Tutorial

Democratizing Artificial Intelligence Imaging Analysis With Automated Machine Learning: Tutorial

Arun James Thirunavukarasu[1,2], BA, MB BChir; Kabilan Elangovan[2], BEng; Laura Gutierrez[2], MD; Yong Li[2], MD; Iris Tan[2], BEng; Pearse A Keane[3], MD; Edward Korot[4,5], MD; Daniel Shu Wei Ting[2,4,6], PhD

[1]University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom

[2]Artificial Intelligence and Digital Innovation Research Group, Singapore Eye Research Institute, Singapore, Singapore

[3]Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom

[4]Byers Eye Institute, Stanford University, Palo Alto, CA, United States

[5]Retina Specialists of Michigan, Grand Rapids, MI, United States

[6]Singapore National Eye Centre, Singapore, Singapore

Corresponding Author:
Arun James Thirunavukarasu, BA, MB BChir
University of Cambridge School of Clinical Medicine
Addenbrooke's Hospital
Hills Rd
Cambridge, CB2 0SP
United Kingdom
Phone: 44 01223 336700
Email: ajt205@cantab.ac.uk

Abstract

Deep learning–based clinical imaging analysis underlies diagnostic artificial intelligence (AI) models, which can match or even exceed the performance of clinical experts, having the potential to revolutionize clinical practice. A wide variety of automated machine learning (autoML) platforms lower the technical barrier to entry to deep learning, extending AI capabilities to clinicians with limited technical expertise, and even autonomous foundation models such as multimodal large language models. Here, we provide a technical overview of autoML with descriptions of how autoML may be applied in education, research, and clinical practice. Each stage of the process of conducting an autoML project is outlined, with an emphasis on ethical and technical best practices. Specifically, data acquisition, data partitioning, model training, model validation, analysis, and model deployment are considered. The strengths and limitations of available code-free, code-minimal, and code-intensive autoML platforms are considered. AutoML has great potential to democratize AI in medicine, improving AI literacy by enabling "hands-on" education. AutoML may serve as a useful adjunct in research by facilitating rapid testing and benchmarking before significant computational resources are committed. AutoML may also be applied in clinical contexts, provided regulatory requirements are met. The abstraction by autoML of arduous aspects of AI engineering promotes prioritization of data set curation, supporting the transition from conventional model-driven approaches to data-centric development. To fulfill its potential, clinicians must be educated on how to apply these technologies ethically, rigorously, and effectively; this tutorial represents a comprehensive summary of relevant considerations.

(J Med Internet Res 2023;25:e49949) doi: 10.2196/49949

KEYWORDS

machine learning; automated machine learning; autoML; artificial intelligence; democratization; autonomous AI; imaging; image analysis; automation; AI engineering

Introduction

Automated machine learning (autoML) is the product of attempts to broaden artificial intelligence (AI) engineering capability beyond those with technical and computational expertise [1]. Machine learning (ML) is a form of AI that describes the computational process of leveraging data to improve performance in a defined task, thereby developing sophisticated models without explicit programming. More recently, deep learning (DL) has emerged as a powerful form of ML capable of interpreting unstructured data, such as images, language, and speech [2,3]. In DL, layers of representation are

developed that iteratively manipulate input data until useful features emerge, permitting the processing of highly complicated data sets. These layers are composed of tuned artificial neurons; computationally encoded mathematical functions that together comprise a deep neural network. Results across medicine have been impressive, with the production of many models with expert or beyond-expert accuracy, sensitivity, and specificity [4]. AutoML acts to extend automation even further through various aspects of algorithm development, including hyperparameter optimization and neural architecture search [1].

Many autoML platforms have been developed in industry and academia, with recent innovation producing platforms capable of DL, compatible with unstructured input data such as medical images (Figure 1) [5,6]. These platforms, with different requirements, capabilities, and limitations, may be categorized based on the spectrum of user coding requirements (Table 1). To capitalize on the potential of AI, more users must be able to harness DL and other ML techniques, leveraging the health care data that continues to be accrued at an accelerating rate [7]. This reduction in the requirement for expertise and computational requirements constitutes the "democratization" of AI technology [5,8]. Democratization refers to the broadening of access to technology conferred by reduced technical and hardware requirements.

When effectively deployed, DL has the potential to improve patient safety, quality of health care, and cost-effectiveness [9-11]. These improvements are based on the accuracy, speed, and reproducibility of DL algorithms, which can exceed that of humans with extensive training [9,10]. Accurate and reliable computational models may complement skilled human assessments as a part of novel systems with equivalent or superior performance to conventional practice with the additional benefit of being less expensive [11]. Successful projects benefit from interdisciplinary collaboration, with clinical and technical expertise brought to bear [12]. Clinicians, computer scientists, and data scientists work together, and their time and resources are scarce [13-15]. Collaboration also introduces complications regarding communication, particularly where individuals' expertise differs, and regarding sharing patient data for which privacy is closely governed [16]. Increasing the accessibility of high-performance DL for clinicians with autoML may ameliorate these issues [6]. Through the democratization of DL, a greater number of AI-literate clinicians can contribute to the research, implementation, and governance of these systems [17]. Moreover, emerging AI models with the capability to leverage application programming interfaces as tools could facilitate AI building itself at a rapidly accelerating rate; early examples include GPT-4, PaLM 2, and LLaMA 2 [18].

