Title: Clinical performance of automated machine learning: a systematic review

Running head: Clinical performance of automated machine learning

Authors: Arun James Thirunavukarasu[1,2,*] MB BChir, Kabilan Elangovan[1] BEng, Laura Gutierrez[1] MD, Refaat Hassan[2] MB BChir, Yong Li[1,3] MD, Ting Fang Tan[1] MBBS, Haoran Cheng[1,3,4] MPH, Zhen Ling Teo[5] FRCOphth, Gilbert Lim[1] PhD, Daniel Shu Wei Ting[1,3,5,*] PhD

Affiliations:

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

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

[3]Duke-NUS Medical School, National University of Singapore, Singapore

[4]Rollins School of Public Health, Emory University, Atlanta, Georgia, United States of America

[5]Singapore National Eye Centre, Singapore

***Corresponding author details**:
A/Prof Daniel Ting MD (1st Hons) PhD
Associate Professor, Duke-NUS Medical School
Director, AI Office, Singapore Health Service
Address: Singapore Eye Research Institute (SERI), The Academia, 20 College Road, Level 6 Discovery Tower, Singapore, 169856
Email: daniel.ting@duke-nus.edu.sg

Dr Arun Thirunavukarasu BA MB BChir
Academic Foundation Doctor, Oxford University Clinical Academic Graduate School
Email: ajt205@cantab.ac.uk

Abstract:

Introduction

Automated machine learning (autoML) removes technical and technological barriers to building artificial intelligence models. We aimed to summarise the clinical applications of autoML, assess the capabilities of utilised platforms, evaluate the quality of the evidence trialling autoML, and gauge the performance of autoML platforms relative to conventionally developed models, as well as each other.

Methods

This review adhered to a PROSPERO-registered protocol (CRD42022344427). The Cochrane Library, Embase, MEDLINE, and Scopus were searched from inception to 11 July 2022. Two researchers screened abstracts and full texts, extracted data and conducted quality assessment. Disagreement was resolved through discussion and as-required arbitration by a third researcher.

Results

In 82 studies, 26 distinct autoML platforms featured. Brain and lung disease were the most common fields of study of 22 specialties. AutoML exhibited variable performance: AUCROC 0.35-1.00, F1-score 0.16-0.99, AUCPR 0.51-1.00. AutoML exhibited the highest AUCROC in 75.6% trials; the highest F1-score in 42.3% trials; and the highest AUCPRC in 83.3% trials. In autoML platform comparisons, AutoPrognosis and Amazon Rekognition performed strongest with unstructured and structured data respectively. Quality of reporting was poor, with a median DECIDE-AI score of 14 of 27.

Conclusions

A myriad of autoML platforms have been applied in a variety of clinical contexts. The

performance of autoML compares well to bespoke computational and clinical benchmarks. Further work is required to improve the quality of validation studies. AutoML may facilitate a transition to data-centric development, and integration with large language models may enable AI to build itself to fulfil user-defined goals.

Keywords: artificial intelligence, automated machine learning; AI; autoML; machine learning; deep learning;

Introduction

In medicine, machine learning (ML) has been applied in a wide variety of contexts ranging from administration to clinical decision support, driven by greater availability of healthcare data and technological development (1–5). Automated machine learning (autoML) enables individuals without extensive computational expertise to access and utilise powerful forms of AI to develop their own models. AutoML thereby enables developers to focus on curating high quality data rather than optimising models manually, facilitating a transition from model-driven to data-driven workflows (6). AutoML has been posited as a means of improving the reproducibility of ML research, and even generating superior model performance relative to conventional ML techniques (7).

AutoML technologies aim to automate some or all of the ML engineering process which otherwise requires advanced data or computer science skills. The first stage is data preparation, involving data integration, transformation, and cleaning. Next is feature selection, where aspects of the data to be utilised in designing the ML model are decided; this may involve data imputation, categorical encoding, and feature splitting (8). Model selection, training, and optimisation are then executed, with model performance evaluation being critical for identification of an optimal solution. AutoML systems use various methods and optimisation techniques to achieve state-of-the-art performance in some or all of the engineering process, such as Bayesian optimisation, random search, grid search, evolutionary based neural architecture selection, and meta-learning (7,9). The optimised model may then be outputted for further work, such as clinical deployment, explainability analysis, or external validation.

AutoML exhibits four major strengths which may support its application in clinical practice and research. Firstly, individual studies have reported comparable performance of autoML to conventionally developed models (10). This raises the possibility of clinical deployment of autoML models and use in pilot studies preceding further model development. Secondly, autoML may improve the reproducibility of ML research by reducing the influence of human technicians who currently engage with an idiosyncratic process of tuning until a satisfactory result is achieved: supporting a transition toward more reproducible data-centric development (6). Thirdly, the reduction in computational experience and hardware conferred by autoML adoption should act as a major democratising force, providing a much larger number of clinicians with access to AI technology (9). Lastly, the time spent on developing models is significantly reduced with autoML, as manual tuning is abolished—this improves efficiency and facilitates an acceleration of exploratory research to establish potential applications of AI (9).

With the myriad of available autoML tools, democratisation of AI beyond those with clinical and computational expertise is feasible, and potential applications are diverse (9,10). However, rigorous validation is necessary to justify deployment. Here, a systematic review was conducted to examine the performance of autoML in clinical settings. We aimed to evaluate the quality of result reporting; describe the specialties and clinical tasks in which autoML has been applied; and compare the performance of autoML platforms with conventionally developed models, as well as each other.

