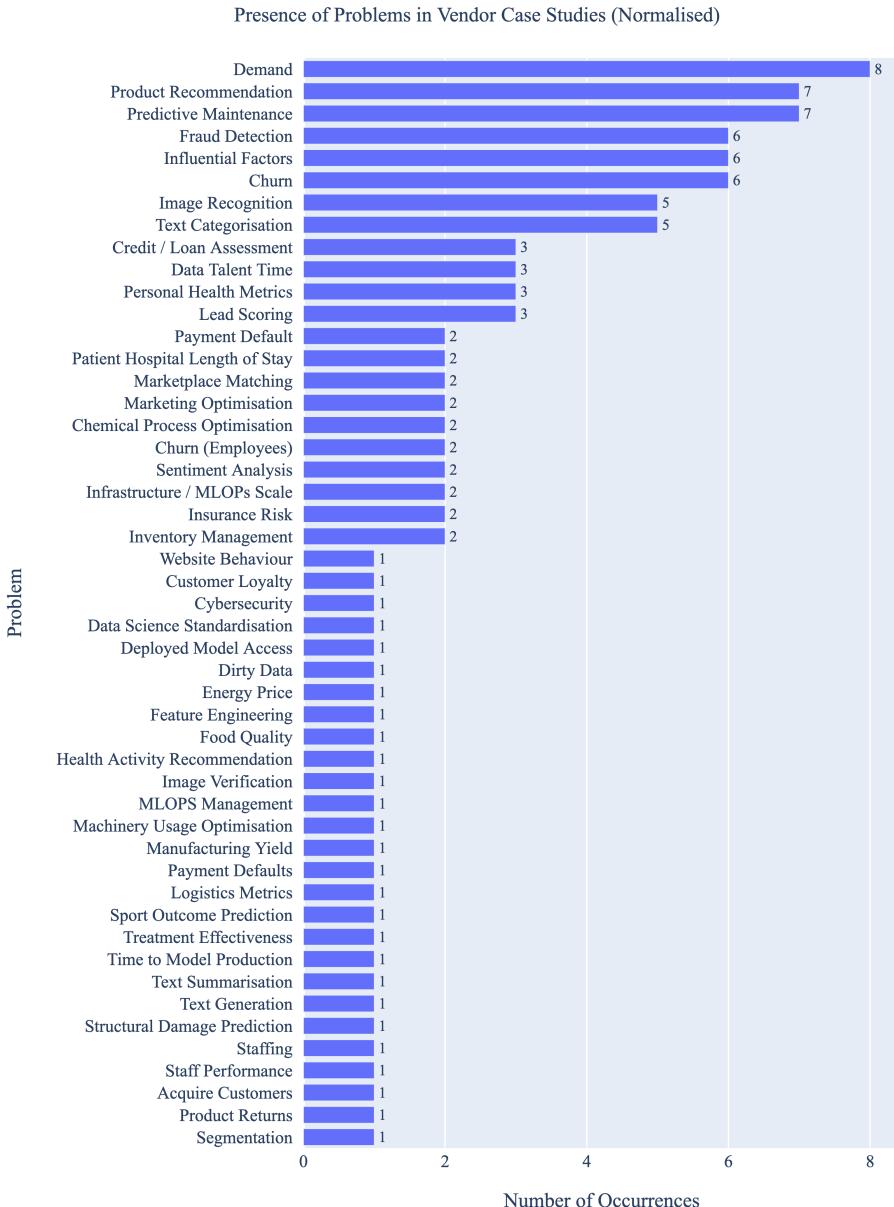


Fig. 13. Problem types suitable for ML. Each one is associated with the number of commercial AutoML websites that dedicate at least one case study to the problem. This survey covers 20 vendors.

Admittedly, it is unclear why so many case studies focus on this problem type, i.e. whether vendors are preparing this promotional material based on the features they have developed or on the demand they expect. Nevertheless, the point here is that traditional technical performance is not the be-all and end-all for such ML tasks, and other factors, including explainability, necessarily influence successful outcomes. In effect, commercial entrepreneurs are exploring just how versatile AutoML

can be. Thus, for those in the broader AutoML research community, it is a timely opportunity to consider what it means for the technology to be ‘performant’ in all the use cases listed within Fig. 13. Such considerations and associated innovations may strongly influence future technological uptake and impact.

6 CRITICAL DISCUSSION AND FUTURE DIRECTIONS

This comprehensive review of AutoML as a technology has comprised multiple analyses representing various angles of attack, all grappling with several fundamental questions. Simply put, as of the early 2020s, what does AutoML technology look like? How is it being offered? How is it being used? And is it any good? In fact, this last question of ‘goodness’ motivated substantial contemplation in Section 3, urging AutoML researchers to look beyond standard technical performance and consider the bigger picture, i.e. the potential requirements of all possible stakeholders in an ML application. Accordingly, equipped with a refined perspective, this monograph has made seminally broad assessments of supply, demand, and service quality. Specifically, surveys of open-source/commercial software and academic/vendor reports have all aided in appraising industrial engagement with AutoML, both present and potential. Nonetheless, with so much content, it is worth summarising/discussing the most prominent insights extracted from the analyses.

Almost no one is touching automated problem formulation and context understanding. As a reminder, such an assessment concerns the first stage of an MLWF, i.e. the conceptual framework introduced in both Section 2 and Fig. 2 to encompass a diversity of high-level ML operations within one inclusive workflow. Accordingly, it makes sense here to structure some of the overarching commentaries according to the main phases of an MLWF. However, there is little to say about the initial stage, when business plans transform into a machine-interpretable problem. Indeed, sophisticated mechanisms do not exist within the commercial sphere, despite several vendors offering templates and code repositories for common use cases. The situation is not much different among open-source packages either, with only a couple trying NLP and adjacent methodologies to interpret user desires efficiently [154, 496]. In effect, virtually all AutoML software starts with the premise that stakeholders have already created a well-defined project plan, for which resourcing is approved and project controls have been set up. This assumption is a poor one in reality, especially as organisations new to ML will likely struggle to translate their business objectives into well-scoped projects [113]. Admittedly, automating the formulation phase is a matter of HCI and may additionally require an AutoML paradigm shift [313], e.g. the design of reasoning mechanisms. After all, business agendas are only rigidly limited by human creativity, so mapping an infinite space of real-world problems into concrete ML processes very much remains an open research question.

Automated data engineering is reasonably well supported, but there is room for improvement. This second phase of an MLWF is the most closely linked to model development, given that data transformations can themselves be bundled into an ML pipeline or a deep neural network, so it makes logical sense that, historically, it was the first to be enveloped by the expanding scope of AutoML. Unsurprisingly, at the current time, there are several tools dedicated solely to data engineering, as listed in Section 4.1.1. Likewise, the majority of surveyed ‘comprehensive’ systems, whether open-source or commercial, provide assistance/automation with basic facets of cleaning or preparation. However, the coverage is incomplete, and packages that, for instance, do not manipulate date-time stamps or tokenise text will struggle to construct the most information-rich version of data for ML modelling. Similarly, Table 22 and Table 66 show that AutoFE, for both feature generation and selection, has not yet been consistently integrated into the broader AutoML ecosystem. This state of the technology means that users still require the technical expertise/support to manually clean, prepare and enhance data for many ML applications. Such an outcome is not

optimal, given that the data-engineering stage of an MLWF is a large time-sink. Moreover, Fig. 7 confirms that there is a high demand for automated data engineering among academically published ML applications. Thus, the idealised form of AutoML services, even for the core data-to-model conversion process, cannot be realised without further developmental progress in this phase.

The focus on model development is most substantial, but notable gaps in coverage exist. First of all, it is no surprise that all comprehensive AutoML software heavily automates the construction of ML models, given that CASH is the capability that initially brought the field to prominence. Virtually all surveyed academic reports of AutoML applications involve some form of model selection. Thus, as detailed by Table 34 and Table 81, AutoML tools, whether open-source or commercial, almost universally support standard forms of supervised ML. Time-series forecasting is also decently well represented. However, the provision of more exotic services is scattered. Stakeholders will need to do their research when shopping for AutoML software if they have problems related to unsupervised learning, computer vision, and NLP. Furthermore, particularly on the commercial side, Section 4.3.8 notes that transparency is an issue; just because a vendor claims to support an ML use case does not mean that the details of this support cover all possible stakeholder requirements. Accordingly, the quality of model outcomes can be tough to guarantee under the broad framework for ‘performant ML’ introduced in Section 3.2.

Thus far, there has been little effort while developing AutoML technologies to ensure trust in ML. Indeed, this assertion is among the most predominant concerns when discussing possible deficiencies within model outcomes. Many factors feed into whether a human is willing to trust the validity/utility of ML, and two are particularly pertinent here: explainability and fairness. Now, Table 47 and Table 93 indicate that existing AutoML software does generally assist interpretability on a global level, e.g. identifying feature importance. However, more sophisticated treatments are rare in commercial products and, according to this survey, are entirely absent in major open-source offerings. This situation means stakeholders wishing to understand ML predictions/prescriptions for individual queries, e.g. distinct people, may be forced to supplement a comprehensive system with the dedicated tools in Section 4.1.2. As for scenario analysis, which could shed further light on problem contexts and constructed ML models by supporting counterfactual thinking, very few AutoML systems incorporate such a mechanism. Of course, explainability can verify whether ML worked free of technical error, but it is also a prerequisite for identifying whether an ML solution is fair. Bias can crop up via human decisions, non-representative data or algorithmic assumptions, so it is a significant deficiency that, among comprehensive AutoML systems, its automated detection and management are rare and practically nonexistent, respectively. Consumers, businesses and general society are becoming increasingly aware of the socioethical impacts that real-world ML applications can have, and regulatory policies and legislation are steadily coming online. Thus, while the topic is understandably complex [313], it does not suit the ethos of AutoML to leave the daunting task of selecting/interpreting bias-and-fairness metrics to a non-expert.

Publicly available AutoML software is hardly ‘cutting edge’, but that may be okay; it is still unclear what contributes most to an optimal ML solution anyway. In Section 4.3.3, a comparative analysis of HPO-mechanism availability within open-source systems and commercial products was accompanied by a discussion asserting that the technical depth of the latter seems relatively shallow. Granted, this assessment is made holistically and is also possibly skewed by the heavy presence of active academics within the open-source community. Nonetheless, it does seem like, for the model-development phase, commercial software is more likely to employ rudimentary grid/random search than any bandit-based refinement of Bayesian optimisation. As for more sophisticated concepts explored within academia, their presence is felt sparsely across *all* offerings, i.e. meta-learning, ensemble techniques, ML pipeline optimisation, and so on. Now, as mentioned

in Section 5.2, the absence of theoretical advances in a package can be due to (1) ignorance, (2) an inability to implement, or (3) intentional rejection. In this last case, it is possible that such an attitude could be justified. Maybe technical depth is not actually required for an AutoML system to support performant ML. Many modern SOTA approaches in data science burn up increasing amounts of resources to squeeze out decreasing increments of improvement, often all focussed on accuracy metrics that, as this review has argued, may not be critically important. Even if they were, perhaps ML applications would be better served by a data-centric approach rather than fixating on fine-tuning models for a static dataset. Intelligently controlling data flow, not just through AutoFE but also by the selective presentation of samples, may potentially impact model outcomes more than CASH. Whatever the case, many commercial AutoML providers are presently surviving without needing to employ SOTA techniques. Perhaps operational efficiencies gained across an entire MLWF make up for any model-development shortcomings. Alternatively, AutoML may be so innovative among its target industries that any rudimentary approaches still provide immense value. Only time will tell if developers start following academic research more closely once the easy pickings dry up.

The limited provision of automated deployment, monitoring and maintenance is a predominantly commercial affair. After all, as a sweeping generalisation, academia appears to be preoccupied with the static accuracy of ML models. This assertion is supported by Fig. 8, which shows the metrics that overwhelmingly matter when publishing AutoML applications via academic journals. Essentially, all scientific interest seems to end with the model-development phase of an MLWF. One may argue that what follows tends to be dismissed as a mundane and messy engineering problem. This perspective may also explain why free AutoML services that are not directly linked to academia likewise do not bother assisting with the deployment stage. In contrast, vendors, whose clientele can include particularly non-technical stakeholders, are contractually obligated to provide end-to-end support. Outside of insight-gathering applications, an ML model object suits no business organisation if it cannot be put into production to supply end-users with predictions/prescriptions. Accordingly, most surveyed commercial systems have found it necessary to venture into the space of MLOps. That said, Table 105 questions the contemporary sophistication of these services, e.g. in terms of testing and updating deployments. The situation is similar for the final phase of an MLWF: monitoring and maintenance. Many surveyed vendors do actually have basic mechanisms in place to alert clients about degrading model performance, triggering retraining if desired. However, Table 109 reveals a dearth of systems that actively monitor data for concept drift, let alone proactively update an ML model. Of course, it is presently unclear what the best policy for triggering adaptation is [580, 583], and vendors are also likely to prefer humans in the loop to limit liability for poor model outcomes. Nonetheless, there is a long way to go before AutoML services can be claimed to support continuous learning, a prerequisite for next-generation AutonoML frameworks [308, 313].

Making a business out of AutoML drives different priorities compared with providing it for free. To support this point and underscore several preceding insights, Table 124 highlights just how differently this monograph rated open-source and commercial AutoML systems for major criteria. This comparison includes considerations of the deployment and management effort (DM), plus governance and security (G), though it is noted that all free providers effectively ignore these facets and score zero, bringing down all their overall scores. Nevertheless, taking into account all the criteria listed in the table, it is worth noting that five packages each obtained over half of the available points, namely Dataiku (0.734), DataRobot (0.729), H2O (0.635), Microsoft (0.615), and SageMaker (0.521). They are all commercial. So, after first accounting for the skewing effect of these five, we find there is little overall difference between open-source and closed-source tools in terms of operational/technical efficiencies (E). In fact, when it comes to matters of technical

Table 124. Summary of how the surveyed open-source and commercial AutoML systems scored for major criteria. Includes (1) total scores incorporating all the included criteria and (2) average scores for the open-source and commercial software tools. Each score is relative to the maximum achievable and is contained within the 0 to 1 range, with 0 meaning that none of the assessed aspects have been met and 1 that all of them are perfectly fulfilled. Colour coding: the darker green colour means better performance. Abbreviations: E - efficiency, DD - dirty data, CC - completeness & currency, EX - explainability, EU - ease of use, DM - deployment and management effort, and G - governance and security.

Name	E Total	DD Total	CC Total	EX Total	EU Total	DM Total	G Total	Overall Proportion
Auto ViML	0.452	0.500	0.417	0.222	0.286	0.000	0.000	0.333
Auto-PyTorch	0.323	0.500	0.458	0.000	0.286	0.000	0.000	0.281
auto-sklearn	0.226	0.625	0.667	0.000	0.429	0.000	0.000	0.323
AutoGluon	0.290	0.375	0.667	0.222	0.429	0.000	0.000	0.344
AutoKeras	0.194	0.375	0.708	0.000	0.429	0.000	0.000	0.302
AutoML Alex	0.258	0.750	0.542	0.000	0.429	0.000	0.000	0.313
carefree-learn	0.290	0.500	0.542	0.000	0.429	0.000	0.000	0.302
FLAML	0.258	0.375	0.583	0.000	0.286	0.000	0.000	0.281
GAMA	0.161	0.375	0.417	0.222	1.000	0.000	0.000	0.281
HyperGBM	0.355	0.500	0.458	0.000	0.857	0.000	0.000	0.333
Hyperopt-sklearn	0.387	0.375	0.417	0.000	0.286	0.000	0.000	0.281
lgel	0.226	0.625	0.583	0.000	0.857	0.000	0.000	0.333
Lightwood	0.194	0.375	0.417	0.222	0.286	0.000	0.000	0.240
Ludwig	0.226	0.500	0.583	0.222	0.571	0.000	0.000	0.323
Mijar	0.258	0.500	0.542	0.222	0.429	0.000	0.000	0.313
mlr3automl	0.258	0.500	0.417	0.000	0.286	0.000	0.000	0.250
OBOE	0.290	0.750	0.375	0.000	0.286	0.000	0.000	0.271
PyCaret	0.258	0.875	0.667	0.222	0.429	0.000	0.000	0.375
TPOT	0.290	0.625	0.750	0.000	0.571	0.000	0.000	0.375
Average score - open-source systems	0.273	0.526	0.537	0.082	0.466	0.000	0.000	0.308
Name	E Total	DD Total	CC Total	EX Total	EU Total	DM Total	G Total	Overall Proportion
Alteryx	0.306	0.875	0.333	0.000	0.714	0.429	0.667	0.391
Auger	0.452	0.750	0.375	0.000	0.714	0.357	0.000	0.406
B2Metric	0.161	0.500	0.458	0.000	0.286	0.429	0.667	0.313
Big Squid	0.323	0.625	0.333	0.556	0.571	0.143	0.000	0.354
BigML	0.226	0.500	0.583	0.222	0.857	0.214	1.000	0.406
cnrvg.io	0.323	0.500	0.375	0.111	0.857	0.714	0.333	0.427
Compellon	0.161	0.375	0.083	0.222	0.286	0.000	0.000	0.146
D2iQ	0.161	0.250	0.208	0.111	0.714	0.357	0.667	0.260
Databricks	0.452	0.125	0.458	0.222	0.571	0.214	1.000	0.396
Dataiku	0.790	0.250	0.792	0.778	0.857	0.643	1.000	0.734
DataRobot	0.839	0.250	0.708	0.667	0.714	0.786	1.000	0.729
Deep Cognition	0.226	0.125	0.250	0.222	0.714	0.143	0.000	0.240
Einblick	0.226	0.125	0.500	0.222	0.571	0.000	0.667	0.292
Google	0.387	0.250	0.542	0.667	0.857	0.357	1.000	0.490
H2O	0.613	0.250	0.833	0.556	0.857	0.643	0.000	0.635
IBM	0.355	0.125	0.333	0.778	0.714	0.571	1.000	0.448
KNIME	0.226	0.125	0.458	0.000	0.857	0.429	1.000	0.354
Microsoft	0.613	0.250	0.625	0.556	0.857	0.643	1.000	0.615
MyDataModels	0.194	0.250	0.250	0.333	0.429	0.000	0.000	0.208
Number Theory	0.290	0.125	0.542	0.222	0.286	0.571	0.000	0.365
RapidMiner	0.323	0.125	0.563	0.444	0.857	0.500	1.000	0.464
SageMaker	0.387	0.125	0.542	0.778	0.857	0.571	1.000	0.521
SAS	0.323	0.125	0.500	0.444	0.714	0.500	1.000	0.438
Spell	0.194	0.000	0.417	0.111	0.714	0.357	0.667	0.302
TIMi	0.032	0.000	0.167	0.222	0.143	0.000	0.000	0.083
Average score - commercial systems	0.343	0.280	0.449	0.338	0.663	0.383	0.587	0.397

depth, e.g. cleaning/preparing dirty data (DD) and providing coverage of ML approaches that is complete and current (CC), free AutoML services are generally ahead. One can even hypothesise that, if benchmarking technical performance was included in this survey, open-source systems would likely have the greater range of options to pursue SOTA ML results. However, when it comes

to explainability (EX) and ease of use (EU), critical requirements for democratisation, vendors are solidly ahead, scoring 0.338 versus 0.082 and 0.663 versus 0.466, respectively. Admittedly, the quantification in this monograph is open to debate, and we do not claim there are well-defined thresholds for service quality or business value. Nonetheless, it is clear that when the stakes are high, and the survival of AutoML software is closely coupled with stakeholder engagement, certain aspects attain an importance they did not previously have.

There is no singular roadmap for the future of AutoML technology. Admittedly, this particular finding may seem at odds with previous suggestions that imply some logical order to advancing the field. However, as discussed in Section 4.3.8, there are many actors with varying origins and diverse agendas telling this story. So, while global trends and directions may be emergently evident, the localised threads of development are much more unpredictable and chaotic; there may yet be surprises in store for what ‘AutoML Tech’ comes to mean. For instance, concerning the latter stages of an MLWF, it is clear that different parties are converging on assisted/automated MLOps from two different directions. Some are pushing forwards from core AutoML practices such as CASH, while others are entering the space by expanding the scope of generic DevOps platforms. It is not even clear which entities will outcompete each other. The latter tend to have well-established infrastructures to lean on, while, as recently mentioned, it is not apparent whether any advanced model-selection methodologies offered by the former constitute a genuine advantage. Thus, from an industrial consumer perspective, will AutoML platforms offer MLOps? Or will MLOps platforms offer AutoML? Then there are higher-level questions. It is clear that an integrated comprehensive system, both end-to-end and generally applicable, is theoretically ideal in terms of broad utility [308]. Yet AutoML tools dedicated solely to individual MLWF phases/processes, as in Section 4.1, and technical/industrial domains, as in Section 5.2, benefit from being lightweight and focussed. Maybe it is more practical for sufficiently technical stakeholders to stitch together a patchwork of AutoML services as desired, assuming commercial tools are transparent enough to allow this, leaving comprehensive systems to those who need/want to relinquish finer control. Again, this review cannot pass judgement here, as it is unclear how much a dedicated developmental focus compensates for restricted utility and the technical challenges of subsequent integration.