Below, we show how to use autoML for medical image analysis for clinicians and other interested allied health care professionals, with step-by-step illustrations of workflow, important considerations, and requirements. The strengths and weaknesses of code-free, code-minimal, and code-intensive platforms are discussed, in addition to the capabilities and limitations of each platform. The potential use cases of autoML in education, primary research, and clinical practice are outlined. The technical and ethical best practices are emphasized throughout to encourage maintained or even improved standards of development and reporting, as a broader subset of clinicians gain access to AI and as developers look to adopt a data-centric approach to constructing models [19].

Figure 1. Relationship between autoML, deep learning, machine learning, and artificial intelligence. autoML: automated machine learning.

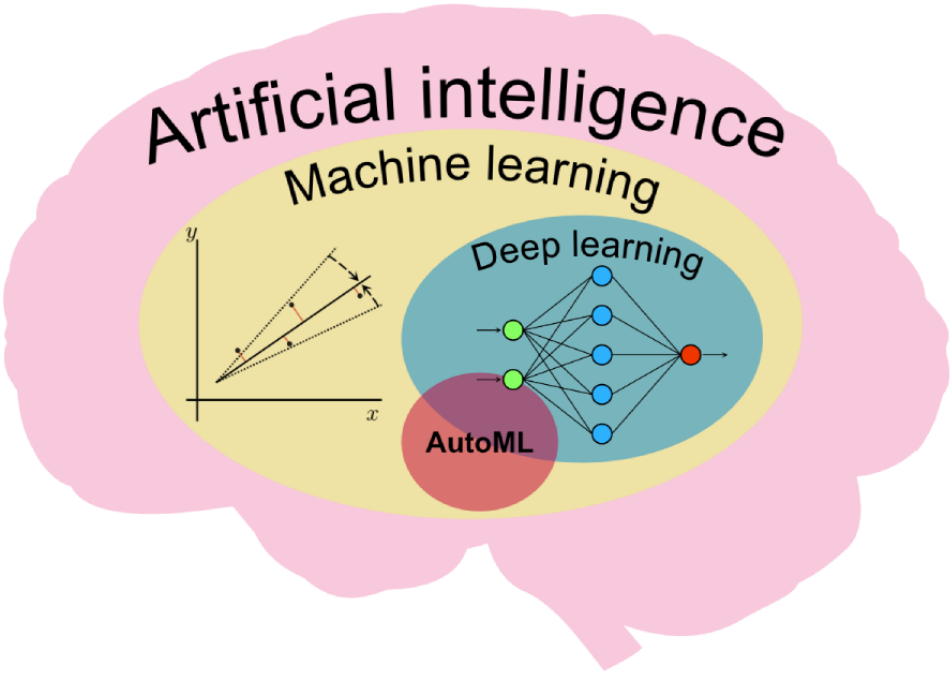


Table 1. Summary of autoML platforms facilitating deep learning for images, with an appraisal of accessibility and portability.

AutoML[a] platform	Accessibility			Portability	
	Cost	Code requirement	Computing location	Exportability	Explainability
Amazon recognition	Chargeable	None	Cloud	No	No
Apple create ML[b]	Free on specific devices	None	Local	To Apple devices	No
Auto-PyTorch	Free	Coding required	Local	Yes	No
AutoGluon	Free	Coding required	Local	Yes	Yes
AutoGOAL	Free	Coding required	Local	Yes	No
AutoKeras	Free	Coding required	Local	Yes	No
Baidu EasyDL	Chargeable	None	Local or cloud	To edge devices	No
Clarifai	Free (chargeable features)	None	Cloud	To edge devices	No
Google Cloud AutoML Vision	Chargeable	None	Cloud	To edge devices	Yes
Huawei ExeML	Chargeable	None	Cloud	No	No
H2O.ai Hydrogen torch	Chargeable	None	Local or cloud	Yes	Yes
H2O R/Python packages	Free	Coding required	Local	Yes	Yes
H2O.ai Driverless Al	Chargeable	None	Local or cloud	Yes	Yes
KNIME	Free (chargeable features)	None	Local	Yes	No
MATLAB	Chargeable	Coding required	Local	Yes	No
MedicMind	Free (chargeable features)	None	Cloud	Yes	Yes
Microsoft Azure AutoML	Chargeable	None	Cloud	To edge devices	No
Neuro-T	Chargeable	None	Local	Yes	No
Sony prediction one	Chargeable	None	Local or cloud	Yes	Yes

[a]AutoML: automated machine learning.