Methods

The reporting of this study adheres to PRISMA guidance, and the systematic review protocol was prospectively registered on PROSPERO (identifier CRD42022344427) (11,12). The protocol was amended to use a second quality assessment tool (PROBAST) in addition to DECIDE-AI, as described below.

Data sources and searches

The Cochrane Library, Embase (via OVID), MEDLINE (via OVID), and Scopus were searched from inception up to 11 July 2022, with no initial restrictions on publication status or type. Our search strategy isolated autoML in clinical contexts with the use of Boolean operators, as detailed in Supplementary Material 1. Before screening, duplicates were removed using Zotero version 6.0.14 (Corporation for Digital Scholarship, Vienna, Virginia, US); and Rayyan (11,13).

Study selection

Abstract screening was conducted in Rayyan by two independent researchers, with a separate third arbitrator with autoML expertise resolving cases of disagreement (13). Full-text screening was similarly conducted by two researchers with a separate arbitrator, in Microsoft Excel for Mac version 16.57 (Microsoft Corporation, Redmond, Washington, US). The explicit, hierarchical criteria for inclusion during abstract and full-text screening are listed below in descending order, with full details provided in Supplementary Material 2:

1. Is written in the English language.
2. Is a peer-reviewed primary research article.
3. Is not a retracted article.
4. Utilises automated machine learning.

5. AutoML is applied in a clinical context.

Data extraction and quality assessment

For articles satisfying the inclusion criteria, data extraction was conducted by two researchers; with a first clinical researcher's work verified by a second computational researcher. Quality assessment was conducted by a single researcher, using implicit criteria based on the Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI) framework (14). Risk of bias and concerns regarding applicability were assessed similarly by two researchers using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) framework and guidance questions (15).

Other data collected included citation details; autoML platform; processing location (cloud or local); code intensity of the autoML platform; technical features of the autoML platform; clinical task autoML applied towards; medical or surgical specialty defined anatomically where possible; sources of data used to train and test models; training and validation dataset size; dataset format (*i.e.* structured or unstructured); evaluation metrics used to gauge performance; and benchmark figures if presented such as with comparisons to expert clinician or conventional ML performance. Specifically, figures for area under the receiver operator characteristic curve (AUCROC), F1-score, and area under the precision-recall curve (AUCPR) were gathered. If F1-score was not provided but precision (positive predictive value) and recall (sensitivity) were, F1-score was calculated as the harmonic mean of the two metrics. If metrics were not stated in text form but were clearly plotted in graphical form, figures were manually interpolated using WebPlotDigitizer v4.6.0 (Ankit

Rohatgi, Pacifica, California, USA). Metrics were excluded if the source or modality of the tested model was unclear (16). Where two researchers disagreed, resolution was achieved through discussion or as-required arbitration by a third researcher.

Data synthesis and analysis

A narrative synthesis was conducted because meta-analysis was precluded by heterogeneity of datasets, platforms, and use-cases. Quantitative comparisons of autoML models was based on performance metrics (F1-score, AUCROC, AUCPR) to judge the clinical utility of applied autoML (17). AutoML platforms were compared on the same basis where platforms were applied to an identical task with the same data. A statistically significant difference in metrics was defined as featuring non-overlapping 95% confidence intervals. To establish the congruence between studies' conclusions and their presented data, the discussion and conclusion sections of each study were appraised by a single researcher to identify if autoML was compared to conventional techniques, and if so whether the comparison favoured autoML, conventional techniques, or neither. AutoML platforms were compared in terms of their requirements and capabilities, with researchers contacted to clarify any questions regarding code intensity, processing location, or data structure. Figures were produced with R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria) (18–20), and Affinity Designer version 1.10.4 (Pantone LLC, Carlstadt, New Jersey, USA).

Results

Record inclusion

Of 2417 records initially identified, 82 were included in the final analysis (Figure 1) (10,16,21–100). In rare cases, research reports referred to autoML or similar terms in the broader context of 'ML that automates', despite not utilising autoML technology: these articles were excluded under criterion 4 (101,102). Other borderline cases considered to be outside the scope of this review based on criterion 5 involved uses of autoML in clinical contexts, but without contributing to patient diagnosis, management, or prognosis. These included a surgical video identification and prediction of biological sex from medical images (103,104).

Characteristics of included studies

The characteristics of the 82 included studies are summarised in Figure 2 and Supplementary Material 3. AutoML first entered the medical literature in 2018 and has been growing in impact ever since: 1 paper in 2018; 7 in 2019; 21 in 2020; 35 in 2021; 18 by July 11th 2022. Use-cases are diverse, but diagnostic tasks (53 studies) were more common than management (four studies) or prognostic (25 studies) tasks. The most common specialties in which autoML was used were pulmonology and neurology. Structured (*e.g.* tabulated) and unstructured (*e.g.* imaging) data were used similarly commonly. Dataset size varied widely, between 31 to 2,185,920 for training; 8 to 2,185,920 for internal validation; and 27 to 34,128 for external validation.