Ultimately, further studies based on a broader vision of ‘performant ML’ are required to honestly assess whether AutoML technology is living up to its full potential. That is not to say that the endeavour of automating high-level ML operations has not already made a splash in its last decade of mainstream dissemination. Numerous enterprising developers have jumped on the bandwagon, offering a diverse supply of AutoML products. Demand is likewise surging if the yearly numbers of published AutoML applications charted in Fig. 3 are anything to go by. Nonetheless, this means that it is even *more* crucial in this phase of initial contact between technology and industry that the best first impression is achieved and maintained. Already, this review has identified numerous gaps in the services offered across both the open-source sphere and commercial marketplace. Admittedly, the impact of these deficiencies will obviously differ. For instance, not being able to cover certain types of ML problem merely limits the prospective clientele for associated tools. On the other hand, obsessing with technical metrics at the cost of appreciating the holistic stakeholder experience may frustrate engaged users and damage future uptake. Developers should thus carefully consider the broader criteria introduced in Section 3.2 when assessing their implementations. Of course, further deliberation and debate are invited so such factors can be quantified as objectively as possible; a scientific consensus can only make ensuing benchmarks for AutoML services much more authoritative. Finally, while this monograph has primarily attempted to assess the current state of AutoML technology via literature surveys, its limitations are clear. For instance, vendor-promoted case studies are a poor substitute for academic reports, often exhibiting bias, spin, and a lack of verifiable details. Instead, directly approaching

AutoML users is likely a better litmus test for demand, although such surveys should also consider a diversity of potential stakeholders beyond just data scientists [198, 550]. Fundamentally, the AutoML community cannot optimise the translation of theory to technology if it cannot effectively observe and identify what industry likes and what industry wants. Simply put, there is more work to be done.

7 CONCLUSIONS

This monograph has reviewed the technological emergence and industrial uptake of AutoML as they appear in the early 2020s. It is the first to comprehensively assess how this translation of academic theory to mainstream practice has fared, distinguishing itself from its forerunners [195, 308] and other overviews in the literature that focus instead on the fundamental concepts behind AutoML. Unsurprisingly, such an endeavour grapples with many questions. Who are the people that are likely to engage with or be impacted by AutoML, both presently and potentially? What do they want from AutoML? How is the technology currently being supplied? What is it capable of? And what is it being used for? Commentary on these topics has been heavily informed by surveys of documented codebases, promotional material, and application reports.

Before undertaking any analyses, it was first necessary to motivate and define a lens through which the state of AutoML technology could be appraised. The following is a summary of this preamble:

- Section 2 formalised the nature of an ML application, i.e. a collection of activities aiming to create/exploit a data-driven solution to a problem of descriptive/predictive/prescriptive analytics via ML techniques. Specifically, the section introduced the notion of an MLWF, designating a systematic way to organise tasks common to an ML application within one encompassing but segmented workflow representation. Such a conceptual framework makes it possible to assess which of the following operational phases are targeted by AutoML technology: problem formulation & context understanding, data engineering, model development, deployment, and monitoring & maintenance.
- Section 3.1 introduced the notion of an ML stakeholder, i.e. a person or party with direct or indirect interests in the processes and outcomes of an ML application. The section then elaborated on the likely desires and requirements each type of stakeholder brings to the table when engaging with ML. Such considerations allow assessments of AutoML to extend beyond the data-scientist experience and, in particular, a fixation on technical metrics for ML model validity.
- Section 3.2 proposed an extensive assessment framework to determine how well an AutoML service supports ‘performant ML’. The associated criteria are based on a holistic view of what matters to stakeholders and are grouped into the following categories: efficiency, dirty data, completeness & currency, explainability, ease of use, deployment & management effort, and governance.
- Section 3.3 finally condensed the assessment framework based on stakeholder requirements to generate insight and commentary about what industry presently sees as the role of AutoML. Overall, there appear to be three primary goals for the technology as it relates to data science practices: enhancement, democratisation, and standardisation.

Having established a comprehensive reference frame for what makes AutoML ‘good’ within a real-world setting, this monograph then proceeded to present and dissect several surveys of existing tools and application reports. The following summarises the analyses specific to AutoML software:

- Section 4.1 began appraising supply by examining AutoML tools that are termed ‘dedicated’; these do not intrinsically feature the core processes of generating/managing ML model objects. Instead, the packages are meant to be used with other software, focussing only on specific segments and responsibilities within an MLWF, such as (1) data and feature engineering, (2) bias, fairness and explainability, and (3) HPO. Overall, these ‘detachable’ tools can provide flexibility and control to a technical user, but integration challenges may hinder broader uptake.
- Section 4.2 considered open-source AutoML software termed ‘comprehensive’; these implementations take charge of training ML models at a low level while also automating higher-level operations. Essentially, this survey found that the field of free options available to interested stakeholders is broad, with nuanced discussions arising for each of the aforementioned criteria. However, there is a general sense that these systems, often tied to academia, are frequently inspired to innovate at technical depth, e.g. with sophisticated model selection. The trade-off is that almost all have no interest in the productionisation stages of an MLWF, and it is also rare to see HCI options for non-technical stakeholders receive serious developmental attention.
- Section 4.3 continued assessing the supply of comprehensive AutoML systems, although now focussing on the commercial sphere. Again, many products ready for organisational use were identified, even though the capabilities and features their developers invest in vary. Unsurprisingly, transparency was an obstacle to this analysis, but the details suggest that commercial AutoML goes broad, not deep, contrasting with open-source offerings. For instance, vendors seeking non-technical clients are motivated to focus on accessible UIs and support some form of MLOps, rather than implement a SOTA CASH mechanism. However, the survey results also indicated that no one particular agenda dominates the space of AutoML technology; developers are scattered in their origins, approaches, and priorities.

The following summarises the analyses specific to AutoML applications:

- Section 5.1 began appraising demand by examining academically published AutoML projects run for real-world problems. The rate of these applications being published was found to be accelerating, and, although usage appears to be dominated by the health & biomedical sectors, the spectrum of industries involved affirms the broad utility of AutoML. However, the academic context of these publications obviously distorts what could be surmised about the technology. In particular, there was a jarring discrepancy between the contemplative ways authors promoted AutoML and the routine ways it was assessed to be performant; judgements predominantly fell back on technical metrics for model validity.
- Section 5.2 briefly acknowledged a rising trend, i.e. the increasing number of AutoML tools being specialised for particular use cases, such as niche forms of data analysis or associations with specific industries. This phenomenon is arguably driven by expectations of demand, as expending effort on specialised functionality has little to gain if it is not used. For now, the trend is too nascent to earn significant commentary, but it reaffirms a finding in this monograph that there are many ways for AutoML technology to evolve ‘in the wild’.
- Section 5.3 finally capped off analysing demand by inspecting AutoML vendor websites and associated case studies. Although such promotional material should be reasonably well representative of what customers want from their commercial AutoML services, the usefulness of the reports was obviously limited by a lack of peer review and detail. Nonetheless, vendors were found to be targeting a selection of industries somewhat dissimilar to those linked with academically published AutoML applications, which again highlights the general utility of the technology. Moreover, the diverse array of problem types that AutoML is servicing

further suggests that high-quality outcomes depend on more than just technical accuracy metrics.

Ultimately, the value in this review is ideally found in the broader perspective it promotes for evaluating a *technology* as opposed to a science. By this stage, preceding monographs have already argued strongly that the theoretical potential of AutoML [308] and AutoDL [195] is immense. An enterprising researcher may envisage logical pathways – not necessarily easy ones – towards inspirational ambitions, including AutonoML systems capable of continuous learning, dynamic self-assembly, and general applicability. However, AutoML implementations ready for societal use are clearly nowhere near such lofty heights. This review even finds suggestions that publicly available products lag behind contemporary SOTA techniques, implying developer ignorance, an inability to implement, or deliberate rejection. Indeed, if fundamental AutoML advances are being actively dismissed by developers, despite researchers finding them promising, then this disconnect is a troublesome one and should be understood; failure to do so will only cripple the translation and impact of AutoML. The fact is, for better or worse, technology is what showcases a science to investors and heavily influences, if not determines, how many resources are cycled back for further research and development. Accordingly, the AutoML community cannot neglect to examine how industry is engaging with the technology, and this requires a solid grasp of what ‘performance’ means in the general case. Real-world ML applications involving a diversity of stakeholders are subject to many more complicated needs and pressures than a sanitised benchmark challenge tackled by a group of expert ML practitioners.

Granted, to the credit of AutoML developers, this review has found many promising implementations of AutoML software, whether open-source or commercial, end-to-end or dedicated, generic or specialised. Nonetheless, at the same time, numerous gaps have been identified in the services provided. Perhaps these do not matter, and clients will still appreciate any added value that ML can extract from their business contexts, provided the platforms they use look professional. Alternatively, complacency may lead to flagging public interest in AutoML, and the technology could end up being perceived as a gimmick. All that is clear is that the lessons learned and decisions made during this initial contact between AutoML technology and industry will set the tone for future societal engagement.

REFERENCES

- [1] 2021. Aible [Commercial Software]. Retrieved from <https://www.aible.com>.
- [2] 2021. Algolytics [Commercial Software]. Retrieved from <https://algolytics.com/products/abm/>.
- [3] 2021. AlternativeTo. Retrieved December 5, 2021 from <https://alternativeto.net>
- [4] 2021. Alteryx [Commercial Software]. Retrieved from <https://www.alteryx.com/analytics-process-automation>.
- [5] 2021. Auger [Commercial Software]. Retrieved from <https://auger.ai>.
- [6] 2021. B2Metric [Commercial Software]. Retrieved from <https://b2metric.com>.
- [7] 2021. Big Squid [Commercial Software]. Retrieved from <https://bigsquid.ai/kraken>.
- [8] 2021. BigML [Commercial Software]. Retrieved from <https://bigml.com>.
- [9] 2021. Cloud AutoML (Google) [Commercial Software]. Retrieved from <https://cloud.google.com/automl>.
- [10] 2021. cnvrg.Io [Commercial Software]. Retrieved from <https://cnvrg.io>.
- [11] 2021. Comet [Commercial Software]. Retrieved from <https://www.comet.ml/site>.
- [12] 2021. Compellon [Commercial Software]. Retrieved from <https://www.compellon.com>.
- [13] 2021. D2iQ (Kaptain) [Commercial Software]. Retrieved from <https://d2iq.com/products/kaptain>.
- [14] 2021. Databricks [Commercial Software]. Retrieved from <https://databricks.com/product/automl>.
- [15] 2021. Dataiku [Commercial Software]. Retrieved from <https://www.dataiku.com>.
- [16] 2021. Datarobot Alternatives. Retrieved December 5, 2021 from <https://www.gartner.com/reviews/market/data-science-machine-learning-platforms/vendor/datarobot/alternatives>
- [17] 2021. DataRobot [Commercial Software]. Retrieved from <https://www.datarobot.com>.
- [18] 2021. Deep Cognition [Commercial Software]. Retrieved from <https://deepcognition.ai>.
- [19] 2021. Determined AI [Commercial Software]. Retrieved from <https://www.determined.ai>.

- [20] 2021. DMway [Commercial Software]. Retrieved from <http://dmway.com/>.
- [21] 2021. Domino [Commercial Software]. Retrieved from <https://www.dominodatalab.com/product/domino-data-science-platform/>.
- [22] 2021. dotData [Commercial Software]. Retrieved from <https://dotdata.com>.
- [23] 2021. Einblick [Commercial Software]. Retrieved from <https://einblick.ai/product/>.
- [24] 2021. Hazy [Commercial Software]. Retrieved from <https://hazy.com>.
- [25] 2021. Iguazio [Commercial Software]. Retrieved from <https://www.iguazio.com>.
- [26] 2021. KDnuggets. <https://www.kdnuggets.com>
- [27] 2021. KNIME [Commercial Software]. Retrieved from <https://www.knime.com>.
- [28] 2021. Kortical [Commercial Software]. Retrieved from <https://kortical.com>.
- [29] 2021. Microsoft Azure Automl [Commercial Software]. Retrieved from <https://azure.microsoft.com/en-us/services/machine-learning/automatedml>.
- [30] 2021. MOSTLY AI [Commercial Software]. Retrieved from <https://mostly.ai>.
- [31] 2021. MyDataModels [Commercial Software]. Retrieved from <https://www.mydatamodels.com>.
- [32] 2021. Neptune.Ai [Commercial Software]. Retrieved from [https://neptune.ai/](https://neptune.ai).
- [33] 2021. neuralstudio.Ai [Commercial Software]. Retrieved from <https://neuralstudio.ai>.
- [34] 2021. Number Theory [Commercial Software]. Retrieved from <http://numbertheory.ai>.
- [35] 2021. OptiScorer [Commercial Software]. Retrieved from <https://optiscorer.com>.
- [36] 2021. Pecan [Commercial Software]. Retrieved from <https://www.pecan.ai>.
- [37] 2021. Revision.io [Commercial Software]. Retrieved from <https://revision.io>.
- [38] 2021. Qubole [Commercial Software]. Retrieved from <https://www.qubole.com/platform/open-data-lake-platform/machine-learning/>.
- [39] 2021. RapidMiner [Commercial Software]. Retrieved from <https://rapidminer.com>.
- [40] 2021. SageMaker (AWS) [Commercial Software]. Retrieved from <https://aws.amazon.com/sagemaker/>.
- [41] 2021. Sklearn documentation - List of all Estimators. Retrieved May 28, 2021 from https://scikit-learn.org/stable/modules/generated/sklearn.utils.all_estimators.html#sklearn-utils-all-estimators
- [42] 2021. SparkCognition [Commercial Software]. Retrieved from <https://www.sparkcognition.com/products/darwin/>.
- [43] 2021. Spell [Commercial Software]. Retrieved from <https://spell.ml/platform>.
- [44] 2021. TAZI [Commercial Software]. Retrieved from [https://www.tazi.ai/](https://www.tazi.ai).
- [45] 2021. TIMi [Commercial Software]. Retrieved from [https://timi.eu/timi/timi-modeler/](https://timi.eu/timi/timi-modeler).
- [46] 2021. Top 10 DataRobot Alternatives & Competitors. Retrieved December 5, 2021 from <https://www.g2.com/products/datarobot/competitors/alternatives>
- [47] 2021. Viya (SAS) [Commercial Software]. Retrieved from https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html.
- [48] 2021. Watson Studio (IBM) [Commercial Software]. Retrieved from <https://www.ibm.com/au-en/cloud/watson-studio/autoai>.
- [49] 2021. Xpanse [Commercial Software]. Retrieved from <https://xpanse.ai>.
- [50] Md. Abdullah-Al-Kafi, Israt Jahan Tasnova, Md. Wadud Islam, and Sumit Kumar Banshal. 2022. Performances of Different Approaches for Fake News Classification: An Analytical Study. In *Advanced Network Technologies and Intelligent Computing (Communications in Computer and Information Science)*, Isaac Woungang, Sanjay Kumar Dhurandher, Kiran Kumar Pattanaik, Anshul Verma, and Pradeepika Verma (Eds.). Springer International Publishing, Cham, 700–714. https://doi.org/10.1007/978-3-030-96040-7_53
- [51] Julius Adebayo. 2021. Fairml [Computer Software]. Retrieved from <https://github.com/adebayoj/fairml>.
- [52] Adesola Adegbeye and Michael Kampouridis. 2021. Machine Learning Classification and Regression Models for Predicting Directional Changes Trend Reversal in FX Markets. *Expert Systems with Applications* 173 (July 2021), 114645. <https://doi.org/10.1016/j.eswa.2021.114645>
- [53] Adesola Adegbeye, Michael Kampouridis, and Fernando Otero. 2021. Improving Trend Reversal Estimation in Forex Markets under a Directional Changes Paradigm with Classification Algorithms. *International Journal of Intelligent Systems* 36, 12 (2021), 7609–7640. <https://doi.org/10.1002/int.22601>
- [54] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. PMLR, 60–69.
- [55] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.
- [56] Farshad Ahmadighohandizi and Kari Systä. 2018. Application Development and Deployment for IoT Devices. In *Advances in Service-Oriented and Cloud Computing (Communications in Computer and Information Science)*, Alexander Lazovik and Stefan Schulte (Eds.). Springer International Publishing, Cham, 74–85. https://doi.org/10.1007/978-3-319-72125-5_6

- [57] Sajid Ahmed, Rafsanjani Muhammad, Zahid Hossain Khan, Sheikh Adilina, Alok Sharma, Swakkhar Shatabda, and Abdollah Dehzangi. 2021. ACP-MHCNN: An Accurate Multi-Headed Deep-Convolutional Neural Network to Predict Anticancer Peptides. *Scientific Reports* 11, 1 (Dec. 2021), 23676. <https://doi.org/10.1038/s41598-021-02703-3>
- [58] Shaojie Ai, Jia Song, and Guobiao Cai. 2021. A Real-Time Fault Diagnosis Method for Hypersonic Air Vehicle with Sensor Fault Based on the Auto Temporal Convolutional Network. *Aerospace Science and Technology* 119 (Dec. 2021), 107220. <https://doi.org/10.1016/j.ast.2021.107220>
- [59] Aible. 2021. True AutoML. Retrieved September 29, 2021 from <https://www.aible.com/blog/true-automl>
- [60] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2623–2631.
- [61] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2021. Optuna [Computer Software]. Retrieved from <https://github.com/optuna/optuna>.
- [62] Ahmed Alaa and Mihaela Schaar. 2018. AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning. In *International Conference on Machine Learning*. 139–148.
- [63] Ahmed M. Alaa. 2021. AutoPrognosis [Computer Software]. Retrieved from <https://github.com/ahmedmalaa/AutoPrognosis>.
- [64] Abbas Raza Ali, Marcin Budka, and Bogdan Gabrys. 2015. A Review of Meta-Level Learning in the Context of Multi-Component, Multi-Level Evolving Prediction Systems. *arXiv preprint arXiv:2007.10818* (2015).
- [65] Abbas Raza Ali, Bogdan Gabrys, and Marcin Budka. 2018. Cross-Domain Meta-Learning for Time-Series Forecasting. 126 (2018), 9–18. <https://doi.org/10.1016/j.procs.2018.07.204>
- [66] Sarah Alnaghmish, Naja Alrashed, Faisal Aleissa, Shahad Althobaiti, Dongyu Liu, Mansour Alsaleh, and Kalyan Veeramachaneni. 2020. Cardea: An Open Automated Machine Learning Framework for Electronic Health Records. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 536–545.
- [67] Fernando Alonso-Martín, Juan José Gamboa-Montero, José Carlos Castillo, Álvaro Castro-González, and Miguel Ángel Salichs. 2017. Detecting and Classifying Human Touches in a Social Robot Through Acoustic Sensing and Machine Learning. *Sensors* 17, 5 (May 2017), 1138. <https://doi.org/10.3390/s17051138>
- [68] Alteryx. 2021. Compose [Computer Software]. Retrieved from <https://github.com/alteryx/compose>.
- [69] Alteryx. 2021. EvalML [Computer Software]. Retrieved from <https://github.com/alteryx/evalml>.
- [70] Alteryx. 2021. Featuretools [Computer Software]. Retrieved from <https://github.com/alteryx/featuretools>.
- [71] Amazon Web Services. 2021. Adatune [Computer Software]. Retrieved from <https://github.com/awslabs/adatune>.
- [72] Amazon Web Services. 2021. AutoGluon [Computer Software]. Retrieved from <https://github.com/awslabs/autogluon>.
- [73] Ruhul Amin, Chowdhury Rafeed Rahman, Sajid Ahmed, Md Habibur Rahman Sifat, Md Nazmul Khan Liton, Md Moshiur Rahman, Md Zahid Hossain Khan, and Swakkhar Shatabda. 2020. iPromoter-BnCNN: A Novel Branched CNN-based Predictor for Identifying and Classifying Sigma Promoters. *Bioinformatics* 36, 19 (Dec. 2020), 4869–4875. <https://doi.org/10.1093/bioinformatics/btaa609>
- [74] Anaconda. 2020. *The State of Data Science 2020*. Technical Report. <https://www.anaconda.com/state-of-data-science-2020>
- [75] AnalyticsVidhya. 2021. AnalyticsVidhya blog posts tagged with 'automl'. Retrieved December 5, 2021 from <https://www.analyticsvidhya.com/blog/tag/automl/>
- [76] Alec W. (Alec Wayne) Anderson. 2017. *Deep Mining : Scaling Bayesian Auto-Tuning of Data Science Pipelines*. Thesis. Massachusetts Institute of Technology.
- [77] Juan S. Angarita-Zapata, Gina Maestre-Gongora, and Jenny Fajardo Calderín. 2021. A Bibliometric Analysis and Benchmark of Machine Learning and AutoML in Crash Severity Prediction: The Case Study of Three Colombian Cities. *Sensors* 21, 24 (Jan. 2021), 8401. <https://doi.org/10.3390/s21248401>
- [78] Juan S. Angarita-Zapata, Antonio D. Masegosa, and Isaac Triguero. 2020. General-Purpose Automated Machine Learning for Transportation: A Case Study of Auto-sklearn for Traffic Forecasting. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems (Communications in Computer and Information Science)*, Marie-Jeanne Lesot, Susana Vieira, Marek Z. Reformat, João Paulo Carvalho, Anna Wilbik, Bernadette Bouchon-Meunier, and Ronald R. Yager (Eds.). Springer International Publishing, Cham, 728–744. https://doi.org/10.1007/978-3-030-50143-3_57
- [79] Juan S. Angarita-Zapata, Isaac Triguero, and Antonio D. Masegosa. 2018. A Preliminary Study on Automatic Algorithm Selection for Short-Term Traffic Forecasting. In *Intelligent Distributed Computing XII (Studies in Computational Intelligence)*, Javier Del Ser, Eneko Osaba, Miren Nekane Bilbao, Javier J. Sanchez-Medina, Massimo Vecchio, and Xin-She Yang (Eds.). Springer International Publishing, Cham, 204–214. https://doi.org/10.1007/978-3-319-99626-4_18
- [80] Talha Anwar and Hassan Anwar. 2021. Beef Quality Assessment Using AutoML. In *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. 1–4. <https://doi.org/10.1109/MAJICC53071.2021.9526256>