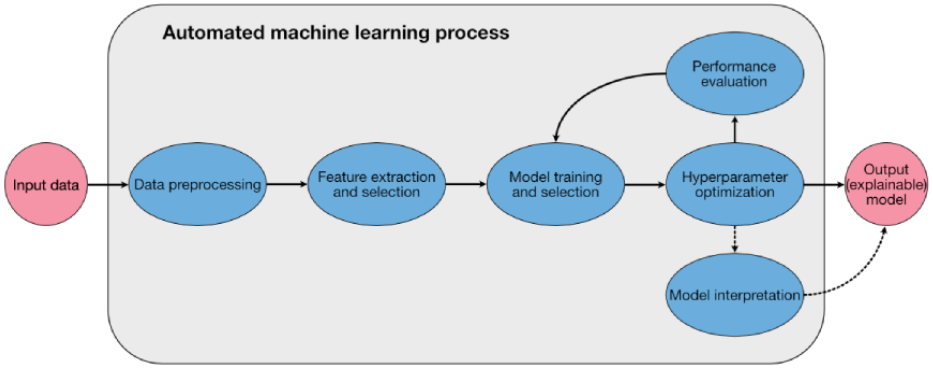
[b]ML: machine learning.

Technical Overview

In general, autoML technology executes part or all of the ML engineering process without users' input (Figure 2). Without autoML, these tasks require skilled data or computer scientists. Through a process of trial and error, informed by prior experience, these experts attempt to find an optimal neural network structure and hyperparameters to solve clinical problems, such as disease diagnosis, treatment planning, or prognosis prediction. AutoML has been applied primarily to classification tasks thus far, where an algorithm seeks to correctly identify ("classify") images exhibiting one of a defined set of potential conditions or diseases ("classes") [5,6]. A wide variety of ML algorithms such as k nearest neighbors, support vector machine, random forest, neural network, naive Bayes, and logistic regression exist for classification, from which an autoML platform may select depending on comparative performance [20]. This is an example of supervised learning, as input data must be labeled by defined classes. In contrast, autoML for unsupervised and reinforcement learning is relatively nascent [21,22].

To achieve performance comparable to bespoke ML models trained by computer scientists, autoML platforms use a variety of methods and optimization techniques including Bayesian optimization, random search, grid search, evolutionary-based neural architecture selection, and meta-learning [23]. An optimal model may then be outputted for internal or external validation, interpretation, and deployment. Many platforms, with various accessibility, technical features, and portability, have been developed in academia and industry. When deciding which platform to use, researchers and clinicians should consider their capabilities, requirements, and aims.

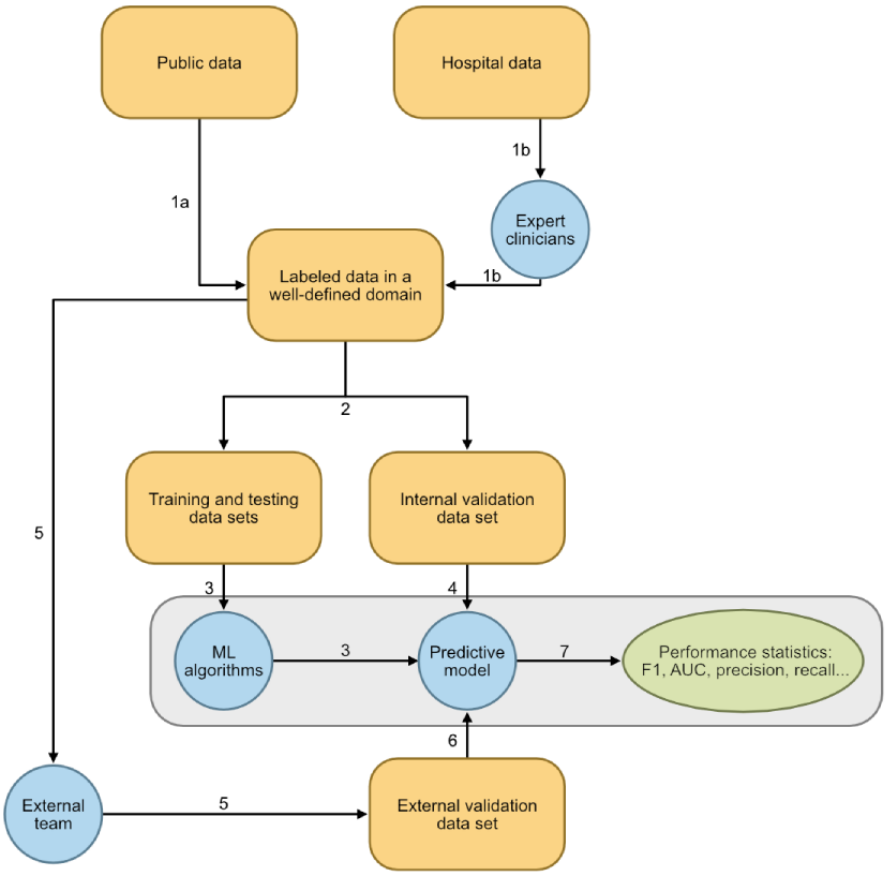
Figure 2. A conceptual diagram of automated machine learning, which may generate a predictive model from input data in the form of medical images. Data preprocessing entails the processing of inputs to augment and simplify the data, and "clean" data into a compatible format. Feature extraction involves the identification of the elements of the input data, which provide the most discriminative power. Model selection, training, and optimization summarize the process of training a myriad of potential deep learning architectures, selecting the best-performing architecture, and optimizing hyperparameters such as time to train or the number of iterations using training data. To judge which model is optimal, and to report its effectiveness, performance evaluation is required. Some automated machine learning platforms facilitate interpretation, allowing the deduction of how decisions are reached.