Quality of reporting is summarised in Figure 2F, with individual scores reported in Supplementary Material 4. The median number of fulfilled DECIDE-AI criteria was 14 out of 27, with the highest score being 19 out of 27. Nine criteria were fulfilled by over 90% of

included studies. Thirteen criteria were not fulfilled in over half of the included studies: (III) Research governance, (3) Participants, (5) Implementation; (6) Safety and errors in the results; (7) Human factors; (8) Ethics; (VI) Patient involvement; (9) Participants; (10) Implementation; (11) Modifications; (13) Safety and errors in results; (14) Human factors; and (16) Safety and errors in the discussion. Of these, three criteria were not fulfilled by any of the 82 included studies: (8) Ethics; (VI) Patient involvement; and (13) Safety and errors in the results.

Risk of bias and concerns regarding applicability are summarised in Figure 2G. The most common sources of bias were retrospective study design often using publically available datasets, rather than testing autoML models in prospective trials to validate clinical performance and establish generalisability; and failure to provide an appropriate bespoke computational or clinical benchmark to demonstrate the performance of autoML—conferring unclear or high risk of bias in PROBAST appraisal (Supplementary Material 5). In many cases, this was because autoML was used as a tool, rather than the study being a trial of autoML technology, but a statement was made in the discussion or conclusion regarding the effectiveness of autoML in 27 of 47 studies (57%) judged to have a high or unclear risk of bias in the analysis.

AutoML performance relative to other modalities

The reporting of performance metrics varied widely between papers, likely representing the inherent limitations of applied autoML platforms. 79 studies (96%) provided AUCROC (Figure 3), F1-score (Figure 4), or AUCPR (Supplementary Material 6) as a measure of performance. Of these, 35 studies (44%) exhibited a computational or clinical benchmark to

compare autoML performance against, and 21 studies (27%) provided 95% confidence intervals for estimates of performance metrics. Of twelve studies (15%) with benchmark comparisons and confidence intervals, autoML exhibited statistically significantly superior AUCROC in six of 17 trials (35%); significantly superior F1-score in zero of one trial (0%); and significantly superior AUCPR in zero of two trials (0%). In studies with benchmark comparisons and confidence intervals, autoML did not exhibit the lowest AUCROC, F1-score, or AUCPR in any trial. In all studies comparing modalities, autoML exhibited the highest AUCROC in 28 of 37 trials (76%); the highest F1-score in eleven of 26 trials (42%); and the highest AUCPRC in ten of 12 trials (83%). AutoML exhibited the lowest AUCROC in five of 37 trials (14%); the lowest F1-score in six of 26 trials (23%); and the lowest AUCPR in two of twelve trials (17%). For autoML models, AUCROC ranged from 0.346-1.000 (scores of 0.5 are equivalent to chance; maximum score = 1); F1-score ranged from 0.128-0.992 (maximum score = 1); and AUCPR ranged from 0.280-1.000 (maximum score = 1).

57 studies (70%) compared autoML to other conventional modelling methods in the prose of their discussion or conclusion. Of these, 28 suggested that autoML was superior to conventional methods; 29 suggested that autoML was comparable to conventional methods; and none suggested that autoML was inferior to conventional methods. Only 35 studies provided a quantitative comparison in their results, as described above (Figure 3, Figure 4, Supplementary Material 6). Conclusions of comparable effectiveness were justified by congruence with reported performance metrics in 16 of 29 studies (55%); conclusions of superior effectiveness of autoML were justified in eleven of 28 studies (39%).

Comparative performance of AutoML platforms

A comparative summary of the autoML platforms validated in the literature is presented in Table 1. Platforms vary greatly in their accessibility, technical features, and portability. While performance in different tasks cannot be compared, five studies directly compared distinct autoML platforms in the same task. Of these, one study (20%) provided AUCROC metrics, which favoured AutoPrognosis over TPOT to prognosticate mortality in cystic fibrosis.(23) Four studies (80%) provided F1-score metrics for a total of nine trials (Figure 5): prognosticating mortality in cystic fibrosis; predicting invasion depth of gastric neoplasms from endoscopic photography; diagnosing referrable diabetic retinopathy from fundus photography; diagnosing age-related macular degeneration, central serous retinopathy, macular hole, and diabetic retinopathy from optical coherence tomography (OCT); diagnosing choroidal neovascularisation, diabetic macular oedema, and drusen from OCT; and classifying spine implants from lumbar spine radiographs (23,29,50,97).

AutoPrognosis (structured data) and Rekognition (unstructured data) exhibited the strongest performance as they were superior to every platform they were compared with, although this was only to TPOT for AutoPrognosis, and Rekognition was compared with fewer platforms than Cloud AutoML. Two studies (40%) reporting five trials provided AUCPR metrics for prognosticating mortality in cystic fibrosis and classifying electrocardiogram traces (23,32). Here, performance favoured AutoPrognosis over TPOT; and AutoDAL-SOAR over USDM, AER, Auto-Weka, Auto-Sklearn, and ASSL+US. While not all platforms can be compared against one another due to incompatibility with data structure, many possible combinations were not trialled and the number of comparative trials was small, making it difficult to establish comparative performance.

Confidence in conclusions

Confidence in conclusions is tempered by high risk of bias, particularly in retrospective study design and limited metrics facilitating statistical comparisons. However, as autoML did not exhibit statistically significantly worse performance than conventional techniques in any trial and exhibited lower performance metrics than conventional trials in a minority of studies, there is high confidence in the conclusion that autoML technology facilitates production of models with comparable performance to conventional techniques such as bespoke computational approaches. Given the low number of studies providing confidence intervals to enable statistical comparison of models' performance within trials, conclusions regarding the superiority of autoML relative to conventional techniques have low confidence. In addition, conclusions cannot be assumed to generalise to all use cases and datasets: performance is highly context-specific, as demonstrated by the large variability observed in AUCROC (Figure 3), F1-score (Figure 4), and AUCPR (Supplementary Material 6). Confidence in the superior performance of AutoPrognosis with structured data is very low, as there were very few comparative trials; and low for the superior performance of Rekognition with unstructured data, as the number of comparative trials was low—though not as low as for structured data—and as there were no data for many possible platform comparisons.