- [81] Information Governance ANZ. 2019. *IG INDUSTRY SURVEY July 2019 Report*. Technical Report. <https://www.infogovanz.com/wp-content/uploads/2020/01/IGANZ2019ReportFinal.pdf>
- [82] Information Governance ANZ. 2021. *IG INDUSTRY REPORT May 2021*. Technical Report. https://www.infogovanz.com/wp-content/uploads/2021/05/InfoGov_IndustrySurvey_MAY2021.pdf
- [83] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv preprint arXiv:1909.03012* (2019).
- [84] Data To AI Lab at The Massachusetts Institute of Technology. 2021. BTB [Computer Software]. Retrieved from <https://github.com/MLBazaar/BTB>.
- [85] Data To AI Lab at The Massachusetts Institute of Technology. 2021. Cardea [Computer Software]. Retrieved from <https://github.com/MLBazaar/Cardea>.
- [86] The Hachmann Lab at The University at Buffalo. 2021. ChemML [Computer Software]. Retrieved from <https://github.com/hachmannlab/chemml>.
- [87] Auger. 2021. How Does Auger.AI Provide Such Fast And Accurate AutoML? – Auger. Retrieved December 14, 2021 from <https://auger.ai/how-does-auger-ai-provide-such-fast-and-accurate-automl/>
- [88] Standards Australia. 2020. *An Artificial Intelligence Standards Roadmap: Making Australia's Voice Heard*. Technical Report. https://www.standards.org.au/getmedia/ede81912-55a2-4d8e-849f-9844993c3b9d/R_1515-An-Artificial-Intelligence-Standards-Roadmap-soft.pdf.aspx
- [89] AutoML Groups Freiburg and Hannover. 2015. RoBO [Computer Software]. Retrieved from <https://github.com/automl/RoBO>.
- [90] AutoML Groups Freiburg and Hannover. 2016. SMAC3 [Computer Software]. Retrieved from <https://github.com/automl/SMAC3>.
- [91] AutoML Groups Freiburg and Hannover. 2017. HpBandSter [Computer Software]. Retrieved from <https://github.com/automl/HpBandSter>.
- [92] AutoML Groups Freiburg and Hannover. 2021. Auto-PyTorch [Computer Software]. Retrieved from <https://github.com/automl/Auto-PyTorch>.
- [93] AutoML Groups Freiburg and Hannover. 2021. auto-sklearn [Computer Software]. Retrieved from <https://github.com/automl/auto-sklearn>.
- [94] AutoML Groups Freiburg and Hannover. 2021. Pyautoweka [Computer Software]. Retrieved from <https://github.com/automl/pyautoweka>.
- [95] AWS. 2021. Getting Started with Amazon SageMaker - Amazon Web Services. Retrieved December 14, 2021 from <https://aws.amazon.com/sagemaker/getting-started/>
- [96] AWS. 2021. Sagemaker Documentation: Notebooks Sharing. Retrieved December 14, 2021 from <https://docs.aws.amazon.com/sagemaker/latest/dg/notebooks-sharing.html>
- [97] AWS. 2021. What Is DevOps? - Amazon Web Services (AWS). Retrieved May 28, 2021 from <https://aws.amazon.com/devops/what-is-devops/>
- [98] Nidhal Baccouri. 2020. Igel [Computer Software]. Retrieved from <https://github.com/nidhaloff/igel>.
- [99] Yang Bai, Yang Li, Yu Shen, Mingyu Yang, Wentao Zhang, and Bin Cui. 2022. AutoDC: An Automatic Machine Learning Framework for Disease Classification. *Bioinformatics* 38, 13 (July 2022), 3415–3421. <https://doi.org/10.1093/bioinformatics/btac334>
- [100] Bowen Baker. 2021. MetaQNN [Computer Software]. Retrieved from <https://github.com/bowenbaker/metaqnn>.
- [101] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. 2017. Designing Neural Network Architectures Using Reinforcement Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=S1c2cvqee>
- [102] Rashid Bakirov, Damien Fay, and Bogdan Gabrys. 2021. Automated adaptation strategies for stream learning. *Machine Learning* (2021), 1–34.
- [103] Rashid Bakirov, Bogdan Gabrys, and Damien Fay. 2015. On Sequences of Different Adaptive Mechanisms in Non-Stationary Regression Problems. In *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE. <https://doi.org/10.1109/ijcnn.2015.7280779>
- [104] Rashid Bakirov, Bogdan Gabrys, and Damien Fay. 2017. Multiple Adaptive Mechanisms for Data-Driven Soft Sensors. 96 (2017), 42–54. <https://doi.org/10.1016/j.compchemeng.2016.08.017>
- [105] Rashid Bakirov, Bogdan Gabrys, and Damien Fay. 2018. Generic Adaptation Strategies for Automated Machine Learning. *arXiv preprint arXiv:1812.10793v2* (2018).
- [106] Maximilian Balandat, Brian Karrer, Daniel R Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. 2019. Botorch: Programmable Bayesian Optimization in Pytorch. *arXiv preprint arXiv:1910.06403*

- (2019).
- [107] Ali Haider Bangash. 2020. Leveraging AutoML to Provide NAFLD Screening Diagnosis: Proposed Machine Learning Models. *medRxiv preprint medRxiv:2020.10.20.20216291v1* (2020).
- [108] Ali Haider Bangash, Ali Haider Shah, Arshiya Fatima, Saiqa Zehra, Syed Mohammad Mehmood Abbas, Hashir Fahim Khawaja, Muhammad Ashraf, and Adil Baloch. 2021. Amalgamation of Auto Machine Learning and Ensemble Approaches to Achieve State-of-the-Art Post-Heart Failure Survival Predictions. *American Heart Journal* 242 (Dec. 2021), 153. <https://doi.org/10.1016/j.ahj.2021.10.021>
- [109] Lisa Barbadora. 2021. Aktana Acquires Ople.AI's Machine Learning Automation Technology. Retrieved December 7, 2021 from <https://www.globenewswire.com/en/news-release/2021/10/27/2321826/0/en/Aktana-Acquires-Ople-AI-s-Machine-Learning-Automation-Technology.html>
- [110] Enrique Barreiro, Cristian R. Munteanu, Maykel Cruz-Monteagudo, Alejandro Pazos, and Humbert González-Díaz. 2018. Net-Net Auto Machine Learning (AutoML) Prediction of Complex Ecosystems. 8, 1 (2018). <https://doi.org/10.1038/s41598-018-30637-w>
- [111] Márcio P Basgalupp, Rodrigo C Barros, Alex GC de Sá, Gisele L Pappa, Rafael G Mantovani, ACPLF de Carvalho, and Alex A Freitas. 2020. An Extensive Experimental Evaluation of Automated Machine Learning Methods for Recommending Classification Algorithms. *Evolutionary Intelligence* (2020), 1–20.
- [112] Saša Baškarada and Andy Koronios. 2017. Unicorn Data Scientist: The Rarest of Breeds. *Program: electronic library and information systems* 51, 1 (2017), 65–74.
- [113] Markus Bauer, Clemens van Dinther, and Daniel Kiefer. 2020. Machine learning in SME: An empirical study on enablers and success factors. In *AMCIS 2020 Virtual Conference, August 10–14, 2020*.
- [114] Filipe Baumeister, Marcelo Werneck Barbosa, and Rodrigo Richard Gomes. 2020. What Is Required to Be a Data Scientist?: Analyzing Job Descriptions With Centering Resonance Analysis. *International Journal of Human Capital and Information Technology Professionals (IJHCITP)* 11, 4 (2020), 21–40. <https://doi.org/10.4018/IJHCITP.2020100102>
- [115] Philipp E. Bayer, Jakob Petereit, Monica Furaste Danilevici, Robyn Anderson, Jacqueline Batley, and David Edwards. 2021. The Application of Pangenomics and Machine Learning in Genomic Selection in Plants. *The Plant Genome* 14, 3 (2021), e20112. <https://doi.org/10.1002/tpg2.20112>
- [116] Andrew L. Beam, Arjun K. Manrai, and Marzyeh Ghassemi. 2020. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA* 323, 4 (Jan. 2020), 305–306. <https://doi.org/10.1001/jama.2019.20866>
- [117] James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *International Conference on Machine Learning*. PMLR, 115–123.
- [118] Yamins D. Cox D. D. Bergstra, J. 2021. Hyperopt [Computer Software]. Retrieved from <https://github.com/hyperopt/hyperopt>.
- [119] Markus Bertl, Peeter Ross, and Dirk Draheim. 2021. Predicting Psychiatric Diseases Using AutoAI: A Performance Analysis Based on Health Insurance Billing Data. In *Database and Expert Systems Applications (Lecture Notes in Computer Science)*, Christine Strauss, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil (Eds.). Springer International Publishing, Cham, 104–111. https://doi.org/10.1007/978-3-030-86472-9_9
- [120] Besim Bilalli, Alberto Abelló, Tomàs Aluja-Banet, Rana Faisal Munir, and Robert Wrembel. 2018. PRESANT: data pre-processing assistant. In *International Conference on Advanced Information Systems Engineering*. Springer International Publishing, 57–65.
- [121] Martin Binder, Bernd Bischl, and Alexander Hanf. 2020. mlr3automl [Computer Software]. Retrieved from <https://github.com/a-hanf/mlr3automl>.
- [122] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).
- [123] Bernd Bischl, Jakob Richter, Jakob Bossek, Daniel Horn, Janek Thomas, and Michel Lang. 2018. mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. *arXiv preprint arXiv:1703.03373* (2018).
- [124] Rex Black. 2002. *Managing the Testing Process*. John Wiley & Sons.
- [125] Brad Boehmke and Brandon Greenwell. 2019. *Hands-On Machine Learning with R*. Chapman and Hall/CRC, Boca Raton. <https://doi.org/10.1201/9780367816377>
- [126] Raymond Bond, Ansgar Koene, Alan Dix, Jennifer Boger, Maurice D. Mulvenna, Mykola Galushka, Bethany Waterhouse Bradley, Fiona Browne, Hui Wang, and Alexander Wong. 2019. Democratisation of Usable Machine Learning in Computer Vision. *arXiv preprint arXiv:1902.06804v1* (2019).
- [127] Nigel Bosch. 2021. AutoML Feature Engineering for Student Modeling Yields High Accuracy, but Limited Interpretability. *Journal of Educational Data Mining* 13, 2 (Aug. 2021), 55–79. <https://doi.org/10.5281/zenodo.5275314>
- [128] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

- [129] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D. Sculley. 2017. The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. In *2017 IEEE International Conference on Big Data (Big Data)*. 1123–1132. <https://doi.org/10.1109/BigData.2017.8258038>
- [130] John S. Breese, David Heckerman, and Carl Kadie. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 43–52.
- [131] Lawrence R. Brindise and Robert J. Steele. 2018. Machine Learning-based Pre-discharge Prediction of Hospital Readmission. In *2018 International Conference on Computer, Information and Telecommunication Systems (CITS)*. 1–5. <https://doi.org/10.1109/CITS.2018.8440171>
- [132] Dmitry S. Bulgarevich, Susumu Tsukamoto, Tadashi Kasuya, Masahiko Demura, and Makoto Watanabe. 2019. Automatic Steel Labeling on Certain Microstructural Constituents with Image Processing and Machine Learning Tools. *Science and Technology of Advanced Materials* 20, 1 (Dec. 2019), 532–542. <https://doi.org/10.1080/14686996.2019.1610668>
- [133] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf>
- [134] Israel Campero Jurado, Andrejs Fedjajevs, Joaquin Vanschoren, and Aarnout Brombacher. 2022. Interpretable Assessment of ST-Segment Deviation in ECG Time Series. *Sensors* 22, 13 (Jan. 2022), 4919. <https://doi.org/10.3390/s22134919>
- [135] Capgemini. 2020. *AI and the Ethical Conundrum*. Technical Report. <https://www.capgemini.com/wp-content/uploads/2020/10/AI-and-the-Ethical-Conundrum-Report.pdf>
- [136] Clay Carper, Aaron McClellan, and Craig C. Douglas. 2021. I-80 Closures: An Autonomous Machine Learning Approach. In *Computational Science – ICCS 2021 (Lecture Notes in Computer Science)*, Maciej Paszynski, Dieter Kranzlmüller, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M.A. Sloot (Eds.). Springer International Publishing, Cham, 286–291. https://doi.org/10.1007/978-3-030-77977-1_22
- [137] Dylan Cashman, Shah Rukh Humayoun, Florian Heimerl, Kendall Park, Subhajit Das, John Thompson, Bahador Saket, Abigail Mosca, John Stasko, Alex Endert, Michael Gleicher, and Remco Chang. 2019. A User-Based Visual Analytics Workflow for Exploratory Model Analysis. *Computer Graphics Forum* 38, 3 (2019), 185–199. <https://doi.org/10.1111/cgf.13681> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13681>
- [138] José Carlos Castillo, Fernando Alonso-Martín, David Cáceres-Domínguez, María Malfaz, and Miguel A. Salichs. 2019. The Influence of Speed and Position in Dynamic Gesture Recognition for Human-Robot Interaction. *Journal of Sensors* 2019 (Feb. 2019), e7060491. <https://doi.org/10.1155/2019/7060491>
- [139] Mayra Ruiz Castro, Beatrice Van der Heijden, and Emma L Henderson. 2020. Catalysts in Career Transitions: Academic Researchers Transitioning into Sustainable Careers in Data Science. *Journal of Vocational Behavior* 122 (2020), 103479.
- [140] Casey G. Cegielski and L. Allison Jones-Farmer. 2016. Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study. *Decision Sciences Journal of Innovative Education* 14, 1 (Jan. 2016), 91–118. <https://doi.org/10.1111/dsji.12086>
- [141] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.
- [142] Sayan Chakraborty, Smit Shah, Kiumars Soltani, Anna Swigart, Luyao Yang, and Kyle Buckingham. 2020. Building an Automated and Self-Aware Anomaly Detection System. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 1465–1475.
- [143] Simon Chan, Philip Treleaven, and Licia Capra. 2013. Continuous Hyperparameter Optimization for Large-Scale Recommender Systems. In *2013 IEEE International Conference on Big Data*. IEEE. <https://doi.org/10.1109/bigdata.2013.6691595>
- [144] Arunima Chaudhary, Alayt Issak, Kiran Kate, Yannis Katsis, Abel Valente, Dakuo Wang, Alexandre Evfimievski, Sairam Gurajada, Ban Kawas, Cristiano Malossi, Lucian Popa, Tejaswini Pedapati, Horst Samulowitz, Martin Wistuba, and Yunyao Li. 2021. AutoText: An End-to-End AutoAI Framework for Text. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 18 (May 2021), 16001–16003.
- [145] Dihao Chen. 2021. Advisor [Computer Software]. Retrieved from <https://github.com/tobegin3hub/advisor>.
- [146] Jiaoyan Chen, Freddy Lécué, Jeff Z. Pan, Shumin Deng, and Huajun Chen. 2021. Knowledge Graph Embeddings for Dealing with Concept Drift in Machine Learning. *Journal of Web Semantics* 67 (Feb. 2021), 100625. <https://doi.org/10.1016/j.websem.2020.100625>
- [147] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*. 785–794.

- [148] Yiren Chen, Yaming Yang, Hong Sun, Yujing Wang, Yu Xu, Wei Shen, Rong Zhou, Yunhai Tong, Jing Bai, and Ruofei Zhang. 2020. AutoADR: Automatic Model Design for Ad Relevance. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2365–2372.
- [149] Zhenshuo Chen, Eoin Brophy, and Tomas Ward. 2021. Malware Classification Using Static Disassembly and Machine Learning. *arXiv preprint arXiv:2201.07649* (2021).
- [150] Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zytek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni. 2022. VBridge: Connecting the Dots Between Features and Data to Explain Healthcare Models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 378–388. <https://doi.org/10.1109/TVCG.2021.3114836> arXiv:cs/2108.02550
- [151] Krzysztof Chomiak and Michal Miklus. 2021. Harnessing Value from Data Science in Business: Ensuring Explainability and Fairness of Solutions. *arXiv preprint arXiv:2108.07714* (2021).
- [152] Rishabh Choudhary and Hemant Kumar Gianey. 2017. Comprehensive Review On Supervised Machine Learning Algorithms. In *2017 International Conference on Machine Learning and Data Science (MLDS)*. 37–43. <https://doi.org/10.1109/MLDS.2017.11>
- [153] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. 2018. Time Series FeatuRe Extraction on Basis of Scalable Hypothesis Tests (Tsfresh – A Python Package). *Neurocomputing* 307 (Sept. 2018), 72–77. <https://doi.org/10.1016/j.neucom.2018.03.067>
- [154] Kartik Chugh. 2020. Otto [Computer Software]. Retrieved from <https://github.com/KartikChugh/Otto>.
- [155] Rob Cimperman. 2006. *UAT Defined: A Guide to Practical User Acceptance Testing (Digital Short Cut)*. Pearson Education.
- [156] Cisco AI. 2021. Amla [Computer Software]. Retrieved from <https://github.com/CiscoAI/amlia>.
- [157] Marc Claesen. 2021. Optunity [Computer Software]. Retrieved from <https://github.com/claesennm/optunity>.
- [158] CleverInsight. 2020. Cognito [Computer Software]. Retrieved from <https://github.com/CleverInsight/cognito>.
- [159] William G. Cochran. 1954. Some Methods for Strengthening the Common X^2 Tests. *Biometrics* 10, 4 (1954), 417–451. <https://doi.org/10.2307/3001616>
- [160] COIN-OR Foundation. 2021. Rbfopt [Computer Software]. Retrieved from <https://github.com/coin-or/rbfopt>.
- [161] The European Commission. 2021. *Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Technical Report. <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence>
- [162] Complex Adaptive Systems (CAS) Lab - University of Technology Sydney. 2022. AutoWeka4MCPS [Computer Software]. Retrieved from <https://github.com/UTS-CASLab/autoweka>.
- [163] Fabricio Magalhães Cordeiro, Gutemberg Borges França, Francisco Leite de Albuquerque Neto, and Ismail Gultepe. 2021. Visibility and Ceiling Nowcasting Using Artificial Intelligence Techniques for Aviation Applications. *Atmosphere* 12, 12 (Dec. 2021), 1657. <https://doi.org/10.3390/atmos12121657>
- [164] Ricardo Corral-Corral, Jesús A. Beltrán, Carlos A. Brizuela, and Gabriel Del Rio. 2017. Systematic Identification of Machine-Learning Models Aimed to Classify Critical Residues for Protein Function from Protein Structure. *Molecules* 22, 10 (Oct. 2017), 1673. <https://doi.org/10.3390/molecules22101673>
- [165] Alberto Costa and Giacomo Nannicini. 2018. RBFOpt: An Open-Source Library for Black-Box Optimization with Costly Function Evaluations. *Mathematical Programming Computation* 10, 4 (Dec. 2018), 597–629. <https://doi.org/10.1007/s12532-018-0144-7>
- [166] Carlos Costa and Maribel Yasmina Santos. 2017. The Data Scientist Profile and Its Representativeness in the European E-Competence Framework and the Skills Framework for the Information Age. *International Journal of Information Management* 37, 6 (Dec. 2017), 726–734. <https://doi.org/10.1016/j.ijinfomgt.2017.07.010>
- [167] Daniel Crankshaw, Gur-Eyal Sela, Corey Zumar, Xiangxi Mo, Joseph E. Gonzalez, Ion Stoica, and Alexey Tumanov. 2018. InferLine: ML Inference Pipeline Composition Framework. *arXiv preprint arXiv:1812.01776v1* (2018).
- [168] CrowdFlower. 2016. *2016 Data Science Report*. Technical Report. https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf
- [169] Wojciech M. Czarnecki, Sabina Podlewska, and Andrzej J. Bojarski. 2015. Robust Optimization of SVM Hyperparameters in the Classification of Bioactive Compounds. *Journal of Cheminformatics* 7, 1 (Aug. 2015), 38. <https://doi.org/10.1186/s13321-015-0088-0>
- [170] Vincent D. Warmerdam. 2021. Scikit-Fairness [Computer Software]. Retrieved from <https://github.com/koaning/scikit-fairness>.
- [171] Luiz Antonio da Ponte Junior, Débora Christina Muchaluat Saade, Alexandre Plastino de Carvalho, Rita de Cássia Alves, Liana Catarina Lima Portugal, Letícia de Oliveira, and Mirtes Garcia Pereira. 2020. Identifying Post-Traumatic Stress Symptoms Using Physiological Signals and Data Mining. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 233–238. <https://doi.org/10.1109/CBMS49503.2020.00051>