Workflow

The process of applying DL to medical image analysis generally involves gathering high-quality data, training a model, and evaluating its performance; the reporting of these processes has been standardized [24]. The process of applying autoML is comparable, despite being less technically demanding, and still relies on careful selection and labeling of representative data sets for the designated use case (Figure 3). These data-driven principles are applicable across different imaging modalities, including X-ray, ultrasound, computed tomography, magnetic resonance imaging, optical coherence tomography, fundus photography, and angiography [25]. The algorithms can be trained very quickly using autoML with expert or even supraexpert ability to classify medical images [6].

Figure 3. The process of developing a deep learning model with automated ML. Minimally expected processes facilitated by an automated ML platform are within the gray-shaded region. However, platforms variably assist with other processes. 1A: public data set curation; 1B: private data set curation; 2: partitioning; 3: training; 4: internal validation; 5: curation of independent data set; 6: external validation; 7: performance evaluation and results presentation. AUC: area under the curve; ML: machine learning.



Data Set Curation

Data Management

For trained algorithms to be appropriate for clinical deployment, it is helpful to use data in the same format as in routine clinical practice. This may be complicated where different formats exist within a single imaging modality, such as from different machines or scanning protocols; an ML model may struggle to classify images due to insufficient salient clinical features relative to structural differences in diverse input. Converting images to a common format may involve manual formatting, down-sampling resolution, or labeling classes with folders or file names. Larger data sets are often assumed to be less susceptible to weaknesses but are not a solution in themselves [26]. In some cases, larger data sets lead to worse performance in a classification task, as seen in published studies attempting to distinguish cognitively normal, mild cognitive impairment, and patients with Alzheimer disease [26]. However, the sample size must be sufficient to represent the general population to which algorithms may be applied, and requirements are more stringent where disease features vary widely or where differences between patients are subtle.

Publicly Available Data (Step 1A)

Many data sets curated to support ML research exist, covering a wide variety of imaging modalities and clinical diagnoses [27,28]. These permit researchers to download and use data to train and test ML models including autoML, but attention to specific permissions is required to avoid breaching privacy or copyright regulations. Data sets often make specific requests of researchers using their files, such as citing a source study in any resultant published work. Furthermore, they often have licensing limits for the development of commercial algorithms derived from the posted data. Where privacy permissions are restricted, robust data security is essential to ensure patient information is not shared with unauthorized parties.

Public data sets are often prelabeled according to a condition of interest. For researchers, it is essential to determine the source of labels and adjudge whether the labeling process is suitably rigorous to use for their specific project, and if the "ground truth" (data set labels that are assumed to be true during training) is sufficiently accurate. If quality standards are not met, such as if labeling was conducted by insufficiently skilled clinicians, if there was no arbitration or adjudication process, or if mistakes are apparent, researchers need to arrange relabeling or correct the wrong labels themselves. By improving the quality of input data, algorithm performance may be significantly stronger, with fewer errors expected [29].

Private Data (Step 1B)

Private data sets may be curated by researchers as part of an autoML project. This requires institutional review board ethical approval, as well as informed consent from patients to use their data in an explicitly defined research context. As discussed above, data security is of paramount importance to ensure consent is not breached regarding the persons given access to patients' data. Sharing data with commercial autoML platforms may be prohibited, necessitating the use of local autoML platforms, such as H2O, Apple CreateML, or autoKeras. Deidentification of data may be a requirement to obtain research ethical approval, retaining only essential elements to facilitate labeling and classification [30].

Labeling should be undertaken by the research team. While certain autoML platforms assist with labeling through active learning, these are often cloud-based services that may require additional institutional approval before use. Labeling may be based on incontrovertible pathologic, genomic, or clinical outcome ground truth or a less accurate approximation based on annotations by expert clinicians, preferably conducted prospectively as clinicians may thereby incorporate contemporaneous clinical and laboratory data into their decision-making. Rigorous arbitration and adjudication are required to minimize labeling errors, which may otherwise significantly impair model performance [30,31].

Partitioning (Step 2)

Data partitioning is a critical preliminary step when building any ML model in order to evaluate model performance fairly on an unseen, representative data set independent from the images used to train the model. Data must be split into training and testing partitions, and there is a wide range of partitioning algorithms with different strengths and limitations summarized as round-robin, hash, range, and random schemes [32]. As a rule of thumb, 80% of available data may be used for training, with 20% used for testing (internal validation). Most autoML platforms facilitate the upload of separate partitions corresponding to training and testing; otherwise, the platforms themselves split data accordingly. Automatic partitioning with obtuse algorithms results in training and testing data sets that may not be known to the user, precluding the establishment of representativeness of these data sets. Therefore, it is best practice to manually partition data sets with patient-level splits, using reproducible and documented methods.