Discussion

This study shows that autoML has been trialled in a wide variety of diagnostic, patient management, and prognostic tasks. AutoML has been used in many clinical specialties, most commonly in brain and lung imaging. Performance of autoML models generally

compares well to bespoke computational and clinical benchmarks, often exhibiting superior performance. However, available studies and appraised risk of bias preclude conclusion of autoML providing universally superior performance to conventional modelling; relative and absolute performance vary widely with the applied platform, use case, and data source. The strength of the evidence base supporting use of different autoML platforms is highly heterogenous, with some platforms exhibiting results more supportive of equivalence or superiority to conventional techniques than others. Few studies compared different autoML platforms to determine which provide optimal performance for a given task. Despite these knowledge gaps, a high number of non-comparative studies suggests that autoML is already being applied as a statistical tool, comparable to bespoke machine learning coding packages or statistical software.

There are five main deficiencies in the quality of the autoML evidence base. First, inconsistency in performance metrics may be a consequence of restrictions imposed by autoML platforms but observed variation between studies using similar platforms also suggests that selective reporting is common. Reporting comprehensive metrics is essential, particularly in the context of diagnostic algorithms, as some metrics are a function of prevalence or model threshold (17). Second, explainability analysis is challenging for similar reasons regarding portability, but is possible with emerging technological solutions (22). In addition, some platforms incorporate inbuilt explainability, such as by providing salience maps for DL models. Issues regarding 'black box' algorithms are accentuated in autoML research, leading to a third limitation: a lack of ethical consideration—such as regarding algorithmic fairness—by all the included studies.

Fourth, inconsistent use of benchmarking represents a form of publication bias leading to erroneous conclusions of equivalent or superior autoML performance relative to conventional bespoke computational methods or clinicians. Many studies relied on historical controls or provided no benchmark at all. To confidently conclude that autoML performance compares well to bespoke models—and particularly to 'state-of-the-art techniques'—a researcher with computational aptitude should have an opportunity to maximise performance. Finally, models should be deployed on separate datasets which were not used in testing or training, for external validation; this demonstrates generalisability, a critical component of clinical potential. Without external validation, overfitting to the datasets provided may lead to inflated estimates of performance (105). External validation is limited on many autoML platforms by a lack of ability to batch test on new data, or to export models for analysis and deployment.

Limitations

This systematic review was limited by three issues: **(1)** PROBAST had to be adapted to apply it in non-diagnostic applications of autoML—we employed DECIDE-AI as a domain-specific quality indicator to mitigate this limitation, and utilised PROBAST in the context of trialling autoML technology rather than in validating models for clinical application. Development of more domain-specific tools to optimise AI-related systematic reviews is underway, and will be a welcome development (106,107). **(2)** Confidence in conclusions was affected by high risk of bias, a common theme in AI research more broadly (108). We provide comprehensive indicators of quality, risk of bias, and concerns regarding applicability to facilitate contextualisation of performance metrics. **(3)** It is difficult to draw conclusions for autoML as a modality because platforms are variable in their features,

performances, and requirements—future reviews may focus on individual platforms, although the number of studies featuring most platforms is very small.

Implications

Researchers applying a platform without providing benchmark comparators for the purposes of primary research or clinical work should justify their decision with validation data demonstrating that their approach is acceptable. Evidence should be contextually relevant, preferably pertaining to the same clinical task. While it is apparent that autoML has already begun to be applied in clinical research as a statistical tool, it is important that these tools are demonstrated to produce accurate, reliable, and fair models. Studies purported as evidence of validation of autoML are often limited by retrospective design, high risk of bias, and unfulfillment of conventional reporting standards—comparable to research regarding other AI technologies (109). Future comparative studies should address the limitations discussed above to convince researchers, clinicians, and policy makers that autoML platforms may be applied in lieu of bespoke modelling.

When reporting AI algorithms tasked with a certain clinical job, it would be helpful to avoid ambiguity in terminology. We would suggest a complete restriction of the terms 'automated machine learning', or 'autoML' for those algorithms built with technology that automates some or all parts of the process of the engineering process—all conventional ML models process data without human guidance, so description of these technologies as automated is redundant. Similar terms such as 'automated artificial intelligence', 'automated machine learning', and 'automated deep learning' are also redundant in the context of bespoke computational models. A simple alternative term for conventional ML

projects would be 'automatic'—these systems may automate a particular task, but their development is not automated, the defining feature of autoML.

The reduced barrier to entry in terms of computational expertise and hardware requirements conferred by many autoML platforms makes them a powerful contributor to democratisation of AI technology: a far greater number of clinicians and scientists are capable of ML development through use of these platforms. AutoML could be an invaluable resource for teaching, as individuals can more rapidly develop hands-on experience, learn by trial-and-error, and thereby develop intuitive understanding of the capabilities and limitations of ML. AutoML could also be applied in pilot studies, enabling clinicians with domain-specific expertise to explore possibilities for ML research—facilitating prioritisation of allocation of scarce resources such as GPU access and expert computer scientists. Validated platforms may be applied more broadly, including in patient care. Moreover, autoML is well placed to respond to calls to inculcate data-centric AI as opposed to model-centric development; focusing effort on curating high quality data, which limits development more often than code or model infrastructure (6). Acceleration in this process may be facilitated by large language models as their emerging capability to leverage plugins will allow autoML to facilitate AI building itself to fulfil user-defined aims (110).