- [172] Jessica Dafflon, Walter H. L. Pinaya, Federico Turkheimer, James H. Cole, Robert Leech, Mathew A. Harris, Simon R. Cox, Heather C. Whalley, Andrew M. McIntosh, and Peter J. Hellyer. 2019. Analysis of an Automated Machine Learning Approach in Brain Predictive Modelling: A Data-Driven Approach to Predict Brain Age from Cortical Anatomical Measures. *arXiv preprint arXiv:1910.03349* (2019).
- [173] Jessica Dafflon, Walter H. L. Pinaya, Federico Turkheimer, James H. Cole, Robert Leech, Mathew A. Harris, Simon R. Cox, Heather C. Whalley, Andrew M. McIntosh, and Peter J. Hellyer. 2020. An Automated Machine Learning Approach to Predict Brain Age from Cortical Anatomical Measures. *Human Brain Mapping* 41, 13 (2020), 3555–3566. <https://doi.org/10.1002/hbm.25028>
- [174] Zehua Dai, Li Wang, and Shanshui Yang. 2020. Data Mining-Based Model Simplification and Optimization of an Electrical Power Generation System. *IEEE Transactions on Transportation Electrification* 6, 4 (Dec. 2020), 1665–1678. <https://doi.org/10.1109/TTE.2020.2995745>
- [175] Brian d’Alessandro, Cathy O’Neil, and Tom LaGatta. 2017. Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification. *Big Data* 5, 2 (June 2017), 120–134. <https://doi.org/10.1089/big.2016.0048>
- [176] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F Ilyas, Mourad Ouzzani, and Nan Tang. 2013. NADEEF: A Commodity Data Cleaning System. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. 541–552.
- [177] Data Analytics - Qatar Computing Research Institute, HBKU. 2021. NADEEF [Computer Software]. Retrieved from <https://github.com/daqcri/NADEEF>.
- [178] Data Analytics Lab at Texas A&M University. 2020. TODS [Computer Software]. Retrieved from <https://github.com/datamllab/tods>.
- [179] Data Science for Social Good. 2021. Aequitas [Computer Software]. Retrieved from <https://github.com/dssg/aequitas>.
- [180] Databricks. 2021. Collaborative Notebooks - Databricks. Retrieved December 1, 2021 from <https://databricks.com/product/collaborative-notebooks>
- [181] DataCanvas. 2020. HyperGBM [Computer Software]. Retrieved from <https://github.com/DataCanvasIO/HyperGBM>.
- [182] DataQuest. 2021. Data Analyst Skills – 8 Skills You Need to Get a Job. Retrieved May 28, 2021 from <https://www.dataquest.io/blog/data-analyst-skills/>
- [183] DataRobot. 2021. What Is MLOps? Retrieved May 28, 2021 from <https://www.datarobot.com/lp/what-is-mlops/>
- [184] Ben Dattner, Tomas Chamorro-Premuzic, Richard Buchband, and Lucinda Schettler. 2019. The Legal and Ethical Implications of Using AI in Hiring. *Harvard Business Review* 25 (2019).
- [185] Thomas H Davenport and DJ Patil. 2012. Data Scientist. *Harvard business review* 90, 5 (2012), 70–76.
- [186] Laidy De Armas Jacomino, Miguel Angel Medina-Pérez, Raúl Monroy, Danilo Valdes-Ramirez, Carlos Morell-Pérez, and Rafael Bello. 2021. Dwell Time Estimation of Import Containers as an Ordinal Regression Problem. *Applied Sciences* 11, 20 (Jan. 2021), 9380. <https://doi.org/10.3390/app11209380>
- [187] Alfonso de la Vega, Diego García-Saiz, Marta Zorrilla, and Pablo Sánchez. 2019. How Far Are We from Data Mining Democratisation? A Systematic Review. *arXiv preprint arXiv:1903.08431v1* (2019).
- [188] Andrea De Mauro, Marco Greco, Michele Grimaldi, and Paavo Ritala. 2018. Human Resources for Big Data Professions: A Systematic Classification of Job Roles and Required Skill Sets. *Information Processing & Management* 54, 5 (Sept. 2018), 807–817. <https://doi.org/10.1016/j.ipm.2017.05.004>
- [189] Axel de Romblay. 2017. MLBox [Computer Software]. Retrieved from <https://github.com/AxeldeRomblay/MLBox>.
- [190] Alex G. C. de Sá, Cristiano G. Pimenta, Gisele L. Pappa, and Alex A. Freitas. 2020. A Robust Experimental Evaluation of Automated Multi-Label Classification Methods. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference (GECCO ’20)*. Association for Computing Machinery, New York, NY, USA, 175–183. <https://doi.org/10.1145/3377930.3390231>
- [191] Richard D De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S Baumer, Andrew Bray, Thomas C Bressoud, Lance Bryant, Lei Z Cheng, Amanda Francis, Robert Gould, Albert Y. Kim Kim, Matt12 Kretchmar, Qin Lu, Ann Moskol, Deborah Nolan, Roberto Pelayo, Sean Raleigh, Ricky J. Sethi, Mutiara Sondjaja, Neelesh Tiruviluamala, Paul X. Uhlig, Talitha M. Washington, Curtis L. Wesley, David White, and Ping Ye. 2017. Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application* 4 (2017), 15–30.
- [192] James Dean, Matthias Scheffler, Thomas A. R. Purcell, Sergey V. Barabash, Rahul Bhowmik, and Timur Bazhirov. 2021. Interpretable Machine Learning for Materials Design. *arXiv preprint arXiv:2112.00239* (2021).
- [193] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. 2017. The Data Civilizer System. In *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*. www.cidrdb.org. <http://cidrdb.org/cidr2017/papers/p44-deng-cidr17.pdf>
- [194] Angel Díaz-Pacheco, Carlos A. Reyes-García, and Vanesa Chicatto-Gasperín. 2021. Granule-Based Fuzzy Rules to Assist in the Infant-Crying Pattern Recognition Problem. *Sādhāraṇā* 46, 4 (Oct. 2021), 199. <https://doi.org/10.1007/s12046-021-01736-8>

- [195] Xuanyi Dong, David Jacob Kedziora, Katarzyna Musial, and Bogdan Gabrys. 2021. Automated Deep Learning: Neural Architecture Search Is Not the End. *arXiv preprint arXiv:2112.09245* (2021).
- [196] Xuanyi Dong, Lu Liu, Katarzyna Musial, and Bogdan Gabrys. 2020. NATS-Bench: Benchmarking NAS Algorithms for Architecture Topology and Size. *arXiv preprint arXiv:2009.00437v4* (2020).
- [197] Xuanyi Dong, Mingxing Tan, Adams Wei Yu, Daiyi Peng, Bogdan Gabrys, and Quoc V. Le. 2020. AutoHAS: Efficient Hyperparameter and Architecture Search. *arXiv preprint arXiv:2006.03656* (2020).
- [198] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in Automl: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 297–307.
- [199] Sebastien Dubois. 2017. DeepMining [Computer Software]. Retrieved from <https://github.com/sds-dubois/DeepMining>.
- [200] Celestine Dunner, Thomas Parnell, Kubilay Atasu, Manolis Sifalakis, and Haralampos Pozidis. 2017. Understanding and Optimizing the Performance of Distributed Machine Learning Applications on Apache Spark. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE. <https://doi.org/10.1109/bigdata.2017.8257942>
- [201] Andre Ebert, Michael Till Beck, Andy Mattausch, Lenz Belzner, and Claudia Limhoff-Popien. 2017. Qualitative Assessment of Recurrent Human Motion. In *2017 25th European Signal Processing Conference (EUSIPCO)*. 306–310. <https://doi.org/10.23919/EUSIPCO.2017.8081218>
- [202] Maryam Edalati, Ali Shariq Imran, Zenun Kastrati, and Sher Muhammad Daudpota. 2022. The Potential of Machine Learning Algorithms for Sentiment Classification of Students' Feedback on MOOC. In *Intelligent Systems and Applications (Lecture Notes in Networks and Systems)*, Kohei Arai (Ed.). Springer International Publishing, Cham, 11–22. https://doi.org/10.1007/978-3-030-82199-9_2
- [203] Radwa Elshawi, Mohamed Maher, and Sherif Sakr. 2019. Automated Machine Learning: State-of-The-Art and Open Challenges. *arXiv preprint arXiv:1906.02287* (2019).
- [204] Hugo Jair Escalante. 2020. Automated Machine Learning – a Brief Review at the End of the Early Years. *arXiv preprint arXiv:2008.08516* (2020).
- [205] Marcos Antonio Espinoza Mina and Doris Del Pilar Gallegos Barzola. 2019. Data Scientist: A Systematic Review of the Literature. In *Technology Trends (Communications in Computer and Information Science)*, Miguel Botto-Tobar, Guillermo Pizarro, Miguel Zúñiga-Prieto, Mayra D'Armas, and Miguel Zúñiga Sánchez (Eds.). Springer International Publishing, Cham, 476–487. https://doi.org/10.1007/978-3-030-05532-5_35
- [206] Esteban Real, Chen Liang, David R. So, Quoc V. Le. 2021. Automl-Zero [Computer Software]. Retrieved from https://github.com/google-research/google-research/tree/master/automl_zero.
- [207] Elton F. de S. Soares, Carlos Alberto V. Campos, and Sidney C. de Lucena. 2019. Online Travel Mode Detection Method Using Automated Machine Learning and Feature Engineering. *Future Generation Computer Systems* 101 (Dec. 2019), 1201–1212. <https://doi.org/10.1016/j.future.2019.07.056>
- [208] Facebook. 2021. Nevergrad [Computer Software]. Retrieved from <https://github.com/facebookresearch/nevergrad>.
- [209] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. <https://doi.org/10.1145/604045.604056>
- [210] Fairlearn. 2021. Fairlearn [Computer Software]. Retrieved from <https://github.com/fairlearn/fairlearn>.
- [211] Rasool Fakoor, Jonas W Mueller, Nick Erickson, Pratik Chaudhari, and Alexander J Smola. 2020. Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 8671–8681. <https://proceedings.neurips.cc/paper/2020/file/62d75fb2e3075506e8837d8f55021ab1-Paper.pdf>
- [212] Suliman Mohamed Fati, Amgad Muneer, Nur Arifin Akbar, and Shakirah Mohd Taib. 2021. A Continuous Cuffless Blood Pressure Estimation Using Tree-Based Pipeline Optimization Tool. *Symmetry* 13, 4 (April 2021), 686. <https://doi.org/10.3390/sym13040686>
- [213] Tom Fawcett. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters* 27, 8 (June 2006), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [214] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. <https://doi.org/10.1145/2783258.2783311>
- [215] Luís Ferreira, André Pilastri, Carlos Martins, Pedro Santos, and Paulo Cortez. 2021. A Scalable and Automated Machine Learning Framework to Support Risk Management. In *Agents and Artificial Intelligence (Lecture Notes in Computer Science)*, Ana Paula Rocha, Luc Steels, and Jaap van den Herik (Eds.). Springer International Publishing, Cham, 291–307. https://doi.org/10.1007/978-3-030-71158-0_14

- [216] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *International Conference on Machine Learning*. PMLR, 1126–1135.
- [217] Sean W. Fleming, David C. Garen, Angus G. Goodbody, Cara S. McCarthy, and Lexi C. Landers. 2021. Assessing the New Natural Resources Conservation Service Water Supply Forecast Model for the American West: A Challenging Test of Explainable, Automated, Ensemble Artificial Intelligence. *Journal of Hydrology* 602 (Nov. 2021), 126782. <https://doi.org/10.1016/j.jhydrol.2021.126782>
- [218] The Institute for Ethical Machine Learning. 2021. Awesome-Production-Machine-Learning. Retrieved March 10, 2021 from <https://github.com/EthicalML/awesome-production-machine-learning>
- [219] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research* 13, 70 (2012), 2171–2175.
- [220] Fortin, Félix-Antoine and Rainville, François-Michel De and Gardner, Marc-André and Parizeau, Marc and Gagné, Christian. 2021. Deap [Computer Software]. Retrieved from <https://github.com/DEAP/deap>.
- [221] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An Intersectional Definition of Fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.
- [222] Jeremy Fowers, Kalin Ovtcharov, Michael Papamichael, Todd Massengill, Ming Liu, Daniel Lo, Shlomi Alkalay, Michael Haselman, Logan Adams, Mahdi Ghandi, Stephen Heil, Prerak Patel, Adam Sapek, Gabriel Weisz, Lisa Woods, Sitaram Lanka, Steven K. Reinhardt, Adrian M. Caulfield, Eric S. Chung, and Doug Burger. 2018. A Configurable Cloud-Scale DNN Processor for Real-Time AI. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE. <https://doi.org/10.1109/isca.2018.00012>
- [223] Luca Franceschi. 2021. FAR-HO [Computer Software]. Retrieved from <https://github.com/lucfra/FAR-HO>.
- [224] Peter I Frazier. 2018. A Tutorial on Bayesian Optimization. *arXiv preprint arXiv:1807.02811* (2018).
- [225] Adrian Furnham and Hua Chu Boo. 2011. A Literature Review of the Anchoring Effect. *The Journal of Socio-Economics* 40, 1 (Feb. 2011), 35–42. <https://doi.org/10.1016/j.socloc.2010.10.008>
- [226] Daviti Gachechiladze, Filippo Lanubile, Nicole Novielli, and Alexander Serebrenik. 2017. Anger and Its Direction in Collaborative Software Development. In *2017 IEEE/ACM 39th International Conference on Software Engineering: New Ideas and Emerging Technologies Results Track (ICSE-NIER)*, 11–14. <https://doi.org/10.1109/ICSE-NIER.2017.18>
- [227] Soledad Galli. 2021. Feature-engine [Computer Software]. Retrieved from https://github.com/solegalli/feature_engine.
- [228] Spyridon Garmpis, Manolis Maragoudakis, and Aristogiannis Garmpis. 2022. Assisting Educational Analytics with AutoML Functionalities. *Computers* 11, 6 (June 2022), 97. <https://doi.org/10.3390/computers11060097>
- [229] Marc Garnica-Caparrós and Daniel Memmert. 2021. Understanding Gender Differences in Professional European Football through Machine Learning Interpretability and Match Actions Data. *Scientific Reports* 11, 1 (May 2021), 10805. <https://doi.org/10.1038/s41598-021-90264-w>
- [230] Gartner. 2021. *2021-2023 Emerging Technology Roadmap for Large Enterprises*. Technical Report. <https://www.gartner.com/en/publications/emerging-technology-roadmap-for-large-enterprises>
- [231] Sharma Gaurav. 2021. Complete Guide to Feature Engineering: Zero to Hero. Retrieved November 10, 2021 from <https://www.analyticsvidhya.com/blog/2021/09/complete-guide-to-feature-engineering-zero-to-hero/>
- [232] Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural Networks and the Bias/Variance Dilemma. *Neural computation* 4, 1 (1992), 1–58.
- [233] Alexander Gerling, Holger Ziekow, Andreas Hess, Ulf Schreier, Christian Seiffer, and Djaffar Ould Abdeslam. 2022. Comparison of Algorithms for Error Prediction in Manufacturing with Automl and a Cost-Based Metric. *Journal of Intelligent Manufacturing* 33, 2 (Feb. 2022), 555–573. <https://doi.org/10.1007/s10845-021-01890-0>
- [234] Jenny Gesley, Tariq Ahmad, Edouardo Soares, Ruth Levush, Gustavo Guerra, James Martin, Kelly Buchanan, Laney Zhang, Sayuri Umeda, Astghik Grigoryan, Nicolas Boring, Elin Hofverberg, George Sadek, Hanibal Goitom, Clare Feikhert-Ahalt, and Graciela Rodriguez-Ferrand. 2019. *Regulation of Artificial Intelligence in Selected Jurisdictions*. Technical Report. Law Library of Congress, Global Legal Research Directorate. <https://tile.loc.gov/storage-services/service/ll/lglrd/2019668143/2019668143.pdf>
- [235] Omid Gheibi, Danny Weyns, and Federico Quin. 2021. Applying Machine Learning in Self-Adaptive Systems: A Systematic Literature Review. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 15, 3 (2021), 1–37.
- [236] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine* 178, 11 (Nov. 2018), 1544–1547. <https://doi.org/10.1001/jamainternmed.2018.3763>
- [237] Pieter Gijsbers, Erin LeDell, Janek Thomas, Sébastien Poirier, Bernd Bischl, and Joaquin Vanschoren. 2019. An Open Source AutoML Benchmark. *arXiv preprint arXiv:1907.00909* (2019).
- [238] Gijsbers, Pieter. 2021. GAMA [Computer Software]. Retrieved from <https://github.com/PGijsbers/gama>.
- [239] Blue Yonder GmbH. 2016. tsfresh [Computer Software]. Retrieved from <https://github.com/blue-yonder/tsfresh>.
- [240] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A Survey of Deep Learning Techniques for Autonomous Driving. *Journal of Field Robotics* 37, 3 (2020), 362–386. <https://doi.org/10.1002/rob.21918>

- [241] Asela Gunawardana and Guy Shani. 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *The Journal of Machine Learning Research* 10 (Dec. 2009), 2935–2962.
- [242] H2O.ai. 2021. H2o-3 [Computer Software]. Retrieved from <https://github.com/h2oai/h2o-3>.
- [243] Mojtaba Haghightlari, Ching-Yen Shih, and Johannes Hachmann. 2019. Thinking Globally, Acting Locally: On the Issue of Training Set Imbalance and the Case for Local Machine Learning Models in Chemistry. *ChemRxiv preprint ChemRxiv:10.26434/chemrxiv.8796947.v1* (2019).
- [244] Mojtaba Haghightlari, Gaurav Vishwakarma, Doaa Altarawy, Ramachandran Subramanian, Bhargava U Kota, Aditya Sonpal, Srirangaraj Setlur, and Johannes Hachmann. 2020. ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 10, 4 (2020), e1458.
- [245] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Thakkar Shraddha, Rebecca Kusko, Susanna-Assunta Sansone, Weida Tong, Russ D. Wolfinger, Christopher E. Mason, Wendell Jones, Joaquin Dopazo, Cesare Furlanello, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush, Hugo J. W. L. Aerts, and Massive Analysis Quality Control (MAQC) Society Board of Directors. 2020. Transparency and Reproducibility in Artificial Intelligence. *Nature* 586, 7829 (2020), E14–E16.
- [246] Tuomas Halvari, Jukka K Nurminen, and Tommi Mikkonen. 2020. Testing the Robustness of AutoML Systems. *arXiv preprint arXiv:2005.02649* (2020).
- [247] Tao Han, Francisco Nauber Bernardo Gois, Ramsés Oliveira, Luan Rocha Prates, and Magda Moura de Almeida Porto. 2021. Modeling the Progression of COVID-19 Deaths Using Kalman Filter and AutoML. *Soft Computing* (Jan. 2021). <https://doi.org/10.1007/s00500-020-05503-5>
- [248] Torben Hansing, Mario Michael Krell, and Frank Kirchner. 2016. hyperSPACE: Automated Optimization of Complex Processing Pipelines for pySPACE. In *NIPS Workshop on Bayesian Optimization. NIPS Workshop on Bayesian Optimization (BayesOPT2016), December 5-10, Barcelona, Spain*.
- [249] Marc Hanussek, Matthias Blohm, and Maximilien Kintz. 2020. Can AutoML Outperform Humans? An Evaluation on Popular OpenML Datasets Using AutoML Benchmark. In *2020 2nd International Conference on Artificial Intelligence, Robotics and Control*. 29–32.
- [250] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems* 29 (2016).
- [251] Ayush Hariharan, Ankit Gupta, and Trisha Pal. 2020. Camplad: Cybersecurity Autonomous Machine Learning Platform for Anomaly Detection. In *Future of Information and Communication Conference*. Springer, 705–720.
- [252] Harvard Intelligent Probabilistic Systems Group. 2021. Spearmint [Computer Software]. Retrieved from <https://github.com/HIPS/Spearmint>.
- [253] Taylor Hatmaker. 2021. Reddit Is Buying Machine Learning Platform Spell. Retrieved June 18, 2022 from <https://social.techcrunch.com/2022/06/16/reddit-spell-machine-learning/>
- [254] Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F Cooper, and Gilles Clermont. 2013. Outlier Detection for Patient Monitoring and Alerting. *Journal of biomedical informatics* 46, 1 (2013), 47–55.
- [255] Chaoyang He. 2020. FedNas [Computer Software]. Retrieved from <https://github.com/chaoyanghe/FedNAS>.
- [256] Chaoyang He, Murali Annavaram, and Salman Avestimehr. 2020. FedNAS: Federated Deep Learning via Neural Architecture Search. *arXiv preprint arXiv:2004.08546* (2020).
- [257] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A Survey of the State-of-the-Art. *Knowledge-Based Systems* 212 (Jan. 2021), 106622. <https://doi.org/10.1016/j.knosys.2020.106622> arXiv:1908.00709
- [258] Yujian He. 2021. carefree-learn [Computer Software]. Retrieved from <https://github.com/carefree0910/carefree-learn>.
- [259] Yujian He. 2022. carefree-learn-deploy [Computer Software]. Retrieved from <https://github.com/carefree0910/carefree-learn-deploy>.
- [260] Christopher Hecht, Jan Figgener, and Dirk Uwe Sauer. 2021. Predicting Electric Vehicle Charging Station Availability Using Ensemble Machine Learning. *Energies* 14, 23 (2021), 7834.
- [261] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22, 1 (Jan. 2004), 5–53. <https://doi.org/10.1145/963770.963772>
- [262] Stephanie C Hicks and Rafael A Irizarry. 2018. A Guide to Teaching Data Science. *The American Statistician* 72, 4 (2018), 382–391.
- [263] Taryn Hodgdon, Rebecca E. Thornhill, Nick D. James, Paul E. Beaulé, Andrew D. Speirs, and Kawan S. Rakhra. 2020. CT Texture Analysis of Acetabular Subchondral Bone Can Discriminate between Normal and Cam-Positive Hips. *European Radiology* 30, 8 (Aug. 2020), 4695–4704. <https://doi.org/10.1007/s00330-020-06781-1>