Training and Internal Validation (Steps 3 and 4)

Reliance on a single data set (especially if small) may result in "overfitting"—where algorithms learn features specific to the images in the training data set only [33]. While performance may be exceptional on the training data set, it is weaker when algorithms are applied to unseen data. To avoid this trap, a small subset of data must be reserved for validation, which acts as a means of observing model performance at each training iteration to guide the process and adjudge when algorithms have been optimally entrained. Separation is key, as algorithms are expected to perform accurately with images previously "seen"; the training data set has features identical to those "learned" by a model to be associated with classification labels. By using a data set entirely separate from the training process, a fairer evaluation of model performance may be obtained and summarized with an array of statistical metrics (discussed in step 7). In practice, ML models improve accuracy on this internal validation data set as the primary indicator of successful training: the final algorithm corresponds to that which performs best on the internal validation data set, where classification accuracy is maximized without overfitting to the training data. Many autoML platforms use cross-validation, where multiple partitioning processes are applied to generate separate testing and training data sets. In this case, model selection is based on

optimal performance across iterations over all the partitioned data sets, further reducing the risk of overfitting [33].

Curation of an Independent Data Set (Step 5)

To externally validate a predictive model, data entirely separated from the training and testing process must be used. External validation demonstrates true generalizability if performance is acceptable with diverse data sets, representative of future cohorts where models may be used with a range of idiosyncratic differences inherent in obtaining images in different clinical environments. While the curation process is identical to that described in steps 1A and 1B, there exists an additional option to collect data prospectively to facilitate a robust analysis of model performance in clinical conditions, with a lower risk of bias [34]. This entails obtaining ethical approval and may require patient consent, as well as the time and clinical training required to collect suitable data. Conducting a prospective pragmatic trial would represent the strongest form of primary evidence for justifying deployment in clinical settings in the future.

External Validation (Step 6)

Using an independent data set on the same model is important to demonstrate generalizability beyond the restricted data used to initially train and test the autoML algorithm. This is most conveniently executed through batch prediction of an external validation data set, but platforms may restrict processing to single images or prohibit the export or deployment of a model without extra costs [6]. While external validation is preferably undertaken by a separate research team to avoid a potential source of bias, initial validation may be done by the same team to improve the veracity of their performance claims [35,36]. If open-source data sets are used, performance with as many data sets as possible should be reported to avoid selection bias resulting from cherry-picking of data where performance is higher; this may be due to the external validation data set being more similar to the training and internal validation data set.

Performance Evaluation and Results Interpretation (Step 7)

Model performance metrics and visualizations are the most important features to users in terms of developing trust in an autoML platform [37]. Many metrics are used in ML research; some are a function of prevalence such as accuracy, area under the precision-recall curve (PRC), and F1 score, whereas others are a function of the model threshold, such as accuracy and F1 score [38]. Threshold refers to the cutoff point of prediction probability above which the model gives one output or another, governing sensitivity and specificity. Many autoML platforms provide just a few performance metrics, in part due to displaying results from an "optimal" model operating at a single threshold, which prohibits calculation of metrics such as area under the precision-recall curve or area under the receiver operating characteristic curve (ROC), and PRC and ROC plots cannot be produced without implementing the model at a range of thresholds. With only a snapshot of performance statistics at one model threshold, it is possible that apparent performance is inflated by condition prevalence or model hypersensitivity. In addition, threshold customizability increases the likelihood of a model being clinically useful. While a particular performance metric may be maximized at 1 threshold, there may be a requirement for tuning, such as to optimize the sensitivity or specificity. Providing more metrics may allow fairer comparison to alternative computational techniques and expert clinician performance, and confusion matrices are an essential tool to judge the use of a model's performance. Certain platforms do provide the customizability to generate PRC and ROC plots, but these often have a greater requirement for coding.

Explainability is an ML research priority due to concerns over delegating responsibility to "black box" models. By understanding how models make successful predictions, the potential risks of delegating decision-making to systems with occult biases are avoided [39]. Clinicians and patients may have more confidence in the so-called explainable AI. The availability of transparency features on autoML platforms is recognized as a key aspect of users' trust and understanding when using these tools [37]. Some platforms have inbuilt explainability features; examples include H2O.ai Driverless AI and Google Cloud AutoML, which provide Grad-CAM and XRAI-derived saliency maps depicting which parts of an image contributed to classification [40]. However, these tools leave an "interpretability gap," which can lead to misleading conclusions [41]. Further explainable AI innovation is required, but this is complicated with platforms that do not facilitate model export and deployment on new batches of data, as with external validation. New tools are being developed to facilitate the interpretation of ML and even autoML models less amenable to export; examples include the What-If Tool which facilitates counterfactual analysis of model performance and individual classification decisions as input data are altered [42].

Capabilities and Limitations: Platform Comparison

The wide variety of autoML platforms offers different capabilities and limitations. In general, platforms may be discussed in terms of their requirement for coding ability, with code-free, code-minimal, and code-intensive examples (Table 1). Platforms may also be parsed by the location of data and processing as either cloud-based or local. Cloud-based solutions may be more secure than local solutions due to industry-standard encryption and International Organization for Standardization compliance audits but require explicit ethical approval to be used with sensitive patient data. Without ethical approval, using cloud-based autoML is limited to open-source data sets, which are now abundant but often lacking in terms of quality labeling and representative populations [27,28]. While local platforms may be preferred for their tendency not to require payment for access, hardware requirements may be prohibitive and security protocols must be sufficiently robust, limiting their role in democratizing AI.