Further work is indicated to improve validation of autoML platforms, either by allowing models to be exported, or by providing more comprehensive internal metrics. Other work should focus on improving the functionality of autoML, specifically on reducing the trade-offs currently implicit in selecting a platform with a given code intensity and computing locus. Using automation to reduce human error to optimise engineering and improve

performance is one ideal: this has been demonstrated with structured data by AutoPrognosis. Increased functionality of code-free platforms while retaining the customisability of code-intense solutions is another ideal: H2O.ai Driverless AI offers the same functionality as the H2O.ai R and Python packages, but with a code-free graphical user interface. Alternatively, maximising accessibility by automating the whole engineering process may be desirable: Dedicaid is a platform requiring just data, with no customisable parameters, but has an 'ethical compass' which flags inappropriate datasets.

Conclusion

AutoML performance is often comparable to bespoke ML and human performance. Many autoML platforms have been developed in academia and industry, with variable strengths and limitations. AutoML may prove especially useful in pilot studies and education, but potential use-cases include primary research and clinical deployment if platforms are rigorously validated. Future autoML research must be more transparently reported, adhere to reporting guidelines, and provide appropriate benchmarks for performance comparisons. Further autoML development should seek to minimise the 'trade-offs' currently inherent in selecting any given platform.

Data availability statement

The raw data from this review may be provided upon request.

Competing interests

All authors declare no competing interests.

Financial support

AJT is supported by The Royal College of Surgeons in Edinburgh (RCSED Bursary 2022), 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.

Author contributions

AJT and DSWT conceived and coordinated the study. AJT, KE, LG, RH, YL, TFT, HC, and GL contributed to data collection. AJT and KE produced visualisations. AJT conducted data analysis. AJT and ZLT drafted the manuscript. KE, LG, RH, YL, TFT, GL, and DSWT edited the manuscript. All authors approved the final draft before submission.

Legends

Table 1—Technological comparison of autoML platforms applied in the studies included in this review. AER = approximated error reduction; ML = machine learning; ASSL = automated semi-supervised learning; WEKA = Waikato Environment for Knowledge Analysis; AutoDAL = automated distributed active learning; AutoDC = automated data-

centric processing; JADBio = Just-Add-Data Bio; KNIME = Konstanz Information Miner; MILO = Machine Intelligence Learning Optimizer; TPOT = Tree-based Pipeline Optimization Tool; USDM = uncertainty sampling with diversity maximization.

Figure 1—PRISMA flow chart depicting the search, screening, and inclusion process of this review.

Figure 2—Collectively summarised characteristics of included studies: (A) Date of publication bar chart; (B) Continent of corresponding author bar chart; (C) Country of corresponding author bar chart; (D) Funding source bar chart; (E) Clinical specialty bar chart; (F) DECIDE-AI score histogram; (G) PROBAST evaluation bar chart; (H) Dataset size histogram with logarithmic X-axis; (I) Data nature bar chart; (J) Data source bar chart. RoB = risk of bias; CrA = concerns regarding applicability.

Figure 3—Forest plot depicting reported AUCROC metrics. SVM = support vector machine; FEV1 = forced expiratory volume in 1 second; PH = proportional hazards; GBM = gradient boosting machine; DRF = distributed random forest; XRT = extremely randomised tree; BMI = body-mass index; ccf-DNA = circulating cell-free DNA; CT = computerised tomography; ARSS = Aneurysm Recanalization Stratification Scale; MCTSI = Modified Computed Tomography Severity Index.

Figure 4—Forest plot depicting reported F1-score metrics. SVM = support vector machine; FEV1 = forced expiratory volume in 1 second; GBM = gradient boosting machine; ARSS = Aneurysm Recanalization Stratification Scale.

Figure 5—Heat map depicting the comparative performance of autoML platforms as applied to the same clinical tasks in terms of F1-score. Shading and numbers correspond to the number of superior performances exhibited by the index platform with respect to the reference platform.

Supplementary Material 1—Systematic review search strategy.

Supplementary Material 2—Inclusion and exclusion criteria, as provided to researchers conducting abstract and full-text screening.

Supplementary Material 3—Tabulated study characteristics.

Supplementary Material 4—Study-level data exhibiting fulfilment of DECIDE-AI reporting standards.

Supplementary Material 5—Study-level data exhibiting appraisal of risk of bias (RoB) and concerns regarding applicability (CrA) using PROBAST.

Supplementary Material 6—Forest plot depicting reported AUCPR metrics. SVM = support vector machine; FEV1 = forced expiratory volume in 1 second; GBM = gradient boosting machine; DRF = distributed random forest; XRT = extremely randomised tree.

References

1. Pianykh OS, Guitron S, Parke D, Zhang C, Pandharipande P, Brink J, et al. Improving healthcare operations management with machine learning. Nat Mach Intell. 2020 May;2(5):266–73.