- [264] Matthew Hoffman, Bobak Shahriari, and Nando Freitas. 2014. On Correlation and Budget Constraints in Model-Based Bandit Optimization with Application to Automatic Machine Learning. In *Artificial Intelligence and Statistics*. 365–374.
- [265] Daniel Homola. 2021. Boruta [Computer Software]. Retrieved from https://github.com/scikit-learn-contrib/boruta_py.
- [266] Moongi Simon Hong, Yu-Ho Lee, Jin-Min Kong, Oh-Jung Kwon, Cheol-Woong Jung, Jaeseok Yang, Myoung-Soo Kim, Hyun-Wook Han, Sang-Min Nam, and Korean Organ Transplantation Registry Study Group. 2022. Personalized Prediction of Kidney Function Decline and Network Analysis of the Risk Factors after Kidney Transplantation Using Nationwide Cohort Data. *Journal of Clinical Medicine* 11, 5 (Jan. 2022), 1259. <https://doi.org/10.3390/jcm11051259>
- [267] The White House. 2021. The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force. Retrieved June 14 2021 from <https://www.whitehouse.gov/ostp/news-upyears/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/>
- [268] Chris Howard and Andy Rowsell-Jones. 2019. *2019 CIO Survey: CIOs Have Awoken to the Importance of AI*. Technical Report. Gartner. <https://www.gartner.com/document/3897266>
- [269] Derek Howard, Marta M Maslej, Justin Lee, Jacob Ritchie, Geoffrey Woppard, and Leon French. 2020. Transfer Learning for Risk Classification of Social Media Posts: Model Evaluation Study. *Journal of medical Internet research* 22, 5 (2020), e15371.
- [270] Yuh-Jong Hu and Shu-Wei Huang. 2017. Challenges of Automated Machine Learning on Causal Impact Analytics for Policy Evaluation. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*. IEEE. <https://doi.org/10.1109/tel-net.2017.8343571>
- [271] Xiaohui Huang, Ze Yu, Xin Wei, Junfeng Shi, Yu Wang, Zeyuan Wang, Jihui Chen, Shuhong Bu, Lixia Li, Fei Gao, Jian Zhang, and Ajing Xu. 2021. Prediction of Vancomycin Dose on High-Dimensional Data Using Machine Learning Techniques. *Expert Review of Clinical Pharmacology* 14, 6 (June 2021), 761–771. <https://doi.org/10.1080/17512433.2021.1911642>
- [272] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. 2011. Sequential Model-Based Optimization for General Algorithm Configuration. In *Learning and Intelligent Optimization (Lecture Notes in Computer Science)*, Carlos A. Coello Coello (Ed.). Springer, Berlin, Heidelberg, 507–523. https://doi.org/10.1007/978-3-642-25566-3_40
- [273] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (Eds.). 2019. *Automated Machine Learning : Methods, Systems, Challenges*. Springer Nature. <https://doi.org/10.1007/978-3-030-05318-5>
- [274] IBM. 2021. *Global AI Adoption Index 2021*. Technical Report. https://filecache.mediарoom.com/mr5mr_ibmnews/190846/IBM%27s%20Global%20AI%20Adoption%20Index%202021_Executive-Summary.pdf
- [275] Joseph R. Imbus, Reese W. Randle, Susan C. Pitt, Rebecca S. Sippel, and David F. Schneider. 2017. Machine Learning to Identify Multigland Disease in Primary Hyperparathyroidism. *Journal of Surgical Research* 219 (Nov. 2017), 173–179. <https://doi.org/10.1016/j.jss.2017.05.117>
- [276] Thorir Mar Ingolfsson, Andrea Cossettini, Simone Benatti, and Luca Benini. 2022. Energy-Efficient Tree-Based EEG Artifact Detection. *arXiv preprint arXiv:2204.09577* (2022).
- [277] DAMA International. 2017. *DAMA-DMBOK: Data Management Body of Knowledge* (second ed.). Technics Publications, LLC.
- [278] Didier Frank Isingizwe, Meng Wang, Wenmao Liu, Dongsheng Wang, Tiejun Wu, and Jun Li. 2021. Analyzing Learning-based Encrypted Malware Traffic Classification with AutoML. In *2021 IEEE 21st International Conference on Communication Technology (ICCT)*. 313–322. <https://doi.org/10.1109/ICCT52962.2021.9658106>
- [279] Amit Kumar Jain, Maharsi Dhada, Ajith Kumar Parlakad, and Bhupesh Kumar Lad. 2020. Product Quality Driven Auto-Prognostics: Low-Cost Digital Solution for SMEs. *IFAC-PapersOnLine* 53, 3 (Jan. 2020), 78–83. <https://doi.org/10.1016/j.ifacol.2020.11.012>
- [280] Saksham Jain, Tayyibah Khanam, Ali Jafar Abedi, and Abid Ali Khan. 2022. Efficient Machine Learning for Malnutrition Prediction among Under-Five Children in India. In *2022 IEEE Delhi Section Conference (DELCON)*. 1–10. <https://doi.org/10.1109/DELCON54057.2022.9753080>
- [281] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.
- [282] Jiayi Liu, Samarth Tripathi, Unmesh Kurup, Mohak Shah. 2021. Auptimizer [Computer Software]. Retrieved from <https://github.com/LGE-ARC-AdvancedAI/auptimizer>.
- [283] José Jiménez. 2021. pyGPGO [Computer Software]. Retrieved from <https://github.com/josejimenezluna/pyGPGO>.
- [284] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- [285] Gunho Jung and Sun-Yong Choi. 2021. Forecasting Foreign Exchange Volatility Using Deep Learning Autoencoder-LSTM Techniques. *Complexity* 2021 (March 2021), e6647534. <https://doi.org/10.1155/2021/6647534>
- [286] Janka Kabathova and Martin Drlik. 2021. Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Applied Sciences* 11, 7 (Jan. 2021), 3130. <https://doi.org/10.3390/app11073130>

- [287] Petr Kadlec and Bogdan Gabrys. 2009. Architecture for Development of Adaptive On-Line Prediction Models. 1, 4 (2009), 241–269. <https://doi.org/10.1007/s12293-009-0017-8>
- [288] Petr Kadlec and Bogdan Gabrys. 2009. Evolving On-Line Prediction Model Dealing with Industrial Data Sets. In *2009 IEEE Workshop on Evolving and Self-Developing Intelligent Systems*. IEEE. <https://doi.org/10.1109/esdis.2009.4938995>
- [289] Petr Kadlec and Bogdan Gabrys. 2009. Soft Sensor Based on Adaptive Local Learning. In *Advances in Neuro-Information Processing*. Springer Berlin Heidelberg, 1172–1179. https://doi.org/10.1007/978-3-642-02490-0_142
- [290] Petr Kadlec and Bogdan Gabrys. 2009. Soft Sensors: Where Are We and What Are the Current and Future Challenges? 42, 19 (2009), 572–577. <https://doi.org/10.3182/20090921-3-tr-3005.00098>
- [291] Petr Kadlec and Bogdan Gabrys. 2010. Adaptive On-Line Prediction Soft Sensing without Historical Data. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE. <https://doi.org/10.1109/ijcnn.2010.5596965>
- [292] Petr Kadlec and Bogdan Gabrys. 2010. Local Learning-Based Adaptive Soft Sensor for Catalyst Activation Prediction. 57, 5 (2010), 1288–1301. <https://doi.org/10.1002/aic.12346>
- [293] Petr Kadlec, Ratko Grbić, and Bogdan Gabrys. 2011. Review of Adaptation Mechanisms for Data-Driven Soft Sensors. 35, 1 (2011), 1–24. <https://doi.org/10.1016/j.compchemeng.2010.07.034>
- [294] Kaggle. 2018. *2018 Kaggle Machine Learning & Data Science Survey*. Technical Report. <https://www.kaggle.com/kaggle/kaggle-survey-2018>
- [295] Kaggle. 2021. *2020 Kaggle Machine Learning & Data Science Survey*. Technical Report. <https://www.kaggle.com/c/kaggle-survey-2020/>
- [296] Li Kai-Yun, this link will open in a new window Link to external site, Niall G. Burnside, this link will open in a new window Link to external site, Raul Sampaio de Lima, this link will open in a new window Link to external site, Miguel Viloslada Peciña, Karli Sepp, Victor Henrique Cabral Pinheiro, Bruno Rucy Carneiro Alves de Lima, Ming-Der Yang, this link will open in a new window Link to external site, Ants Vain, Kalev Sepp, and this link will open in a new window Link to external site. 2021. An Automated Machine Learning Framework in Unmanned Aircraft Systems: New Insights into Agricultural Management Practices Recognition Approaches. *Remote Sensing* 13, 16 (2021), 3190. <https://doi.org/10.3390/rs13163190>
- [297] Faisal Kamiran and Toon Calders. 2012. Data Preprocessing Techniques for Classification without Discrimination. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 1–33. <https://doi.org/10.1007/s10115-011-0463-8>
- [298] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *2012 IEEE 12th International Conference on Data Mining*. 924–929. <https://doi.org/10.1109/ICDM.2012.45>
- [299] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3
- [300] Ameeth Kanawady and Aditya Sane. 2017. Machine Learning for Predictive Maintenance of Industrial Machines Using IoT Sensor Data. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 87–90.
- [301] Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R Collins, Jeff Schneider, Barnabas Poczos, and Eric P Xing. 2020. Tuning Hyperparameters without Grad Students: Scalable and Robust Bayesian Optimisation with Dragonfly. *Journal of Machine Learning Research* 21, 81 (2020), 1–27.
- [302] Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R. Collins, Jeff Schneider, Barnabas Poczos, and Eric P. Xing. 2021. Dragonfly [Computer Software]. Retrieved from <https://github.com/dragonfly/dragonfly>.
- [303] James Max Kanter, Owen Gillespie, and Kalyan Veeramachaneni. 2016. Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 430–439.
- [304] James Max Kanter and Kalyan Veeramachaneni. 2015. Deep Feature Synthesis: Towards Automating Data Science Endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Paris, France, October 19–21, 2015*. IEEE, 1–10.
- [305] Gilad Katz. 2021. ExploreKit [Computer Software]. Retrieved from <https://github.com/giladkatz/ExploreKit>.
- [306] Gilad Katz, Eui Chul Richard Shin, and Dawn Song. 2016. Explorekit: Automatic Feature Generation and Selection. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 979–984.
- [307] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.
- [308] David Jacob Kedziora, Katarzyna Musial, and Bogdan Gabrys. 2020. AutonoML: Towards an Integrated Framework for Autonomous Machine Learning. *arXiv preprint arXiv:2012.12600* (2020).
- [309] David Jacob Kedziora, Tien-Dung Nguyen, Katarzyna Musial, and Bogdan Gabrys. 2022. On Taking Advantage of Opportunistic Meta-knowledge to Reduce Configuration Spaces for Automated Machine Learning. *arXiv preprint arXiv:2208.04376* (2022).

- [310] Keras. 2021. AutoKeras [Computer Software]. Retrieved from <https://github.com/keras-team/autokeras>.
- [311] Zuhair Khayyat, Ihab F Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Nan Tang, and Si Yin. 2015. Bigdansing: A System for Big Data Cleansing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 1215–1230.
- [312] Khabat Khosravi, Rahim Barzegar, Ali Golkarian, Gianluigi Busico, Emilio Cuoco, Micòl Mastrocicco, Nicolò Colombari, Dario Tedesco, Maria Margarita Ntona, and Nerantzis Kazakis. 2021. Predictive Modeling of Selected Trace Elements in Groundwater Using Hybrid Algorithms of Iterative Classifier Optimizer. *Journal of Contaminant Hydrology* 242 (Oct. 2021), 103849. <https://doi.org/10.1016/j.jconhyd.2021.103849>
- [313] Tung Thanh Khuat, David Jacob Kedziora, Katarzyna Musial, and Bogdan Gabrys. 2021. The Roles and Modes of Human Interactions with Automated Machine Learning Systems. *arXiv preprint arXiv:2205.04139* (2021).
- [314] Zolo Kiala, John Odindi, and Onisimo Mutanga. 2022. Determining the Capability of the Tree-Based Pipeline Optimization Tool (TPOT) in Mapping Parthenium Weed Using Multi-Date Sentinel-2 Image Data. *Remote Sensing* 14, 7 (Jan. 2022), 1687. <https://doi.org/10.3390/rs14071687>
- [315] Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. 2017. Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. In *Artificial Intelligence and Statistics*. PMLR, 528–536.
- [316] Nicolas Knudde, Joachim van der Herten, Tom Dhaene, and Ivo Couckuyt. 2017. GPflowOpt: A Bayesian Optimization Library Using tensorflow. *arXiv preprint arXiv:1711.03845* (2017).
- [317] Nicolas Knudde, Joachim van der Herten, Tom Dhaene, and Ivo Couckuyt. 2021. GPflowOpt [Computer Software]. Retrieved from <https://github.com/GPflow/GPflowOpt>.
- [318] Simon Kocbek and Bogdan Gabrys. 2019. Automated Machine Learning Techniques in Prognostics of Railway Track Defects. In *2019 International Conference on Data Mining Workshops (ICDMW)*. 777–784. <https://doi.org/10.1109/ICDMW.2019.00115>
- [319] Simon Kocbek, P. Kocbek, T. Zupanic, G. Štiglic, and B. Gabrys. 2019. Using (Automated) Machine Learning and Drug Prescription Records to Predict Mortality and Polypharmacy in Older Type 2 Diabetes Mellitus Patients. In *ICONIP*. https://doi.org/10.1007/978-3-030-36808-1_68
- [320] Tomoki Kojima, Kazuto Oishi, Naoto Aoki, Yasushi Matsubara, Toshiaki Uete, Yoshihiko Fukushima, Goichi Inoue, Say Sato, Toru Shiraishi, Hiroyuki Hirooka, and Tatsuaki Masuda. 2022. Estimation of Beef Cow Body Condition Score: A Machine Learning Approach Using Three-Dimensional Image Data and a Simple Approach with Heart Girth Measurements. *Livestock Science* 256 (Feb. 2022), 104816. <https://doi.org/10.1016/j.livsci.2021.104816>
- [321] Bergstra J., Komor B., and Eliasmith C. 2021. Hyperopt-sklearn [Computer Software]. Retrieved from <https://github.com/hyperopt/hyperopt-sklearn>.
- [322] Georgios Kostopoulos, Theodor Panagiotakopoulos, Sotiris Kotsiantis, Christos Pierrakeas, and Achilles Kameas. 2021. Interpretable Models for Early Prediction of Certification in MOOCs: A Case Study on a MOOC for Smart City Professionals. *IEEE Access* 9 (2021), 165881–165891.
- [323] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. 2006. Machine Learning: A Review of Classification and Combining Techniques. *Artificial Intelligence Review* 26, 3 (Nov. 2006), 159–190. <https://doi.org/10.1007/s10462-007-9052-3>
- [324] Tim Kraska. 2018. Northstar: An Interactive Data Science System. *Proceedings of the VLDB Endowment* 11, 12 (Aug. 2018), 2150–2164. <https://doi.org/10.14778/3229863.3240493>
- [325] Sean Kross, Roger D Peng, Brian S Caffo, Ira Gooding, and Jeffrey T Leek. 2020. The Democratization of Data Science Education. *The American Statistician* 74, 1 (2020), 1–7.
- [326] Cedric Kulbach, Patrick Philipp, and Steffen Thoma. 2020. Personalized Automated Machine Learning. *ECAI 2020* (2020), 1246–1253. <https://doi.org/10.3233/FAIA200225>
- [327] Miron B Kursa and Witold R Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software* 36, 11 (2010), 1–13.
- [328] Laboratório de Inteligência Computacional (Computational Intelligence Lab) at Universidade Federal de Minas Gerais. 2021. Recipe [Computer Software]. Retrieved from <https://github.com/laic-ufmg/Recipe>.
- [329] Kwei-Herng Lai, Daochen Zha, Guanchu Wang, Junjie Xu, Yue Zhao, Devesh Kumar, Yile Chen, Purav Zumkhawaka, Minyang Wan, Diego Martinez, and Xia Hu. 2021. Tods: An Automated Time Series Outlier Detection System. In *Proceedings of the AaaI Conference on Artificial Intelligence*, Vol. 35. 16060–16062.
- [330] Himabindu Lakkaraju, Julius Adebayo, and Sameer Singh. 2020. Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities. *NeurIPS Tutorial* (2020).
- [331] Hoang Thanh Lam, Johann-Michael Thiebaut, Mathieu Sinn, Bei Chen, Tiep Mai, and Oznur Alkan. 2017. One Button Machine for Automating Feature Engineering in Relational Databases. *arXiv preprint arXiv:1706.00327* (2017).
- [332] Jean-Baptiste Lamare, Tobi Olatunji, and Li Yao. 2020. On the Diminishing Return of Labeling Clinical Reports. *arXiv preprint arXiv:2010.14587* (2020).
- [333] Justin Larocque-Villiers, Patrick Dumond, and David Knox. 2021. Automating Predictive Maintenance Using State-Based Transfer Learning and Ensemble Methods. In *2021 IEEE International Symposium on Robotic and Sensors*

- Environments (ROSE)*. 1–7. <https://doi.org/10.1109/ROSE52750.2021.9611768>
- [334] Xavier Larriva-Novo, Carmen Sánchez-Zas, Víctor A. Villagrá, Mario Vega-Barbas, and Diego Rivera. 2020. An Approach for the Application of a Dynamic Multi-Class Classifier for Network Intrusion Detection Systems. *Electronics* 9, 11 (Nov. 2020), 1759. <https://doi.org/10.3390/electronics9111759>
- [335] Erik Larsen, Korey MacVittie, and John Lilly. 2021. A Survey of Machine Learning Algorithms for Detecting Malware in IoT Firmware. *arXiv preprint arXiv:2111.02388* (2021).
- [336] Erik Larsen, David Noever, and Korey MacVittie. 2021. A Survey of Machine Learning Algorithms for Detecting Ransomware Encryption Activity. *arXiv preprint arXiv:2110.07636* (2021).
- [337] Igor Lazic, Florian Hinterwimmer, Severin Langer, Florian Pohlig, Christian Suren, Fritz Seidl, Daniel Rückert, Rainer Burgkart, and Rüdiger von Eisenhart-Rothe. 2022. Prediction of Complications and Surgery Duration in Primary Total Hip Arthroplasty Using Machine Learning: The Necessity of Modified Algorithms and Specific Data. *Journal of Clinical Medicine* 11, 8 (Jan. 2022), 2147. <https://doi.org/10.3390/jcm11082147>
- [338] Trang T Le, Weixuan Fu, and Jason H Moore. 2020. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. *Bioinformatics* 36, 1 (2020), 250–256.
- [339] Mickael Leclercq, Benjamin Vittrant, Marie Laure Martin-Magniette, Marie Pier Scott Boyer, Olivier Perin, Alain Bergeron, Yves Fradet, and Arnaud Droit. 2019. Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. *Frontiers in Genetics* 10 (2019).
- [340] Doris Jung Lin Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya G Parameswaran. 2019. A Human-in-the-Loop Perspective on AutoML: Milestones and the Road Ahead. *IEEE Data Eng. Bull.* 42, 2 (2019), 59–70.
- [341] Ming-Chang Lee, Jia-Chun Lin, and Ernst Gunnar Gran. 2021. DistTune: Distributed Fine-Grained Adaptive Traffic Speed Prediction for Growing Transportation Networks. *Transportation Research Record* 2675, 10 (Oct. 2021), 211–227. <https://doi.org/10.1177/03611981211011170>
- [342] Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [343] Alex Lekov. 2021. AutoML Alex [Computer Software]. Retrieved from https://github.com/Alex-Lekov/AutoML_Alex.
- [344] Christiane Lemke, Marcin Budka, and Bogdan Gabrys. 2015. Metalearning: A Survey of Trends and Technologies. 44, 1 (2015), 117–130. <https://doi.org/10.1007/s10462-013-9406-y>
- [345] Christiane Lemke and Bogdan Gabrys. 2010. Meta-Learning for Time Series Forecasting and Forecast Combination. 73, 10-12 (2010), 2006–2016. <https://doi.org/10.1016/j.neucom.2009.09.020>
- [346] Christiane Lemke and Bogdan Gabrys. 2010. Meta-Learning for Time Series Forecasting in the NN GC1 Competition. In *International Conference on Fuzzy Systems*. IEEE. <https://doi.org/10.1109/fuzzy.2010.5584001>
- [347] Christiane Lemke, Silvia Riedel, and Bogdan Gabrys. 2012. Evolving Forecast Combination Structures for Airline Revenue Management. 12, 3 (2012), 221–234. <https://doi.org/10.1057/rpm.2012.30>
- [348] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4 (Dec. 2018), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- [349] LF AI Foundation. 2021. AIExplainability360 [Computer Software]. Retrieved from <https://github.com/Trusted-AI/AIX360>.
- [350] LF AI Foundation. 2021. AIF360 [Computer Software]. Retrieved from <https://github.com/Trusted-AI/AIF360>.
- [351] Kai-Yun Li, Raul Sampaio de Lima, Niall G. Burnside, Ele Vahtmäe, Tiit Kutser, Karli Sepp, Victor Henrique Cabral Pinheiro, Ming-Der Yang, Ants Vain, and Kalle Sepp. 2022. Toward Automated Machine Learning-Based Hyperspectral Image Analysis in Crop Yield and Biomass Estimation. *Remote Sensing* 14, 5 (Jan. 2022), 1114. <https://doi.org/10.3390/rs14051114>
- [352] Rita Yi Man Li, Kwong Wing Chau, Herru Ching Yu Li, Fanjie Zeng, Beiqi Tang, and Meilin Ding. 2021. Remote Sensing, Heat Island Effect and Housing Price Prediction via AutoML. In *Advances in Artificial Intelligence, Software and Systems Engineering (Advances in Intelligent Systems and Computing)*, Tareq Ahram (Ed.). Springer International Publishing, Cham, 113–118. https://doi.org/10.1007/978-3-030-51328-3_17
- [353] Linux Foundation AI & Data. 2021. Ludwig [Computer Software]. Retrieved from <https://github.com/ludwig-ai/ludwig>.
- [354] Gengbo Liu, Dan Lu, and James Lu. 2021. Pharm-AutoML: An Open-Source, End-to-End Automated Machine Learning Package for Clinical Outcome Prediction. *CPT: pharmacometrics & systems pharmacology* 10, 5 (2021), 478–488.
- [355] Hanxiao Liu. 2021. Darts [Computer Software]. Retrieved from <https://github.com/quark0/darts>.
- [356] J. Liu, S. Tripathi, U. Kurup, and M. Shah. 2019. Auptimizer - an Extensible, Open-Source Framework for Hyperparameter Tuning. In *2019 IEEE International Conference on Big Data (Big Data)*. 339–348. <https://doi.org/10.1109/BigData47090.2019.9006330>
- [357] Pengjie Liu, Fucheng Pan, Xiaofeng Zhou, Shuai Li, and Liang Jin. 2022. CF-DAML: Distributed Automated Machine Learning Based on Collaborative Filtering. *Applied Intelligence* (2022), 1–25.