Technical features correspond to the development process outlined in Figure 2. The so-called "end-to-end" platforms may be defined as those that automate all these processes; all users are required to do is input labeled data. Most platforms equipped for DL have end-to-end functionality, operating without a requirement for user input. Platforms differ in their permissiveness of model export for further validation, explainability analysis, and potential deployment. In general,

local platforms always facilitate the export of models amenable to explainability analysis, external validation, and deployment at scale. Cloud-based platforms are generally more of a "black box," offering no details as to the model architecture, but some enable model export and batch prediction to facilitate external validation and explainability analysis—often in return for a fee.

Use Cases of AutoML

Medical Education

AutoML for medical image analysis can be an educational tool for clinicians and medical students. By lowering the requirement for coding expertise or GPU access, autoML permits more learners to explore ideas practically rather than merely discussing them in theory. Learners can actively produce and modify models to demonstrate the importance of data set quality, validation, and explainability for themselves [43]. This may also provide learners with the intuition of an ML developer sooner—conferring greater practical expertise than a mere understanding of theory alone when approaching new problems. As autoML promotes interaction with data over coding, further learning through institutional courses or individual initiatives is required to develop the necessary ML expertise to engineer bespoke, fully customizable models. However, focusing on data may best prepare clinicians for future trends in ML development and promote mechanical understanding of algorithms rather than learning tricks to maximize performance.

A transition from model-driven to data-driven techniques—incubated and facilitated by autoML—has been discussed as a means of accelerating development, as arduous engineering is bypassed. This complements the recent drive to inculcate "data-centric AI" (spearheaded by Andrew Ng), where data set curation is focused on rather than optimizing code. As the supply of high-quality data is more often the limiting factor in development than code or model infrastructure, future innovation is likely to focus on the generation and aggregation of training data [19]. The approach is a nascent paradigm in medicine, although promising results have begun to emerge.

Medical Research

Another main use case for autoML is medical image-based primary research, including pilot studies and larger-scale projects. AutoML pilot studies, with relatively low costs in terms of time and money, may be used to gauge whether a research question can be solved with AI and ML techniques. AutoML enables clinicians to perform initial proof of concept studies with private and well-labeled data, generating initial outcomes without a requirement for collaboration with computational experts, enabling optimization of research resource allocation. For example, autoML may help determine whether a certain sample size or quality level of images is useful and practical for developing classification models. This is an alternative to haphazard trial and error, which wastes expertise, time, and resources as infeasibility is determined at a later step. External technical collaborators may be approached with promising interim results generated by autoML, increasing the likelihood of a proposed project succeeding. Applications for research funding may be strengthened by promising pilot study results generated with autoML.

AutoML may also be used independently in primary research. However, to apply autoML in medical image analysis, there are rigorous academic requirements, including standardization of benchmarks, ensuring reproducibility in analysis and the interpretability of outcomes, and following guidelines for research and reporting, such as for Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence (DECIDE-AI) [24]. The technical limitations of certain autoML platforms make adherence to these standards more challenging, but useful results have been and will continue to be produced with autoML technology.

Clinical Deployment

As with all AI research in health care, models are hoped to improve patient care in real-world settings [44]. As studies have demonstrated the performance of models generated by autoML to be comparable or even higher than those generated with conventional AI techniques, there is a basis for exploring the clinical deployment of autoML. Although autoML often restricts explainability and interpretability—discussed above—many Food and Drug Administration-approved models have no explainability [45]. However, to be deployed in direct clinical care, autoML models must meet the same regulatory standards as conventionally developed AI applications. These standards are evolving around the world but involve extensive evaluation and validation and are often expensive application processes to regulatory bodies. For regulatory clearance, emphasis is placed on the intended clinical use; this must be defined precisely to ensure deployment as desired is permissible if an application is accepted. Requirements include demonstration of data traceability and quality, so careful documentation is required to preserve the source of data and ground truth. Model versions must be documented with the data used at each step; only the version accepted by a regulatory body may be deployed. Additional considerations include governance systems, adherence to software development cycle requirements and International Organization for Standardization regulations, integration into clinical workflow, and cybersecurity.

Conclusions

AutoML is an exciting innovation that reduces the barrier to entry for AI development, including DL for medical image classification. With the democratization of AI, it is hoped that the quality and acceptance of AI innovations for patient diagnosis, management, and prognosis will improve, accelerating computational innovation in clinical practice. Specifically, empowering nonspecialists to harness DL technology may enable clinician-driven AI, allowing experts with knowledge of domain-specific pain points to take a more active role in the development of applicable, effective, and useful new tools. The technical limitations of autoML are reducing as corporate and academic developers continue to improve available platforms, although conventional techniques have an edge in terms of capability, customizability, and explainability. This currently limits the potential of autoML, particularly in younger, developing subfields such as multimodal AI and autonomous foundation models [46,47]. Nevertheless, autoML represents an excellent tool for interested clinicians to

develop DL skills, conduct pilot studies and other research, and produce models to improve clinical practice.