2. Park JY, Hsu TC, Hu JR, Chen CY, Hsu WT, Lee M, et al. Predicting Sepsis Mortality in a Population-Based National Database: Machine Learning Approach. J Med Internet Res. 2022 Apr 13;24(4):e29982.

3. Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. BMC Med. 2019 Dec;17(1):143, s12916-019-1382–x.

4. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data. 2019 Jun 19;6(1):54.

5. 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. The Lancet Global Health. 2023 Sep 1;11(9):e1432–43.

6. Chang EY. Knowledge-Guided Data-Centric AI in Healthcare: Progress, Shortcomings, and Future Directions [Internet]. arXiv; 2022 [cited 2023 Jan 18]. Available from: http://arxiv.org/abs/2212.13591

7. Hutter F, Kotthoff L, Vanschoren J, editors. Automated Machine Learning: Methods, Systems, Challenges [Internet]. Cham: Springer International Publishing; 2019 [cited 2022 Jul 12]. (The Springer Series on Challenges in Machine Learning). Available from: http://link.springer.com/10.1007/978-3-030-05318-5

8. Rawat T, Khemchandani V. Feature Engineering (FE) Tools and Techniques for Better Classification Performance. IJIET [Internet]. 2017 [cited 2022 Sep 7];8(2). Available from: http://ijiet.com/wp-content/uploads/2017/05/24.pdf

9. Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine. 2020 Apr 1;104:101822.

10. 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 Sep;1(5):e232–42.

11. Thirunavukarasu A, Gutierrez L, Elangovan K, Zheng F, Li S, Ting D. The applications of automated machine learning in clinical contexts [Internet]. PROSPERO; 2022. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022344427

12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;372:n71.

13. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews. 2016 Dec 5;5(1):210.

14. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022 May 18;377:e070904.

15. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019 Jan 1;170(1):51–8.

16. Shen H, Liu T, Cui J, Borole P, Benjamin A, Kording K, et al. A web-based automated machine learning platform to analyze liquid biopsy data. Lab chip. 2020;20(12):2166–74.

17. Erickson BJ, Kitamura F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiol Artif Intell. 2021 May 12;3(3):e200126.

18. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. Research Synthesis Methods [Internet]. 2020 Apr 26 [cited 2020 May 21];n/a(n/a). Available from: https://doi.org/10.1002/jrsm.1411

19. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019 Nov 21;4(43):1686.

20. Dayim A. forestploter [Internet]. 2023 [cited 2023 Jan 3]. Available from: https://github.com/adayim/forestploter

21. Hasimbegovic E, Papp L, Grahovac M, Krajnc D, Poschner T, Hasan W, et al. A Sneak-Peek into the Physician's Brain: A Retrospective Machine Learning-Driven Investigation of Decision-Making in TAVR versus SAVR for Young High-Risk Patients with Severe Symptomatic Aortic Stenosis. J Pers Med. 2021 Oct 22;11(11):1062.

22. 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 Feb 5;

23. Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Sci rep. 2018;8(1):11242.

24. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One. 2019;14(5):e0213653.

25. An JY, Seo H, Kim YG, Lee KE, Kim S, Kong HJ. Codeless Deep Learning of COVID-19 Chest X-Ray Image Dataset with KNIME Analytics Platform. Healthc inform res. 2021;27(1):82–91.

26. Antaki F, Coussa RG, Kahwati G, Hammamji K, Sebag M, Duval R. Accuracy of automated machine learning in classifying retinal pathologies from ultra-widefield pseudocolour fundus images. Br J Ophthalmol. 2021 Aug 3;bjophthalmol-2021-319030.

27. Antaki F, Kahwati G, Sebag J, Coussa RG, Fanous A, Duval R, et al. Predictive modeling of proliferative vitreoretinopathy using automated machine learning by ophthalmologists without coding experience. Sci Rep. 2020 Nov 11;10(1):19528.

28. Bai Y, Li Y, Shen Y, Yang M, Zhang W, Cui B. AutoDC: an Automatic Machine Learning Framework for Disease Classification. Bioinformatics. 2022;(cw9, 9808944).

29. Bang CS, Lim H, Jeong HM, Hwang SH. Use of Endoscopic Images in the Prediction of Submucosal Invasion of Gastric Neoplasms: Automated Deep Learning Model Development and Usability Study. J Med Internet Res. 2021;23(4):e25167.

30. Bhat GS, Shankar N, Panahi IMS. Automated machine learning based speech classification for hearing aid applications and its real-time implementation on smartphone. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual International Conference. United States; 2020. p. 956–9.

31. Borkowski AA, Viswanadhan NA, Thomas LB, Guzman RD, Deland LA, Mastorides SM. Using Artificial Intelligence for COVID-19 Chest X-ray Diagnosis. Fed Pract. 2020 Sep;37(9):398–404.

32. Chen X, Wujek B. A Unified Framework for Automatic Distributed Active Learning. IEEE trans pattern anal mach intell. 2021;PP(9885960).

33. Chou A, Torres-Espin A, Kyritsis N, Huie JR, Khatry S, Funk J, et al. Expert-augmented automated machine learning optimizes hemodynamic predictors of spinal cord injury outcome. PLoS ONE. 2022;17(4):e0265254.

34. Danilatou V, Nikolakakis S, Antonakaki D, Tzagkarakis C, Mavroidis D, Kostoulas T, et al. Outcome Prediction in Critically-Ill Patients with Venous Thromboembolism and/or Cancer Using Machine Learning Algorithms: External Validation and Comparison with Scoring Systems. IJMS. 2022 Jun 27;23(13):7132.