- [358] Pengjie Liu, Fucheng Pan, Xiaofeng Zhou, Shuai Li, Pengyu Zeng, Shurui Liu, and Liang Jin. 2022. Dsa-PAML: A Parallel Automated Machine Learning System via Dual-Stacked Autoencoder. *Neural Computing and Applications* (2022), 1–22.
- [359] Xingchi Liu, Peizheng Li, and Ziming Zhu. 2022. Bayesian Optimisation-Assisted Neural Network Training Technique for Radio Localisation. *arXiv preprint arXiv:2203.04032* (2022).
- [360] Jonathan Lorraine, Paul Vicol, and David Duvenaud. 2020-08-26/2020-08-28. Optimizing Millions of Hyperparameters by Implicit Differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Silvia Chiappa and Roberto Calandra (Eds.), Vol. 108. PMLR, 1540–1552. <https://proceedings.mlr.press/v108/lorraine20a.html>
- [361] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. 2019. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (Dec. 2019), 2346–2363. <https://doi.org/10.1109/TKDE.2018.2876857>
- [362] Yifeng Lu. 2019. An End-to-End AutoML Solution for Tabular Data at KaggleDays. Retrieved March 17, 2021 from <https://ai.googleblog.com/2019/05/an-end-to-end-automl-solution-for.html>
- [363] Scott Lundberg. 2021. Shap [Computer Software]. Retrieved from <https://github.com/slundberg/shap>.
- [364] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems* 30 (2017).
- [365] Ingrid Lunden. 2020. Intel Acquires SigOpt, a Specialist in Modeling Optimization, to Boost Its AI Business. Retrieved December 7, 2021 from <https://techcrunch.com/2020/10/29/intel-acquires-sigopt-a-specialist-in-modeling-optimization-to-boost-its-ai-business/>
- [366] Chenqiang Luo, Zhendong Zhang, Dongdong Qiao, Xin Lai, Yongying Li, and Shunli Wang. 2022. Life Prediction under Charging Process of Lithium-Ion Batteries Based on AutoML. *Energies* 15, 13 (Jan. 2022), 4594. <https://doi.org/10.3390/en15134594>
- [367] Gang Luo. 2016. PredicT-ML: A Tool for Automating Machine Learning Model Building with Big Clinical Data. 4, 1 (2016). <https://doi.org/10.1186/s13755-016-0018-1>
- [368] Gang Luo, Shan He, Bryan L. Stone, Flory L. Nkoy, and Michael D. Johnson. 2020. Developing a Model to Predict Hospital Encounters for Asthma in Asthmatic Patients: Secondary Analysis. *JMIR Medical Informatics* 8, 1 (Jan. 2020), e16080. <https://doi.org/10.2196/16080>
- [369] Gang Luo, Bryan L Stone, Michael D Johnson, Peter Tarczy-Hornoch, Adam B Wilcox, Sean D Mooney, Xiaoming Sheng, Peter J Haug, and Flory L Nkoy. 2017. Automating Construction of Machine Learning Models with Clinical Big Data: Proposal Rationale and Methods. 6, 8 (2017), e175. <https://doi.org/10.2196/resprot.7757>
- [370] Gang Luo, Bryan L. Stone, Xiaoming Sheng, Shan He, Corinna Koebnick, and Flory L. Nkoy. 2021. Using Computational Methods to Improve Integrated Disease Management for Asthma and Chronic Obstructive Pulmonary Disease: Protocol for a Secondary Analysis. *JMIR Research Protocols* 10, 5 (May 2021), e27065. <https://doi.org/10.2196/27065>
- [371] Zhibin Lv, Donghua Wang, Hui Ding, Bineng Zhong, and Lei Xu. 2020. Escherichia Coli DNA N-4-Methycytosine Site Prediction Accuracy Improved by Light Gradient Boosting Machine Feature Selection Technology. *IEEE Access* 8 (2020), 14851–14859. <https://doi.org/10.1109/ACCESS.2020.2966576>
- [372] K. Maass, A. Aravkin, and M. Kim. 2022. A Hyperparameter-Tuning Approach to Automated Inverse Planning. *Medical Physics* 49, 5 (2022), 3405–3415. <https://doi.org/10.1002/mp.15557>
- [373] Jorge G Madrid, Hugo Jair Escalante, Eduardo F Morales, Wei-Wei Tu, Yang Yu, Lisheng Sun-Hosoya, Isabelle Guyon, and Michèle Sebag. 2019. Towards AutoML in the Presence of Drift: First Results. *arXiv preprint arXiv:1907.10772* (2019).
- [374] Mary B. Makarious, Hampton L. Leonard, Dan Vitale, Hirotaka Iwaki, David Saffo, Lana Sargent, Anant Dadu, Eduardo Salmerón Castaño, John F. Carter, Melina Maleknia, Juan A. Botía, Cornelis Blauwendraat, Roy H. Campbell, Sayed Hadi Hashemi, Andrew B. Singleton, Mike A. Nalls, and Faraz Faghri. 2021. GenoML: Automated Machine Learning for Genomics. *arXiv preprint arXiv:2103.03221* (2021).
- [375] Spyros Makridakis. 2017. The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms. *Futures* 90 (June 2017), 46–60. <https://doi.org/10.1016/j.futures.2017.03.006>
- [376] Elisabetta Manduchi, Joseph D. Romano, and Jason H. Moore. 2021. The Promise of Automated Machine Learning for the Genetic Analysis of Complex Traits. *Human Genetics* (Oct. 2021). <https://doi.org/10.1007/s00439-021-02393-x>
- [377] Sowmya Mangalath Ravindran, Santosh Kumar Moorakkal Bhaskaran, Sooraj K. Ambat, Kannan Balakrishnan, and Manoj Mangattathil Gopalakrishnan. 2022. An Automated Machine Learning Methodology for the Improved Prediction of Reference Evapotranspiration Using Minimal Input Parameters. *Hydrological Processes* 36, 5 (2022), e14571. <https://doi.org/10.1002/hyp.14571>
- [378] Nathan Mantel and William Haenszel. 1959. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute* 22, 4 (April 1959), 719–748. <https://doi.org/10.1093/jnci/22.4.719>
- [379] Ruben Martinez-Cantin. 2014. BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits. *Journal of Machine Learning Research* 15, 1 (2014), 3735–3739.

- [380] Ruben Martinez-Cantin. 2021. Bayesopt [Computer Software]. Retrieved from <https://github.com/rmcantin/bayesopt>.
- [381] Suraya Masrom, Thuraiya Mohd, Nur Syafiqah Jamil, Abdullah Sani Abd. Rahman, and Norhayati Baharun. 2019. Automated Machine Learning Based on Genetic Programming: A Case Study on a Real House Pricing Dataset. In *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. 48–52. <https://doi.org/10.1109/AiDAS47888.2019.8970916>
- [382] Steve Koshy Mathew and Yu Zhang. 2020. Acoustic-Based Engine Fault Diagnosis Using WPT, PCA and Bayesian Optimization. *Applied Sciences* 10, 19 (Jan. 2020), 6890. <https://doi.org/10.3390/app10196890>
- [383] Caio Eduardo Falcão Matos, Johnatan Carvalho Souza, João Otávio Bandeira Diniz, Geraldo Braz Junior, Anselmo Cardoso de Paiva, João Dallyson Sousa de Almeida, Simara Vieira da Rocha, and Aristófanes Correa Silva. 2019. Diagnosis of Breast Tissue in Mammography Images Based Local Feature Descriptors. *Multimedia Tools and Applications* 78, 10 (May 2019), 12961–12986. <https://doi.org/10.1007/s11042-018-6390-x>
- [384] Ganjour Mazaev, Agusmian Partogi Ompusunggu, Georges Tod, Guillaume Crevecoeur, and Sofie Van Hoecke. 2020. Data-Driven Prognostics of Alternating Current Solenoid Valves. In *2020 Prognostics and Health Management Conference (PHM-Besançon)*. 109–115. <https://doi.org/10.1109/PHM-Besancon49106.2020.00024>
- [385] Cody Mazza-Anthony. 2021. A Five-Step Guide for Conducting Exploratory Data Analysis. Retrieved November 9, 2021 from <https://shopify.engineering/conducting-exploratory-data-analysis>
- [386] Robert T. McGibbon, Carlos X. Hernández, Matthew P. Harrigan, Steven Kearnes, Mohammad M. Sultan, Stanislaw Jastrzebski, Brooke E. Husic, and Vijay S. Pande. 2021. Osprey [Computer Software]. Retrieved from <https://github.com/msmbuilder/osprey>.
- [387] McKinsey. 2020. *Global Survey: The State of AI in 2020* / McKinsey. Technical Report. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020>
- [388] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
- [389] Ning Mei, Michael D Grossberg, Kenneth Ng, Karen T Navarro, and Timothy M Ellmore. 2017. Identifying Sleep Spindles with Multichannel EEG and Classification Optimization. *Computers in biology and medicine* 89 (2017), 441–453.
- [390] Moritz Meiners, Marlene Kuhn, and Jörg Franke. 2021. Manufacturing Process Curve Monitoring with Deep Learning. *Manufacturing Letters* 30 (Oct. 2021), 15–18. <https://doi.org/10.1016/j.mfglet.2021.09.006>
- [391] Hugo Abreu Mendes. 2021. On AutoMLs for Short-Term Solar Radiation Forecasting in Brazilian Northeast. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. 1–6. <https://doi.org/10.1109/ICEET53442.2021.9659788>
- [392] J. Mendikute, J. Plazaola, M. Baskaran, E. Zugasti, L. Aretxabaleta, and J. Aurrekoetxea. 2021. Impregnation Quality Diagnosis in Resin Transfer Moulding by Machine Learning. *Composites Part B: Engineering* 221 (Sept. 2021), 108973. <https://doi.org/10.1016/j.compositesb.2021.108973>
- [393] Rick Merritt. 2021. What Is MLOps? Retrieved May 28, 2021 from <https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/>
- [394] Epimack Michael, He Ma, Hong Li, and Shouliang Qi. 2022. An Optimized Framework for Breast Cancer Classification Using Machine Learning. *BioMed Research International* 2022 (Feb. 2022), e8482022. <https://doi.org/10.1155/2022/8482022>
- [395] Microsoft. 2020. FLAML [Computer Software]. Retrieved from <https://github.com/microsoft/FLAML>.
- [396] Patrick Mikalef, Michail N. Giannakos, Ilias O. Pappas, and John Krogstie. 2018. The Human Side of Big Data: Understanding the Skills of the Data Scientist in Education and Industry. In *2018 IEEE Global Engineering Education Conference (EDUCON)*. 503–512. <https://doi.org/10.1109/EDUCON.2018.8363273>
- [397] Steven Mills, Sylvain Duranton, Maximiliano Santinelli, Guangying Hua, Elias Baltassis, Stephan Thiel, and Olivier Muehlstein. 2021. *Are You Overestimating Your Responsible AI Maturity?* Technical Report. BCG. <https://www.bcg.com/publications/2021/the-four-stages-of-responsible-ai-maturity>
- [398] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. 2021. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. <https://doi.org/10.1109/TPAMI.2021.3059968>
- [399] MindsDB Inc. 2022. Lightwood [Computer Software]. Retrieved from <https://github.com/mindsdb/lightwood>.
- [400] MIT - The Human Data Interaction Project. 2021. ATM [Computer Software]. Retrieved from <https://github.com/HDI-Project/ATM>.
- [401] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. 2018. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv preprint arXiv:1811.07867* (2018).
- [402] MLJAR. 2021. Mljar [Computer Software]. Retrieved from <https://github.com/mljar/mljar-supervised>.
- [403] Yunjeong Mo, Dong Zhao, Jing Du, Matt Syal, Azizan Aziz, and Heng Li. 2020. Automated Staff Assignment for Building Maintenance Using Natural Language Processing. *Automation in Construction* 113 (May 2020), 103150.

- <https://doi.org/10.1016/j.autcon.2020.103150>
- [404] Aparna Mohapatra, Saumendra Pattnaik, Binod Kumar Pattanayak, Srikanta Pattnaik, and Suprava Ranjan Laha. 2022. Software Quality Prediction Using Machine Learning. In *Advances in Data Science and Management (Lecture Notes on Data Engineering and Communications Technologies)*, Samarjeet Borah, Sambit Kumar Mishra, Brojo Kishore Mishra, Valentina Emilia Balas, and Zdzislaw Polkowski (Eds.). Springer Nature, Singapore, 137–146. https://doi.org/10.1007/978-981-16-5685-9_14
- [405] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26, 4 (Aug. 2020), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- [406] Rafsanjani Muhammod, Sajid Ahmed, Dewan Md Farid, Swakkhar Shatabda, Alok Sharma, and Abdollah Dehzangi. 2019. PyFeat: A Python-based Effective Feature Generation Tool for DNA, RNA and Protein Sequences. *Bioinformatics* 35, 19 (Oct. 2019), 3831–3833. <https://doi.org/10.1093/bioinformatics/btz165>
- [407] Viswajit Mulpuru and Nidhi Mishra. 2021. In Silico Prediction of Fraction Unbound in Human Plasma from Chemical Fingerprint Using Automated Machine Learning. *ACS omega* 6, 10 (2021), 6791–6797.
- [408] Willian Muniz Do Nascimento and Luiz Gomes-Jr. 2022. Enabling Low-Cost Automatic Water Leakage Detection: A Semi-Supervised, autoML-based Approach. *Urban Water Journal* 0, 0 (April 2022), 1–11. <https://doi.org/10.1080/1573062X.2022.2056710>
- [409] Reiichiro Nakano. 2021. Xcessiv [Computer Software]. Retrieved from <https://github.com/reiinakano/xcessiv>.
- [410] National Disability Insurance Scheme (NDIS). 2021. *Personalised Budgets - Proposal for a New NDIS Budget Model Technical Information Paper*. Technical Report. <https://www.ndis.gov.au/media/3124/download>
- [411] Renato Negrinho. 2021. DeepArchitect [Computer Software]. Retrieved from https://github.com/negrinho/deep_architect.
- [412] E. W. T. Ngai, Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. 2011. The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature. *Decision Support Systems* 50, 3 (Feb. 2011), 559–569. <https://doi.org/10.1016/j.dss.2010.08.006>
- [413] Tien-Dung Nguyen, David Jacob Kedziora, Katarzyna Musial, and Bogdan Gabrys. 2021. Exploring Opportunistic Meta-knowledge to Reduce Search Spaces for Automated Machine Learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–8. <https://doi.org/10.1109/IJCNN52387.2021.9533431>
- [414] Tien-Dung Nguyen, Katarzyna Musial, and Bogdan Gabrys. 2021. AutoWeka4MCPs-AVATAR: Accelerating automated machine learning pipeline composition and optimisation. *Expert Systems with Applications* 185 (2021), 115643. <https://doi.org/10.1016/j.eswa.2021.115643>
- [415] Tien-Dung Nguyen, Thi Thanh Sang Nguyen, and Nhat-Tan Le. 2018. Calibration of Conductivity Sensor Using Combined Algorithm Selection and Hyperparameter Optimization: A Case Study. In *2018 International Conference on Advanced Technologies for Communications (ATC)*. 296–300. <https://doi.org/10.1109/ATC.2018.8587559>
- [416] Fernando Nogueira. 2021. Bayesian Optimization [Computer Software]. Retrieved from <https://github.com/fmfn/BayesianOptimization>.
- [417] Gabriel Novillo Rangone, Carlos Pizarro, and German Montejano. 2022. Automation of an Educational Data Mining Model Applying Interpretable Machine Learning and Auto Machine Learning. In *Communication and Smart Technologies (Smart Innovation, Systems and Technologies)*, Álvaro Rocha, Daniel Barredo, Paulo Carlos López-López, and Iván Puentes-Rivera (Eds.). Springer, Singapore, 22–30. https://doi.org/10.1007/978-981-16-5792-4_3
- [418] Tinase Nyabako, Brighton M. Mvumi, Tanya Stathers, Shaw Mlambo, and Macdonald Mubayiwa. 2020. Predicting Prostephanus truncatus (Horn) (Coleoptera: Bostrichidae) Populations and Associated Grain Damage in Smallholder Farmers' Maize Stores: A Machine Learning Approach. *Journal of Stored Products Research* 87 (May 2020), 101592. <https://doi.org/10.1016/j.jspr.2020.101592>
- [419] Code of Federal Regulations. 1978. PART 1607 - UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES (1978). Retrieved from <https://www.govinfo.gov/content/pkg/CFR-2014-title29-vol4/xml/CFR-2014-title29-vol4-part1607.xml>.
- [420] Office of the Australian Information Commissioner. 2018. *Guide to Data Analytics and the Australian Privacy Principles*. Technical Report. <https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-data-analytics-and-the-australian-privacy-principles>
- [421] Office of the Prime Minister and Cabinet. 2021. *Budget 2021-22 Fact Sheets - Artificial Intelligence*. Technical Report. <https://web.archive.org/web/20211027035957/https://digitaleconomy.pmc.gov.au/fact-sheets/artificial-intelligence>
- [422] Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole A. Lavender, La Creis Kidd, and Jason H. Moore. 2016. Automating Biomedical Data Science through Tree-Based Pipeline Optimization. In *Applications of Evolutionary Computation*. Springer International Publishing, 123–137. https://doi.org/10.1007/978-3-319-31204-0_9
- [423] Alena Olenko, Jason H Moore, Patryk Orzechowski, Randal S Olson, Junmei Cairns, Pedro J Caraballo, Richard M Weinshilboum, Liewei Wang, and Matthew K Breitenstein. 2018. Considerations for Automated Machine Learning in