Acknowledgments

The authors extend their thanks to Timing Liu for his insights into automated machine learning and artificial intelligence more broadly. AJT is supported by The Royal College of Surgeons in Edinburgh (RCSED Bursary 2022), the Royal College of Physicians (MSEB 2022), and Corpus Christi College, University of Cambridge (Gordon Award 1083874682). DSWT is supported by the National Medical Research Council, Singapore (NMCR/HSRG/0087/2018; MOH-000655-00; MOH-001014-00); Duke-NUS Medical School (Duke-NUS/RSF/2021/0018; 05/FY2020/EX/15-A58); and Agency for Science, Technology and Research (A20H4g2141; H20C6a0032). These funders were not involved in the conception, execution, or reporting of this study.

Data Availability

All data generated or analyzed for the work presented here are included in this published article.

Authors' Contributions

AJT conceived and led the project. DSWT provided academic supervision. LG, KE, YL, PAK, EK, and DSWT provided technical and clinical information and assisted with drafting the manuscript. IT provided regulatory insight and assisted with drafting the manuscript. AJT, EK, KE, and YL verified the features of autoML platforms. AJT and KE produced figures. All authors approved the final draft for submission.

Conflicts of Interest

PAK has acted as a consultant for DeepMind, Roche, Novartis, Apellis and BitFount, and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan and Bayer. EK has acted as a consultant for Genentech and Google Health, and is an equity owner in Reti Health.

References

1. Yao Q, Wang M, Chen Y, Dai W, Li YF, Tu WW, et al. Taking human out of learning applications: a survey on automated machine learning. ArXiv Preprint posted online on December 16, 2019 [FREE Full text] [doi: 10.48550/arXiv.1810.13306]
2. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24-29 [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-444 [doi: 10.1038/nature14539] [Medline: 26017442]
4. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021;4(1):65 [FREE Full text] [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
5. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health 2019;1(5):e232-e242 [FREE Full text] [doi: 10.1016/S2589-7500(19)30108-6] [Medline: 33323271]
6. Korot E, Guan Z, Ferraz D, Wagner SK, Zhang G, Liu X, et al. Code-free deep learning for multi-modality medical image classification. Nat Mach Intell 2021;3(4):288-298 [FREE Full text] [doi: 10.1038/s42256-021-00305-2]
7. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
8. Korot E, Gonçalves MB, Khan SM, Struyven R, Wagner SK, Keane PA. Clinician-driven artificial intelligence in ophthalmology: resources enabling democratization. Curr Opin Ophthalmol 2021;32(5):445-451 [doi: 10.1097/ICU.0000000000000785] [Medline: 34265784]
9. Bash S, Johnson B, Gibbs W, Zhang T, Shankaranarayanan A, Tanenbaum LN. Deep learning image processing enables 40% faster spinal MR scans which match or exceed quality of standard of care: a prospective multicenter multireader study. Clin Neuroradiol 2022;32(1):197-203 [FREE Full text] [doi: 10.1007/s00062-021-01121-2] [Medline: 34846555]
10. Kwon JM, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. J Am Heart Assoc 2018;7(13):e008678 [FREE Full text] [doi: 10.1161/JAHA.118.008678] [Medline: 29945914]
11. Xie Y, Nguyen QD, Hamzah H, Lim G, Bellemo V, Gunasekeran DV, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. Lancet Digit Health 2020;2(5):e240-e249 [FREE Full text] [doi: 10.1016/S2589-7500(20)30060-1] [Medline: 33328056]
12. Xu J, Xue K, Zhang K. Current status and future trends of clinical diagnoses via image-based deep learning. Theranostics 2019;9(25):7556-7565 [FREE Full text] [doi: 10.7150/thno.38065] [Medline: 31695786]

13. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: towards an EU action plan. Genome Med 2016;8(1):71 [FREE Full text] [doi: 10.1186/s13073-016-0323-y] [Medline: 27338147]
14. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317-1318 [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
15. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347-1358 [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
16. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25(1):37-43 [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
17. Keane PA, Topol EJ. AI-facilitated health care requires education of clinicians. Lancet 2021;397(10281):1254 [doi: 10.1016/S0140-6736(21)00722-4] [Medline: 33812482]
18. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940 [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
19. Chang EY. Knowledge-guided data-centric AI in healthcare: progress, shortcomings, and future directions. ArXiv Preprint posted online on April 30, 2023 [FREE Full text] [doi: 10.48550/arXiv.2212.13591]
20. Yuvalı M, Yaman B, Tosun Ö. Classification comparison of machine learning algorithms using two independent CAD datasets. Mathematics 2022;10(3):311 [FREE Full text] [doi: 10.3390/math10030311]
21. Parker-Holder J, Rajan R, Song X, Biedenkapp A, Miao Y, Eimer T, et al. Automated reinforcement learning (AutoRL): a survey and open problems. J Artif Intell Res 2022;74:517-568 [doi: 10.1613/jair.1.13596]
22. Poulakis Y, Doulkeridis C, Kyriazis D. AutoClust: a framework for automated clustering based on cluster validity indices. Manhattan, New York City: IEEE; 2020 Presented at: 2020 IEEE International Conference on Data Mining (ICDM); November 17-20, 2020; Sorrento, Italy p. 1220-1225 URL: https://ieeexplore.ieee.org/document/9338346 [doi: 10.1109/icdm50108.2020.00153]
23. Hutter F, Kotthoff L, Vanschoren J, editors. Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer International Publishing; 2019.
24. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924-933 [FREE Full text] [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
25. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
26. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 2022;5(1):48 [FREE Full text] [doi: 10.1038/s41746-022-00592-y] [Medline: 35413988]
27. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit Health 2021;3(1):e51-e66 [FREE Full text] [doi: 10.1016/S2589-7500(20)30240-5] [Medline: 33735069]
28. Wen D, Khan SM, Xu AJ, Ibrahim H, Smith L, Caballero J, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit Health 2022;4(1):e64-e74 [FREE Full text] [doi: 10.1016/S2589-7500(21)00252-1] [Medline: 34772649]
29. Jain A, Patel H, Nagalapatti L, Gupta N, Mehta S, Guttula S, et al. Overview and importance of data quality for machine learning tasks. New York, NY, United States: Association for Computing Machinery; 2020 Presented at: KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; July 6-10, 2020; CA USA p. 3561-3562 URL: https://dl.acm.org/doi/proceedings/10.1145/3394486 [doi: 10.1145/3394486.3406477]
30. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology 2020;295(1):4-15 [FREE Full text] [doi: 10.1148/radiol.2020192224] [Medline: 32068507]
31. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125(8):1264-1272 [doi: 10.1016/j.ophtha.2018.01.034]
32. Mahmud MS, Huang JZ, Salloum S, Emara TZ, Sadatdiynov K. A survey of data partitioning and sampling methods to support big data analysis. Big Data Min Anal 2020;3(2):85-101 [FREE Full text] [doi: 10.26599/bdma.2019.9020015]
33. Ying X. An overview of overfitting and its solutions. J Phys Conf Ser 2019;1168:022022 [FREE Full text] [doi: 10.1088/1742-6596/1168/2/022022]
34. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
35. Kuehn BM. Striving for a more perfect peer review: editors confront strengths, flaws of biomedical literature. JAMA 2013;310(17):1781-1783 [doi: 10.1001/jama.2013.280660] [Medline: 24193063]

36. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40 [FREE Full text] [doi: 10.1186/1471-2288-14-40] [Medline: 24645774]
37. Drozdal J, Weisz J, Wang D, Dass G, Yao B, Zhao C, et al. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. New York, NY, United States: Association for Computing Machinery; 2020 Presented at: IUI '20: Proceedings of the 25th International Conference on Intelligent User Interfaces; March 17-20, 2020; Cagliari Italy p. 297-307 URL: https://dl.acm.org/doi/proceedings/10.1145/3377325 [doi: 10.1145/3377325.3377501]
38. Erickson BJ, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. Radiol Artif Intell 2021;3(3):e200126 [FREE Full text] [doi: 10.1148/ryai.2021200126] [Medline: 34136815]
39. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022;79:102470 [FREE Full text] [doi: 10.1016/j.media.2022.102470] [Medline: 35576821]
40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2019;128(2):336-359 [doi: 10.1007/s11263-019-01228-7]
41. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745-e750 [FREE Full text] [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
42. Abbas A, O'Byrne C, Fu DJ, Moraes G, Balaskas K, Struyven R, et al. Evaluating an automated machine learning model that predicts visual acuity outcomes in patients with neovascular age-related macular degeneration. Graefes Arch Clin Exp Ophthalmol 2022;260(8):2461-2473 [FREE Full text] [doi: 10.1007/s00417-021-05544-y] [Medline: 35122132]
43. Xie Y, Chen M, Kao D, Gao G, Chen XA. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. New York, NY, United States: Association for Computing Machinery; 2020 Presented at: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; April 25-30, 2020; Honolulu, HI, USA p. 1-13 URL: https://dl.acm.org/doi/proceedings/10.1145/3313831 [doi: 10.1145/3313831.3376807]
44. Tan TF, Thirunavukarasu AJ, Jin L, Lim J, Poh S, Teo ZL, et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. Lancet Glob Health 2023;11(9):e1432-e1443 [FREE Full text] [doi: 10.1016/S2214-109X(23)00323-6] [Medline: 37591589]
45. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food and Drug Administration. 2022. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2023-08-02]
46. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. J R Soc Med 2023;116(5):181-182 [FREE Full text] [doi: 10.1177/01410768231173123] [Medline: 37199678]
47. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265 [FREE Full text] [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]

Abbreviations

AI: artificial intelligence
autoML: automated machine learning
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence
DL: deep learning
ML: machine learning
PRC: precision-recall curve
ROC: receiver operating characteristic

Edited by A Mavragani; submitted 14.06.23; peer-reviewed by YK Suh, A Lemonard, M Wosny, T Nguyen; comments to author 09.08.23; revised version received 21.08.23; accepted 13.09.23; published 12.10.23

Please cite as:
Thirunavukarasu AJ, Elangovan K, Gutierrez L, Li Y, Tan I, Keane PA, Korot E, Ting DSW
Democratizing Artificial Intelligence Imaging Analysis With Automated Machine Learning: Tutorial
J Med Internet Res 2023;25:e49949
URL: https://www.jmir.org/2023/1/e49949
doi: 10.2196/49949
PMID: 37824185