35. Feretzakis G, Sakagianni A, Loupelis E, Kalles D, Skarmoutsou N, Martsoukou M, et al. Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy. Healthc inform res. 2021;27(3):214–21.

36. Ghosh T., Tanwar S., Chumber S., Vani K. Classification of chest radiographs using general purpose cloud-based automated machine learning: pilot study. Egypt J Radiol Nucl Med. 2021;52(1):120.

37. Hu R, Li H, Horng H, Thomasian NM, Jiao Z, Zhu C, et al. Automated machine learning for differentiation of hepatocellular carcinoma from intrahepatic cholangiocarcinoma on multiphasic MRI. Sci rep. 2022;12(1):7924.

38. Ikemura K, Bellin E, Yagi Y, Billett H, Saada M, Simone K, et al. Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study. J Med Internet Res. 2021;23(2):e23458.

39. Ito H, Nakamura Y, Takanari K, Oishi M, Matsuo K, Kanbe M, et al. 'Development of a Novel Scar Screening System with Machine Learning'. Plast Reconstr Surg. 2022;(1306050).

40. Ito Y, Unagami M, Yamabe F, Mitsui Y, Nakajima K, Nagao K, et al. A method for utilizing automated machine learning for histopathological classification of testis based on Johnsen scores. Sci rep. 2021;11(1):9962.

41. Jen KY, Albahra S, Yen F, Sageshima J, Chen LX, Tran N, et al. Automated En Masse Machine Learning Model Generation Shows Comparable Performance as Classic Regression Models for Predicting Delayed Graft Function in Renal Allografts. Transplantation. 2021;105(12):2646–54.

42. Karaglani M, Gourlia K, Tsamardinos I, Chatzaki E. Accurate Blood-Based Diagnostic Biosignatures for Alzheimer's Disease via Automated Machine Learning. J Clin Med. 2020;9(9).

43. Karaglani M, Panagopoulou M, Cheimonidi C, Tsamardinos I, Maltezos E, Papanas N, et al. Liquid Biopsy in Type 2 Diabetes Mellitus Management: Building Specific Biosignatures via Machine Learning. J Clin Med. 2022;11(4).

44. Karhade DS, Roach J, Shrestha P, Simancas-Pallares MA, Ginnis J, Burk ZJS, et al. An Automated Machine Learning Classifier for Early Childhood Caries. Pediatr Dent. 2021;43(3):191–7.

45. Karstoft KI, Tsamardinos I, Eskelund K, Andersen SB, Nissen LR. Applicability of an Automated Model and Parameter Selection in the Prediction of Screening-Level PTSD in Danish Soldiers Following Deployment: Development Study of Transferable Predictive Models Using Automated Machine Learning. JMIR Med Inform. 2020;8(7):e17119.

46. Katsuki M, Kawamura S, Koh A. Easily Created Prediction Model Using Automated Artificial Intelligence Framework (Prediction One, Sony Network Communications Inc., Tokyo, Japan) for Subarachnoid Hemorrhage Outcomes Treated by Coiling and Delayed Cerebral Ischemia. Cureus. 2021 Jun;13(6):e15695.

47. Katsuki M, Matsuo M. Relationship Between Medical Questionnaire and Influenza Rapid Test Positivity: Subjective Pretest Probability, 'I Think I Have Influenza,' Contributes to the Positivity Rate. Cureus. 2021;13(7):e16679.

48. Kim IK, Lee K, Park JH, Baek J, Lee WK. Classification of pachychoroid disease on ultrawide-field indocyanine green angiography using auto-machine learning platform. Br J Ophthalmol. 2021 Jun;105(6):856–61.

49. Koga S, Ghayal NB, Dickson DW. Deep Learning-Based Image Classification in Differentiating Tufted Astrocytes, Astrocytic Plaques, and Neuritic Plaques. J Neuropathol Exp Neurol. 2021;80(4):306–12.

50. Korot E, Guan Z, Ferraz D, Wagner SK, Zhang G, Liu X, et al. Code-free deep learning for multi-modality medical image classification. Nature Machine Intelligence. 2021;3(4):288–98.

51. Kumar M, Ang LT, Png H, Ng M, Tan K, Loy SL, et al. Automated Machine Learning (AutoML)-Derived Preconception Predictive Risk Model to Guide Early Intervention for Gestational Diabetes Mellitus. Int J Environ Res Public Health. 2022;19(11).

52. Lee JH, Kim YT, Lee JB, Jeong SN. A Performance Comparison between Automated Deep Learning and Dental Professionals in Classification of Dental Implant Systems from Dental Imaging: A Multi-Center Study. Diagnostics (Basel). 2020;10(11).

53. Liu Y, Li T, Fan Z, Li Y, Sun Z, Li S, et al. Image-Based Differentiation of Intracranial Metastasis From Glioblastoma Using Automated Machine Learning. Front neurosci. 2022;16(101478481):855990.

54. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: An automated machine learning approach. Laryngoscope. 2020;130(6):1408–13.

55. Luna A, Bernanke J, Kim K, Aw N, Dworkin JD, Cha J, et al. Maturity of gray matter structures and white matter connectomes, and their relationship with psychiatric symptoms in youth. Hum Brain Mapp. 2021;42(14):4568–79.