- Clinical Metabolic Profiling: Altered Homocysteine Plasma Concentration Associated with Metformin Exposure. In *Pacific Symposium on Biocomputing 2018*. World Scientific, 460–471.
- [424] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. 2020. Deep Learning for Financial Applications : A Survey. *Applied Soft Computing* 93 (Aug. 2020), 106384. <https://doi.org/10.1016/j.asoc.2020.106384>
- [425] Meghana Padmanabhan, Pengyu Yuan, Govind Chada, and Hien Van Nguyen. 2019. Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction. *Journal of Clinical Medicine* 8, 7 (July 2019), 1050. <https://doi.org/10.3390/jcm8071050>
- [426] Tobias Pahlberg, Matthew Thurley, Djordje Popovic, and Olle Hagman. 2018. Crack Detection in Oak Flooring Lamellae Using Ultrasound-Excited Thermography. *Infrared Physics & Technology* 88 (Jan. 2018), 57–69. <https://doi.org/10.1016/j.infrared.2017.11.007>
- [427] Nelly Rosaura Palacios Salinas, Mitra Baratchi, Jan N. van Rijn, and Andreas Vollrath. 2021. Automated Machine Learning for Satellite Data: Integrating Remote Sensing Pre-trained Models into AutoML Systems. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track (Lecture Notes in Computer Science)*, Yuxiao Dong, Nicolas Kourtellis, Barbara Hammer, and Jose A. Lozano (Eds.). Springer International Publishing, Cham, 447–462. https://doi.org/10.1007/978-3-030-86517-7_28
- [428] Andrew Parker. 2010. The Spawn of Craigslist. Retrieved April 28, 2021 from <https://thegongshow.tumblr.com/post/345941486/the-spawn-of-craiglist-like-most-vcs-that-focus>
- [429] Preston Parry. 2021. auto_m1l [Computer Software]. Retrieved from https://github.com/ClimbsRocks/auto_m1l.
- [430] Kristian Pastor, Marko Ilić, Jovana Kojić, Marijana Ačanski, and Djura Vujić. 2022. Classification of Cereal Flour by Gas Chromatography – Mass Spectrometry (GC-MS) Liposoluble Fingerprints and Automated Machine Learning. *Analytical Letters* 55, 14 (Sept. 2022), 2220–2226. <https://doi.org/10.1080/00032719.2022.2050921>
- [431] Krutika Patidar. 2020. *Predicting Spin-Symmetry Breaking in Organic Photovoltaic Compounds Using a Data Mining Approach*. Ph.D. Dissertation. State University of New York at Buffalo.
- [432] Nick Payton. 2020. *Insights from Interviewing Dozens of Modelers*. Technical Report. SigOpt. <https://sigopt.com/blog/insights-from-interviewing-dozens-of-modelers/>
- [433] Yam Peleg. 2021. HungaBunga [Computer Software]. Retrieved from <https://github.com/ypeleg/HungaBunga>.
- [434] Marco Pellegrini. 2021. Accurate Prediction of Breast Cancer Survival through Coherent Voting Networks with Gene Expression Profiling. *Scientific Reports* 11, 1 (July 2021), 14645. <https://doi.org/10.1038/s41598-021-94243-z>
- [435] Roger Peng. 2015. The Reproducibility Crisis in Science: A Statistical Counterattack. *Significance* 12, 3 (2015), 30–32. <https://doi.org/10.1111/j.1740-9713.2015.00827.x>
- [436] People+AI Research (PAIR) Initiative at Google. 2021. What-If-Tool [Computer Software]. Retrieved from <https://github.com/PAIR-code/what-if-tool>.
- [437] Gregory Piatetsky. 2019. *KDnuggets Software Poll 2019*. Technical Report. KGnuggets. <https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html>
- [438] Gregory Piatetsky. 2020. *AutoML Poll Results: If You Try It, You'll like It More*. Technical Report. KGnuggets. <https://www.kdnuggets.com/automl-poll-results-if-you-try-it-youll-like-it-more.html/>
- [439] Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. 2021. A Survey on Deep Learning in Medicine: Why, How and When? *Information Fusion* 66 (Feb. 2021), 111–137. <https://doi.org/10.1016/j.inffus.2020.09.006>
- [440] Juan Pineda-Jaramillo and Óscar Arbeláez-Arenas. 2022. Assessing the Performance of Gradient-Boosting Models for Predicting the Travel Mode Choice Using Household Survey Data. *Journal of Urban Planning and Development* 148, 2 (2022), 04022007.
- [441] Geoff Pleiss, Manish Raghuvan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. *Advances in neural information processing systems* 30 (2017).
- [442] Iana S. Polonskaia, Ilya R. Aliev, and Nikolay O. Nikitin. 2021. Automated Evolutionary Design of CNN Classifiers for Object Recognition on Satellite Images. *Procedia Computer Science* 193 (Jan. 2021), 210–219. <https://doi.org/10.1016/j.procs.2021.10.021>
- [443] Claudia Pompa and Travis Burke. 2017. Data Science and Analytics Skills Shortage: Equipping the APEC Workforce with the Competencies Demanded by Employers. *APEC Human Resource Development Working Group* (2017).
- [444] David M. W. Powers. 2020. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *arXiv preprint arXiv:2010.16061* (2020).
- [445] Philipp Probst. 2021. tuneRanger [Computer Software]. Retrieved from <https://github.com/PhilippPro/tuneRanger>.
- [446] Shane J. Prochnow, Nickolas Scott Raterman, Megan Swenberg, Lilia Reddy, Ian Smith, Marina Romanyuk, and Thomas Fernandez. 2022. A Subsurface Machine Learning Approach at Hydrocarbon Production Recovery & Resource Estimates for Unconventional Reservoir Systems: Making Subsurface Predictions from Multimensional Data Analysis. *Journal of Petroleum Science and Engineering* 215 (Aug. 2022), 110598. <https://doi.org/10.1016/j.petrol.2022.110598>

- [447] Rizka Purwanto, Arindam Pal, Alan Blair, and Sanjay Jha. 2021. Man versus Machine: AutoML and Human Experts' Role in Phishing Detection. *arXiv preprint arXiv:2108.12193* (2021).
- [448] Puspus, Xavier. 2021. Parity-Fairness [Computer Software]. Retrieved from <https://github.com/xmpuspus/parity-fairness>.
- [449] PyCaret. 2021. PyCaret [Computer Software]. Retrieved from <https://github.com/pycaret/pycaret>.
- [450] Pymetrics. 2021. Audit-Ai [Computer Software]. Retrieved from <https://github.com/pymetrics/audit-ai>.
- [451] Pytorch. 2021. Botorch [Computer Software]. Retrieved from <https://github.com/pytorch/botorch>.
- [452] Qatar Computing Research Institute. 2020. Data_civilizer_system [Computer Software]. Retrieved from https://github.com/qcri/data_civilizer_system.
- [453] Wenwen Qi, Chong Xu, and Xiwei Xu. 2021. AutoGluon: A Revolutionary Framework for Landslide Hazard Analysis. *Natural Hazards Research* 1, 3 (Sept. 2021), 103–108. <https://doi.org/10.1016/j.nhres.2021.07.002>
- [454] Qlik. 2021. Qlik Acquires Big Squid to Expand Its Industry Leading Augmented Analytics Capabilities with No-Code Automated Machine Learning. Retrieved December 7, 2021 from <https://www.qlik.com/us/company/press-room/press-releases/qlik-acquires-big-squid-to-expand-its-industry-leading-augmented-analytics-capabilities>
- [455] Basheer Qolomany, Ihab Mohammed, Ala Al-Fuqaha, Mohsen Guizani, and Junaid Qadir. 2021. Trust-Based Cloud Machine Learning Model Selection for Industrial IoT and Smart City Services. *IEEE Internet of Things Journal* 8, 4 (Feb. 2021), 2943–2958. <https://doi.org/10.1109/JIOT.2020.3022323>
- [456] Siti Fairuz Mat Radzi, Muhammad Khalis Abdul Karim, M. Iqbal Saripan, Mohd Amiruddin Abd Rahman, Iza Nurzawani Che Isa, and Mohammad Johari Ibahim. 2021. Hyperparameter Tuning and Pipeline Optimization via Grid Search Method and Tree-Based AutoML in Breast Cancer Prediction. *Journal of Personalized Medicine* 11, 10 (Oct. 2021), 978. <https://doi.org/10.3390/jpm11100978>
- [457] RapidMiner. 2019. Automated Machine Learning. Retrieved September 2, 2021 from <https://rapidminer.com/glossary/automated-machine-learning/>
- [458] Ray. 2021. Tune-Sklearn [Computer Software]. Retrieved from <https://github.com/ray-project/tune-sklearn>.
- [459] Esteban Real, Chen Liang, David So, and Quoc Le. 2020. Automl-Zero: Evolving Machine Learning Algorithms from Scratch. In *International Conference on Machine Learning*. PMLR, 8007–8019.
- [460] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zihui Li, Xiaojiang Chen, and Xin Wang. 2021. A Comprehensive Survey of Neural Architecture Search: Challenges and Solutions. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–34.
- [461] Emre Rençberoğlu. 2021. Fundamental Techniques of Feature Engineering for Machine Learning. Retrieved November 10, 2021 from <https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114>
- [462] Roger Resmini, Lincoln Faria da Silva, Petrucio R. T. Medeiros, Adriel S. Araujo, Débora C. Muchaluat-Saade, and Aura Conci. 2021. A Hybrid Methodology for Breast Screening and Cancer Diagnosis Using Thermography. *Computers in Biology and Medicine* 135 (Aug. 2021), 104553. <https://doi.org/10.1016/j.combiomed.2021.104553>
- [463] Rachid Riad, Marine Lunven, Hadrien Titeux, Xuan-Nga Cao, Jennifer Hamet Bagnou, Laurie Lemoine, Justine Montillot, Agnes Sliwinski, Katia Youssouf, Laurent Cleret de Langavant, Emmanuel Dupoux, and Anne-Catherine Bachoud-Lévi. 2022. Predicting Clinical Scores in Huntington's Disease: A Lightweight Speech Test. *Journal of Neurology* (May 2022). <https://doi.org/10.1007/s00415-022-11148-1>
- [464] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016*, 1135–1144.
- [465] Marco Tulio Correia Ribeiro. 2021. Lime [Computer Software]. Retrieved from <https://github.com/marcotcr/lime>.
- [466] Christian Ritter, Thomas Wollmann, Patrick Bernhard, Manuel Gunkel, Delia M. Braun, Ji-Young Lee, Jan Meiners, Ronald Simon, Guido Sauter, Holger Erfle, Karsten Rippe, Ralf Bartenschlager, and Karl Rohr. 2019. Hyperparameter Optimization for Image Analysis: Application to Prostate Tissue Images and Live Cell Data of Virus-Infected Cells. *International Journal of Computer Assisted Radiology and Surgery* 14, 11 (Nov. 2019), 1847–1857. <https://doi.org/10.1007/s11548-019-02010-3>
- [467] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. 2021. Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans. *Nature Machine Intelligence* 3, 3 (March 2021), 199–217. <https://doi.org/10.1038/s42256-021-00307-0>
- [468] Marko Robnik-Šikonja, Miloš Radović, Smiljana Đorović, Bojana Andđelković-Ćirković, and Nenad Filipović. 2018. Modeling Ischemia with Finite Elements and Automated Machine Learning. *Journal of Computational Science* 29 (Nov. 2018), 99–106. <https://doi.org/10.1016/j.jocs.2018.09.017>
- [469] Joseph D Romano, Trang T Le, Weixuan Fu, and Jason H Moore. 2020. Is Deep Learning Necessary for Simple Classification Tasks? *arXiv preprint arXiv:2006.06730* (2020).

- [470] rstecca. 2021. sklearn-deap [Computer Software]. Retrieved from <https://github.com/rstecca/sklearn-deap>.
- [471] Chaitanya K. Rudrabhatla. 2020. Comparison of Zero Downtime Based Deployment Techniques in Public Cloud Infrastructure. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. 1082–1086. <https://doi.org/10.1109/I-SMAC49090.2020.9243605>
- [472] Dymitr Ruta, Bogdan Gabrys, and Christiane Lemke. 2011. A Generic Multilevel Architecture for Time Series Prediction. 23, 3 (2011), 350–359. <https://doi.org/10.1109/tkde.2010.137>
- [473] Ingrid Rye, Alexandra Vik, Marek Kocinski, Alexander S. Lundervold, and Astri J. Lundervold. 2022. Predicting Conversion to Alzheimer’s Disease in Individuals with Mild Cognitive Impairment Using Clinically Transferable Features. *Scientific Reports* 12, 1 (Sept. 2022), 15566. <https://doi.org/10.1038/s41598-022-18805-5>
- [474] Jungwoo Ryoo, Syed Rizvi, William Aiken, and John Kissell. 2014. Cloud Security Auditing: Challenges and Emerging Approaches. *IEEE Security Privacy* 12, 6 (2014), 68–74. <https://doi.org/10.1109/MSP.2013.132>
- [475] Denham Sadley. 2021. NDIS ‘Robo-Plans’ Test Algorithmic Transparency. Retrieved October 1, 2021 from <https://www.innovationaus.com/ndis-robo-plans-test-algorithmic-transparency/>
- [476] Waddah Saeed. 2021. Comparison of Automated Machine Learning Tools for SMS Spam Message Filtering. In *Advances in Cyber Security (Communications in Computer and Information Science)*, Nibras Abdullah, Selvakumar Manickam, and Mohammed Anbar (Eds.). Springer, Singapore, 307–316. https://doi.org/10.1007/978-981-16-8059-5_18
- [477] Bikash Chandra Saha, Joshuva Arockia Dhanraj, M. Sujatha, R. Vallikannu, Mohana Alanazi, Ahmad Almadhor, Ravishankar Sathyamurthy, Kuma Gowwomsa Erko, and V. Sugumaran. 2022. Investigating Rotor Conditions on Wind Turbines Using Integrating Tree Classifiers. *International Journal of Photoenergy* 2022 (June 2022), e5389574. <https://doi.org/10.1155/2022/5389574>
- [478] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2019. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577* (2019).
- [479] Manuel Martin Salvador. 2017. *Automatic and Adaptive Preprocessing for the Development of Predictive Models*. Ph.D. Dissertation.
- [480] Manuel Martin Salvador, Marcin Budka, and Bogdan Gabrys. 2016. Adapting Multicomponent Predictive Systems Using Hybrid Adaptation Strategies with Auto-Weka in Process Industry. In *Workshop on Automatic Machine Learning*. 48–57.
- [481] Manuel Martin Salvador, Marcin Budka, and Bogdan Gabrys. 2016. Towards Automatic Composition of Multicomponent Predictive Systems. In *Hybrid Artificial Intelligent Systems*. Springer International Publishing, 27–39.
- [482] Manuel Martin Salvador, Marcin Budka, and Bogdan Gabrys. 2019. Automatic Composition and Optimization of Multicomponent Predictive Systems With an Extended Auto-WEKA. *IEEE Transactions on Automation Science and Engineering* 16, 2 (April 2019), 946–959. <https://doi.org/10.1109/TASE.2018.2876430> arXiv:1612.08789
- [483] Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, and Sathvik Koneru. 2020. Hazards of Data Leakage in Machine Learning: A Study on Classification of Breast Cancer Using Deep Neural Networks. In *Medical Imaging 2020: Computer-Aided Diagnosis*, Vol. 11314. International Society for Optics and Photonics, 1131416. <https://doi.org/10.1117/12.2549313>
- [484] Nadia N. Sánchez-Pozo, Juan S. Mejía-Ordóñez, Diana C. Chamorro, Dagoberto Mayorca-Torres, and Diego H. Peluffo-Ordóñez. 2021. Predicting High School Students’ Academic Performance: A Comparative Study of Supervised Machine Learning Techniques. In *2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop*. IEEE, 1–6.
- [485] Shubhra Kanti Karmaker Santu, Md Mahadi Hassan, Micah J. Smith, Lei Xu, ChengXiang Zhai, and Kalyan Veeramachaneni. 2020. A Level-wise Taxonomic Perspective on Automated Machine Learning to Date and Beyond: Challenges and Opportunities. *arXiv preprint arXiv:2010.10777* (2020).
- [486] Roger Schaer, Henning Müller, and Adrien Depeursinge. 2016. Optimized Distributed Hyperparameter Search and Simulation for Lung Texture Classification in CT Using Hadoop. 2, 2 (2016), 19. <https://doi.org/10.3390/jimaging2020019>
- [487] Susan Schneegans, Jake Lewis, and Tiffany Straza. 2021. *UNESCO Science Report 2021: The Race against Time for Smarter Development*. Technical Report. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000377433>
- [488] Lars Ole Schwen, Daniela Schacherer, Christian Geißler, and André Homeyer. 2022. Evaluating Generic AutoML Tools for Computational Pathology. *Informatics in Medicine Unlocked* 29 (Jan. 2022), 100853. <https://doi.org/10.1016/j.imu.2022.100853>
- [489] Scikit-Optimize. 2021. Scikit-Optimize [Computer Software]. Retrieved from <https://github.com/scikit-optimize/scikit-optimize>.
- [490] Scottfree Analytics LLC. 2021. AlphaPy [Computer Software]. Retrieved from <https://github.com/ScottfreeLLC/AlphaPy>.
- [491] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems* 28 (2015).

- [492] Frank Sehnke, Martin D Felder, and Anton K Kaifel. 2012. Learn-o-Matic: A Fully Automated Machine Learning Suite for Profile Retrieval Applications. In *Proceedings of the 2012 EUMETSAT Meteorological Satellite Conference*.
- [493] Kanghyeon Seo, Bokjin Chung, Hamsa Priya Panchaseelan, Taewoo Kim, Hyejung Park, Byungmo Oh, Minho Chun, Sunjae Won, Donkyu Kim, Jaewon Beom, Doyoung Jeon, and Jihoon Yang. 2021. Forecasting the Walking Assistance Rehabilitation Level of Stroke Patients Using Artificial Intelligence. *Diagnostics* 11, 6 (June 2021), 1096. <https://doi.org/10.3390/diagnostics11061096>
- [494] Ram Seshadri. 2021. Auto_TS [Computer Software]. Retrieved from https://github.com/AutoViML/Auto_TS.
- [495] Ram Seshadri. 2021. Auto_ViML [Computer Software]. Retrieved from https://github.com/AutoViML/Auto_ViML.
- [496] Palash Shah. 2020. Libra [Computer Software]. Retrieved from <https://github.com/Palashio/libra>.
- [497] Colin Shearer. 2000. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of data warehousing* 5, 4 (2000), 13–22.
- [498] Yunzhi Shi, Raj Biswas, Mehdi Noori, Michael Kilberry, John Oram, Joe Mays, Sachin Kharude, Dinesh Rao, and Xin Chen. 2021. Predicting Road Accident Risk Using Geospatial Data and Machine Learning (Demo Paper). In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21)*. Association for Computing Machinery, New York, NY, USA, 512–515. <https://doi.org/10.1145/3474717.3484253>
- [499] Terence Shin. 2021. An Extensive Step by Step Guide to Exploratory Data Analysis. Retrieved November 9 2021 from <https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e>
- [500] Rui Shu, Tianpei Xia, Jianfeng Chen, Laurie Williams, and Tim Menzies. 2021. How to Better Distinguish Security Bug Reports (Using Dual Hyperparameter Optimization). *Empirical Software Engineering* 26, 3 (April 2021), 53. <https://doi.org/10.1007/s10664-020-09906-8>
- [501] Vashisth Shubhangi, Erick Brethenoux, Farhan Choudhary, and Jim Hare. 2020. *Use Gartner's 3-Stage MLOps Framework to Successfully Operationalize Machine Learning Projects*. Technical Report. <https://www.gartner.com/document/3987104>
- [502] Ashton Sidhu. 2021. Aethos [Computer Software]. Retrieved from <https://github.com/Ashton-Sidhu/aethos>.
- [503] Marion R. Sills, Mustafa Ozkaynak, and Hoon Jang. 2021. Predicting Hospitalization of Pediatric Asthma Patients in Emergency Departments Using Machine Learning. *International Journal of Medical Informatics* 151 (July 2021), 104468. <https://doi.org/10.1016/j.ijmedinf.2021.104468>
- [504] Lincoln Silva, Flávio Seixas, Cristina Fontes, Débora Muchaluat-Saade, and Aura Conci. 2020. A Computational Method for Breast Abnormality Detection Using Thermographs. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 469–474. <https://doi.org/10.1109/CBMS49503.2020.00095>
- [505] Lincoln F. Silva, Alair Augusto S. M. D. Santos, Renato S. Bravo, Aristófanes C. Silva, Débora C. Muchaluat-Saade, and Aura Conci. 2016. Hybrid Analysis for Indicating Patients with Breast Cancer Using Temperature Time Series. *Computer Methods and Programs in Biomedicine* 130 (July 2016), 142–153. <https://doi.org/10.1016/j.cmpb.2016.03.002>
- [506] Amarnpreet Singh, Narina Thakur, and Aakanksha Sharma. 2016. A Review of Supervised Machine Learning Algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. 1310–1315.
- [507] Devendra Singh, Pawan Kumar Pant, Himanshu Pant, and Dinesh C. Dobhal. 2021. Robust Automated Machine Learning (AutoML) System for Early Stage Hepatic Disease Detection. In *Intelligent Data Communication Technologies and Internet of Things (Lecture Notes on Data Engineering and Communications Technologies)*, Jude Hemanth, Robert Bestak, and Joy Iong-Zong Chen (Eds.). Springer, Singapore, 65–76. https://doi.org/10.1007/978-981-15-9509-7_6
- [508] Amin Sleimi, Nicolas Sannier, Mehrdad Sabetzadeh, Lionel Briand, Marcello Ceci, and John Dann. 2021. An Automated Framework for the Extraction of Semantic Legal Metadata from Legal Texts. *Empirical Software Engineering* 26, 3 (March 2021), 43. <https://doi.org/10.1007/s10664-020-09933-5>
- [509] Micah J. Smith, Carles Sala, James Max Kanter, and Kalyan Veeramachaneni. 2020. The Machine Learning Bazaar: Harnessing the ML Ecosystem for Effective System Development. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (June 2020), 785–800. <https://doi.org/10.1145/3318464.3386146> arXiv:1905.08942
- [510] Australian Computer Society. 2021. *DIGITAL PULSE 2021*. Technical Report. <https://www.acs.org.au/insightsandpublications/reports-publications/digital-pulse-2021.html>
- [511] Society For Human Resource Management. 2016. *Jobs of the Future: Data Analysis Skills*. Technical Report. <https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/documents/data-analysis-skills.pdf>
- [512] Evan R. Sparks, Ameet Talwalkar, Daniel Haas, Michael J. Franklin, Michael I. Jordan, and Tim Kraska. 2015. Automating Model Search for Large Scale Machine Learning. In *Proceedings of the Sixth ACM Symposium on Cloud Computing - SoCC 15*. ACM Press. <https://doi.org/10.1145/2806777.2806945>
- [513] Evan R. Sparks, Shivaram Venkataraman, Tomer Kaftan, Michael J. Franklin, and Benjamin Recht. 2017. KeystoneML: Optimizing Pipelines for Large-Scale Advanced Analytics. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE. <https://doi.org/10.1109/icde.2017.109>
- [514] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness

- via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2239–2248. <https://doi.org/10.1145/3219819.3220046>
- [515] Marco Spruit and Miltiadis Lytras. 2018. Applied Data Science in Patient-Centric Healthcare: Adaptive Analytic Systems for Empowering Physicians and Patients. *Telematics and Informatics* 35, 4 (July 2018), 643–653. <https://doi.org/10.1016/j.tele.2018.04.002>
 - [516] Sreeradha.Basu@timesgroup.com. 2018. 50,000+ Data Science, AI Jobs Vacant Due to Shortage of Talent [Careers & Companies]: Market Has Twice the Number of Jobs than Jobseekers, Finds a Report. *The Economic Times (Online)* (2018).
 - [517] Sarthak Srivastava, Radhika N. K., Rajesh Srinivasan, Nishanth K. M. Nambisan, and Sai Siva Gorthi. 2021. Diagnosis of Sickle Cell Anemia Using AutoML on UV-Vis Absorbance Spectroscopy Data. *arXiv preprint arXiv:2111.12711* (2021).
 - [518] Stackoverflow. 2021. *2021 Developer Survey*. Technical Report. <https://insights.stackoverflow.com/survey/2021#technology>
 - [519] Brian Stanton and Jensen Theodore. 2021. *Trust and Artificial Intelligence*. Technical Report. National Institute of Standards and Technology, U.S. Department of Commerce. <https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8332-draft.pdf>
 - [520] Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. 2021. Towards CRISP-ML (Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction* 3, 2 (2021), 392–413.
 - [521] Alexander Y. Sun, Bridget R. Scanlon, Himanshu Save, and Ashraf Rateb. 2021. Reconstruction of GRACE Total Water Storage Through Automated Machine Learning. *Water Resources Research* 57, 2 (2021), e2020WR028666. <https://doi.org/10.1029/2020WR028666>
 - [522] Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002* (2020).
 - [523] Daniela F. Taino, Matheus G. Ribeiro, Guilherme F. Roberto, Geraldo F. D. Zafalon, Marcelo Z. do Nascimento, Thaína A. A. Tosta, Alessandro S. Martins, and Leandro A. Neves. 2021. Analysis of Cancer in Histological Images: Employing an Approach Based on Genetic Algorithm. *Pattern Analysis and Applications* 24, 2 (May 2021), 483–496. <https://doi.org/10.1007/s10044-020-00931-3>
 - [524] Jiaqi Tan, Shutao Xie, Wencong Wu, Peijia Qin, and Tiancheng Ouyang. 2022. Evaluating and Optimizing the Cold Energy Efficiency of Power Generation and Wastewater Treatment in LNG-fired Power Plant Based on Data-Driven Approach. *Journal of Cleaner Production* 334 (Feb. 2022), 130149. <https://doi.org/10.1016/j.jclepro.2021.130149>
 - [525] Sean J Taylor and Benjamin Letham. 2018. Forecasting at Scale. *The American Statistician* 72, 1 (2018), 37–45.
 - [526] TensorFlow. 2021. Adanet [Computer Software]. Retrieved from Retrieved from <https://github.com/tensorflow/adanet>.
 - [527] The Artificial Intelligence Innovation (AII) research laboratory at Cedars-Sinai Medical Center. 2021. TPOT [Computer Software]. Retrieved from <https://github.com/EpistasisLab/tpot>.
 - [528] The Machine Learning Group at The University of Sheffield. 2021. GPyOpt [Computer Software]. Retrieved from <https://github.com/SheffieldML/GPyOpt>.
 - [529] The mlr-org. 2021. mlrMBO [Computer Software]. Retrieved from <https://github.com/mlr-org/mlrMBO>.
 - [530] Janek Thomas. 2021. AutoXGBoost [Computer Software]. Retrieved from <https://github.com/ja-thomas/autogbost>.
 - [531] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2021. Deep Learning's Diminishing Returns: The Cost of Improvement Is Becoming Unsustainable. *IEEE Spectrum* 58, 10 (Oct. 2021), 50–55. <https://doi.org/10.1109/MSPEC.2021.9563954>
 - [532] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 847–855.
 - [533] D. Tighe, T. Lewis-Morris, and A. Freitas. 2019. Machine Learning Methods Applied to Audit of Surgical Outcomes after Treatment for Cancer of the Head and Neck. *British Journal of Oral and Maxillofacial Surgery* 57, 8 (Oct. 2019), 771–777. <https://doi.org/10.1016/j.bjoms.2019.05.026>
 - [534] Tiobe. 2021. Tiobe Index. Retrieved May 26, 2021 from <https://www.tiobe.com/tiobe-index>
 - [535] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 272–283.
 - [536] Anh Truong, Austin Walters, Jeremy Goodfellow, Keegan Hines, C. Bayan Bruss, and Reza Farivar. 2019. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)* (Nov. 2019), 1471–1479. <https://doi.org/10.1109/ICTAI.2019.00209> arXiv:1908.05557

- [537] Chun-Wei Tsai and Zhi-Yan Fang. 2021. An Effective Hyperparameter Optimization Algorithm for DNN to Predict Passengers at a Metro Station. *ACM Transactions on Internet Technology* 21, 2 (March 2021), 32:1–32:24. <https://doi.org/10.1145/3410156>
- [538] Athanasios Tsakonas and Bogdan Gabrys. 2012. GRADIENT: Grammar-driven Genetic Programming Framework for Building Multi-Component, Hierarchical Predictive Systems. 39, 18 (2012), 13253–13266. <https://doi.org/10.1016/j.eswa.2012.05.076>
- [539] Ioannis Tsamardinos, Paulos Charonyktakis, Georgios Papoutsoglou, Giorgos Borboudakis, Kleanthi Lakiotaki, Jean Claude Zenklusen, Hartmut Juhl, Ekaterini Chatzaki, and Vincenzo Lagani. 2022. Just Add Data: Automated Predictive Modeling for Knowledge Discovery and Feature Selection. *NPJ precision oncology* 6, 1 (2022), 1–17.
- [540] Maria Tsiakkaki, Georgios Kostopoulos, Sotiris Kotsiantis, and Omiros Ragos. 2020. Implementing AutoML in Educational Data Mining for Prediction Tasks. *Applied Sciences* 10, 1 (Jan. 2020), 90. <https://doi.org/10.3390/app10010090>
- [541] Kalinda Ukanwa and Roland T Rust. 2021. Algorithmic Discrimination in Service. *USC Marshall School of Business Research Paper* (2021).
- [542] Danilo Valdes-Ramirez, Miguel A. Medina-Pérez, and Raúl Monroy. 2021. An Ensemble of Fingerprint Matching Algorithms Based on Cylinder Codes and Mtriplets for Latent Fingerprint Identification. *Pattern Analysis and Applications* 24, 2 (May 2021), 433–444. <https://doi.org/10.1007/s10044-020-00911-7>
- [543] Wessel A. van Eeden, Chuan Luo, Albert M. van Hemert, Ingrid V. E. Carlier, Brenda W. Penninx, Klaas J. Wardenaar, Holger Hoos, and Erik J. Giltay. 2021. Predicting the 9-Year Course of Mood and Anxiety Disorders with Automated Machine Learning: A Comparison between Auto-Sklearn, Naïve Bayes Classifier, and Traditional Logistic Regression. *Psychiatry Research* 299 (May 2021), 113823. <https://doi.org/10.1016/j.psychres.2021.113823>
- [544] Vanawat, Nimit. 2021. How To Perform Exploratory Data Analysis -A Guide for Beginners. Retrieved November 09, 2021 from <https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/>
- [545] Jake VanderPlas. 2016. *Python Data Science Handbook: Essential Tools for Working with Data.* " O'Reilly Media, Inc".
- [546] Joaquin Vanschoren. 2018. Meta-Learning: A Survey. *arXiv preprint arXiv:1810.03548* (2018).
- [547] D. Venkata Vara Prasad, P. Senthil Kumar, Lokeshwari Y. Venkataramana, G. Prasannamedha, S. Harshana, S. Jahnavi Srividya, K. Harrinei, and Sravya Indraganti. 2021. Automating Water Quality Analysis Using ML and Auto ML Techniques. *Environmental Research* 202 (Nov. 2021), 111720. <https://doi.org/10.1016/j.envres.2021.111720>
- [548] Gaurav Vishwakarma, Mojtaba Haghishatlari, and Johannes Hachmann. 2019. Towards Autonomous Machine Learning in Chemistry via Evolutionary Algorithms. *ChemRxiv preprint ChemRxiv:10.26434/chemrxiv.9782387.v1* (2019).
- [549] Can Wang, Thomas Bäck, Holger H. Hoos, Mitra Baratchi, Steffen Limmer, and Markus Olhofer. 2019. Automated Machine Learning for Short-term Electric Load Forecasting. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 314–321. <https://doi.org/10.1109/SSCI44817.2019.9002839>
- [550] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. 3 (2019), 1–24. Issue CSCW. <https://doi.org/10.1145/3359313>
- [551] Tao Wang, Ximin Wu, and Taiping He. 2019. Trustable and Automated Machine Learning Running with Blockchain and Its Applications. *arXiv preprint arXiv:1908.05725* (2019).
- [552] Zeyuan Wang, Josiah Poon, Shuze Wang, Shiding Sun, and Simon Poon. 2021. A Novel Method for Clinical Risk Prediction with Low-Quality Data. *Artificial Intelligence in Medicine* 114 (April 2021), 102052. <https://doi.org/10.1016/j.artmed.2021.102052>
- [553] Geoffrey I. Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. 2016. Characterizing Concept Drift. *Data Mining and Knowledge Discovery* 30, 4 (July 2016), 964–994. <https://doi.org/10.1007/s10618-015-0448-4>
- [554] Wayne Wei. 2021. Awesome-AutoML. Retrieved March 10, 2021 from <https://github.com/windmaple/awesome-AutoML>
- [555] Jason West. 2018. Teaching Data Science: An Objective Approach to Curriculum Validation. *Computer Science Education* 28, 2 (2018), 136–157.
- [556] Hadley Wickham and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1st ed.). O'Reilly Media, Inc.
- [557] Wikipedia. 2021. Automated machine learning. Retrieved December 5, 2021 from https://en.wikipedia.org/w/index.php?title=Automated_machine_learning&oldid=1054510848
- [558] Leann Williams. 2021. Clearsense Acquires Compellon. Retrieved December 7, 2021 from <https://clearsense.com/insights/clearsense-acquires-plug-and-play-ai-analytics-firm/>
- [559] Manoh L. Williams, Wilian P. James, and Michael T. Rose. 2019. Variable Segmentation and Ensemble Classifiers for Predicting Dairy Cow Behaviour. *Biosystems Engineering* 178 (Feb. 2019), 156–167. <https://doi.org/10.1016/j.biosystemsengineering.2019.01.001>

- [biosystemseng.2018.11.011](#)
- [560] Martin Wistuba, Ambrish Rawat, and Tejaswini Pedapati. 2019. A Survey on Neural Architecture Search. *arXiv preprint arXiv:1905.01392v2* (2019).
- [561] Barbara Wixom, Thilini Ariyachandra, David Douglas, Michael Goul, Babita Gupta, Lakshmi Iyer, Uday Kulkarni, John Mooney, Gloria Phillips-Wren, and Ozgur Turetken. 2014. The Current State of Business Intelligence in Academia: The Arrival of Big Data. *Communications of the Association for Information Systems* 34, 1 (Jan. 2014). <https://doi.org/10.17705/1CAIS.03401>
- [562] Chengrun Yang, Yuji Akimoto, Dae Won Kim, and Madeleine Udell. 2021. OBOE [Computer Software]. Retrieved from <https://github.com/udellgroup/oboe>.
- [563] Li Yang and Abdallah Shami. 2020. On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing* (jul 2020). <https://doi.org/10.1016/j.neucom.2020.07.061>
- [564] Yu Yang, Jia Mao, Richard Nguyen, Annas Tohmeh, and Hen-Geul Yeh. 2022. Feature Construction and Selection for PV Solar Power Modeling. *arXiv preprint arXiv:2202.06226* (2022).
- [565] Quanming Yao, Mengshuo Wang, Yuqiang Chen, Wenyuan Dai, Yu-Feng Li, Wei-Wei Tu, Qiang Yang, and Yang Yu. 2018. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. *arXiv preprint arXiv:1810.13306* (2018).
- [566] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*. PMLR, 325–333.
- [567] Siyang Zeng, Mehrdad Arjomandi, Yao Tong, Zachary C. Liao, and Gang Luo. 2022. Developing a Machine Learning Model to Predict Severe Chronic Obstructive Pulmonary Disease Exacerbations: Retrospective Cohort Study. *Journal of Medical Internet Research* 24, 1 (Jan. 2022), e28953. <https://doi.org/10.2196/28953>
- [568] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
- [569] Cheng Zhang and Zehao Ye. 2020. Water Pipe Failure Prediction Using AutoML. *Facilities* 39, 1/2 (Jan. 2020), 36–49. <https://doi.org/10.1108/F-08-2019-0084>
- [570] Tianxiang Zhang, Jinya Su, Cunjia Liu, and Wen-Hua Chen. 2021. State and Parameter Estimation of the AquaCrop Model for Winter Wheat Using Sensitivity Informed Particle Filter. *Computers and Electronics in Agriculture* 180 (Jan. 2021), 105909. <https://doi.org/10.1016/j.compag.2020.105909>
- [571] Wenqiang Zhang, Peng Ge, Weidong Jin, and Jian Guo. 2018. Radar Signal Recognition Based on TPOT and LIME. In *2018 37th Chinese Control Conference (CCC)*. 4158–4163. <https://doi.org/10.23919/ChiCC.2018.8483165>
- [572] Xiaohang Zhang, Yuqi Li, and Zhengren Li. 2022. Comparative Research of Hyper-Parameters Mathematical Optimization Algorithms for Automatic Machine Learning in New Generation Mobile Network. *Mobile Networks and Applications* 27, 3 (June 2022), 928–935. <https://doi.org/10.1007/s11036-022-01913-x>
- [573] Yuyu Zhang, Mohammad Taha Bahadori, Hang Su, and Jimeng Sun. 2016. FLASH: Fast Bayesian Optimization for Data Analytic Pipelines. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press. <https://doi.org/10.1145/2939672.2939829>
- [574] Zhe Zhang and Daniel B. Neill. 2017. Identifying Significant Predictive Bias in Classifiers. *arXiv preprint arXiv:1611.08292* (2017).
- [575] Yue Zhao, Zain Nasrullah, and Zheng Li. 2019. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research* 20, 96 (2019), 1–7.
- [576] Jiang Zhong, Weili Guo, and Zhenhua Wang. 2016. Study on Network Failure Prediction Based on Alarm Logs. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*. 1–7. <https://doi.org/10.1109/ICBDSC.2016.7460337>
- [577] Bin Zhou, Evyatar Erell, Ian Hough, Alexandra Shtein, Allan C. Just, Victor Novack, Jonathan Rosenblatt, and Itai Kloog. 2020. Estimation of Hourly near Surface Air Temperature Across Israel Using an Ensemble Model. *Remote Sensing* 12, 11 (Jan. 2020), 1741. <https://doi.org/10.3390/rs12111741>
- [578] Liqian Zhou, Zhao Wang, Xiongfei Tian, and Lihong Peng. 2021. LPI-deepGBDT: A Multiple-Layer Deep Framework Based on Gradient Boosting Decision Trees for lncRNA-Protein Interaction Identification. *BMC Bioinformatics* 22, 1 (Oct. 2021), 479. <https://doi.org/10.1186/s12859-021-04399-8>
- [579] Zillow. 2020. Luminaire [Computer Software]. Retrieved from <https://github.com/zillow/luminaire>.
- [580] Indre Zliobaite, Albert Bifet, Mohamed Gaber, Bogdan Gabrys, Joao Gama, Leandro Minku, and Katarzyna Musial. 2012. Next Challenges for Adaptive Learning Systems. 14, 1 (2012), 48. <https://doi.org/10.1145/2408736.2408746>
- [581] Indre Zliobaite and Bogdan Gabrys. 2014. Adaptive Preprocessing for Streaming Data. 26, 2 (2014), 309–321. <https://doi.org/10.1109/tkde.2012.147>
- [582] Marc-André Zöller and Marco F. Huber. 2021. Benchmark and Survey of Automated Machine Learning Frameworks. *Journal of Artificial Intelligence Research* 70 (Jan. 2021), 409–472. <https://doi.org/10.1613/jair.1.11854>

- [583] Indrė Žliobaitė, Marcin Budka, and Frederic Stahl. 2015. Towards Cost-Sensitive Adaptation: When Is It Worth Updating Your Predictive Model? 150 (2015), 240–249. <https://doi.org/10.1016/j.neucom.2014.05.084>

Appendix A FADED HYPERPARAMETER OPTIMISATION TOOLS

Name	GitHub	Ref.
Adatune	https://github.com/awslabs/adatune	[71]
Darts	https://github.com/quark0/darts	[355]
DeepArchitect	https://github.com/negrinho/deep_architect	[411]
FAR-HO	https://github.com/lucfra/FAR-HO	[223]
GPyOpt	https://github.com/SheffieldML/GPyOpt	[528]
Optunity	https://github.com/claesemn/optunity	[157]
Osprey	https://github.com/msmbuilder/osprey	[386]
pyGPGO	https://github.com/josejimenezluna/pyGPGO	[283]
RoBO	https://github.com/automl/RoBO	[89]
sklearn-deap	https://github.com/rsteca/sklearn-deap	[470]
Spearmint	https://github.com/HIPS/Spearmint	[252]

Appendix B FADED AUTOML SYSTEMS

Name	GitHub	Company	Ref.
Adanet	https://github.com/tensorflow/adanet	Google	[526]
Advisor	https://github.com/tobegin3hub/advisor	Personal	[145]
Aethos	https://github.com/Ashton-Sidhu/aethos	Personal	[502]
Amla	https://github.com/CiscoAI/aml	Cisco	[156]
ATM	https://github.com/HDI-Project/ATM	Research	[400]
auto_ml	https://github.com/ClimbsRocks/auto_ml	Personal	[429]
Auto-Weka	https://github.com/automl/pyautoweka	Research	[94]
AutoXGBoost	https://github.com/ja-thomas/autoxgboost	Personal	[530]
DeepMining	https://github.com/sds-dubois/DeepMining	Research	[76, 199]
FedNas	https://github.com/chaoyanghe/FedNAS	Research	[255, 256]
HpBandSter	https://github.com/automl/HpBandSter	Research	[91]
MetaQNN	https://github.com/bowenbaker/metaqnn	Research	[100, 101]
MLBox	https://github.com/AxeldeRomblay/MLBox	Personal	[189]
Recipe	https://github.com/laic-ufmg/Recipe	Research	[328]
tuneRanger	https://github.com/PhilippPro/tuneRanger	Research	[445]
Xcessiv	https://github.com/reinakano/xcessiv	Research	[409]

Appendix C INSUFFICIENTLY DETAILED COMMERCIAL SYSTEMS

Name	Website	Ref.
Aible	https://www.aible.com/	[1]
Algolytics	https://algolytics.com/products/abm/	[2]
DMway	http://dmway.com/	[20]
dotData	https://dotdata.com/	[22]
Kortical	https://kortical.com/	[28]
neuralstudio.ai	https://neuralstudio.ai/	[33]
OptiScorer	https://optiscorer.com/	[35]
Pecan	https://www.pecan.ai/	[36]
Prevision.io	https://prevision.io/	[37]
SparkCognition	https://www.sparkcognition.com/products/darwin/	[42]
TAZI	https://www.tazi.ai/	[44]
Xpanse	https://xpanse.ai/	[49]