56. Mazaki J, Katsumata K, Ohno Y, Udo R, Tago T, Kasahara K, et al. A Novel Predictive Model for Anastomotic Leakage in Colorectal Cancer Using Auto-artificial Intelligence. Anticancer Res. 2021 Nov;41(11):5821–5.

57. Mohsen F, Biswas MR, Ali H, Alam T, Househ M, Shah Z. Customized and Automated Machine Learning-Based Models for Diabetes Type 2 Classification. Stud Health Technol Inform. 2022;295(ck1, 9214582):517–20.

58. Nagy A, Ligeti B, Szebeni J, Pongor S, Gyrffy B. COVIDOUTCOME-estimating COVID severity based on mutation signatures in the SARS-CoV-2 genome. Database (Oxford). 2021;2021(101517697).

59. Narkhede SM, Luther L, Raugh IM, Knippenberg AR, Esfahlani FZ, Sayama H, et al. Machine Learning Identifies Digital Phenotyping Measures Most Relevant to Negative Symptoms in Psychotic Disorders: Implications for Clinical Trials. Schizophr Bull. 2022;48(2):425–36.

60. Nero C, Ciccarone F, Boldrini L, Lenkowicz J, Paris I, Capoluongo ED, et al. Germline BRCA 1-2 status prediction through ovarian ultrasound images radiogenomics: a hypothesis generating study (PROBE study). Sci rep. 2020;10(1):16511.

61. Orlenko A, Kofink D, Lyytikainen LP, Nikus K, Mishra P, Kuukasjarvi P, et al. Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning. Bioinformatics. 2020;36(6):1772–8.

62. Ou C, Liu J, Qian Y, Chong W, Liu D, He X, et al. Automated Machine Learning Model Development for Intracranial Aneurysm Treatment Outcome Prediction: A Feasibility Study. Front Neurol. 2021;12(101546899):735142.

63. Padmanabhan M, Yuan P, Chada G, Nguyen HV. Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction. J Clin Med. 2019;8(7).

64. Panagopoulou M, Cheretaki A, Karaglani M, Balgkouranidou I, Biziota E, Amarantidis K, et al. Methylation Status of Corticotropin-Releasing Factor (CRF) Receptor Genes in Colorectal Cancer. J Clin Med. 2021;10(12).

65. Panagopoulou M, Karaglani M, Manolopoulos VG, Iliopoulos I, Tsamardinos I, Chatzaki E. Deciphering the Methylation Landscape in Breast Cancer: Diagnostic and Prognostic Biosignatures through Automated Machine Learning. Cancers (Basel). 2021;13(7).

66. Papoutsoglou G, Karaglani M, Lagani V, Thomson N, Roe OD, Tsamardinos I, et al. Automated machine learning optimizes and accelerates predictive modeling from COVID-19 high throughput datasets. Sci rep. 2021;11(1):15107.

67. Peng WL, Zhang TJ, Shi K, Li HX, Li Y, He S, et al. Automatic machine learning based on native T1 mapping can identify myocardial fibrosis in patients with hypertrophic cardiomyopathy. Eur Radiol. 2022;32(2):1044–53.

68. Purkayastha S, Zhao Y, Wu J, Hu R, McGirr A, Singh S, et al. Differentiation of low and high grade renal cell carcinoma on routine MRI with an externally validated automatic machine learning algorithm. Sci rep. 2020;10(1):19503.

69. Radzi SFM, Karim MKA, Saripan MI, Rahman MAA, Isa INC, Ibahim MJ. Hyperparameter Tuning and Pipeline Optimization via Grid Search Method and Tree-Based AutoML in Breast Cancer Prediction. J pers med. 2021;11(10).

70. Rallabandi V.P.S., Tulpule K., Gattu M. Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer's disease using structural MRI analysis. Inform Med Unlocked. 2020;18((Rallabandi, Tulpule, Gattu) Excelra Knowledge Solutions Pvt. Ltd., Hyderabad, Telangana, India):100305.

71. Rashidi HH, Dang LT, Albahra S, Ravindran R, Khan IH. Automated machine learning for endemic active tuberculosis prediction from multiplex serological data. Sci rep. 2021;11(1):17900.

72. Rashidi HH, Khan IH, Dang LT, Albahra S, Ratan U, Chadderwala N, et al. Prediction of Tuberculosis Using an Automated Machine Learning Platform for Models Trained on Synthetic Data. J Pathol Inform. 2022;13(101528849):10.

73. Rashidi HH, Makley A, Palmieri TL, Albahra S, Loegering J, Fang L, et al. Enhancing Military Burn- and Trauma-Related Acute Kidney Injury Prediction Through an Automated Machine Learning Platform and Point-of-Care Testing. Arch Pathol Lab Med. 2021;145(3):320–6.

74. Real AD, Real OD, Sardina S, Oyonarte R. Use of automated artificial intelligence to predict the need for orthodontic extractions. Korean j orthod. 2022;52(2):102–11.

75. Ritter Z, Papp L, Zambo K, Toth Z, Dezso D, Veres DS, et al. Two-Year Event-Free Survival Prediction in DLBCL Patients Based on In Vivo Radiomics and Clinical Parameters. Front oncol. 2022;12(101568867):820136.

76. Sakagianni A, Feretzakis G, Kalles D, Koufopoulou C, Kaldis V. Setting up an Easy-to-Use Machine Learning Pipeline for Medical Decision Support: A Case Study for COVID-19 Diagnosis Based on Deep Learning with CT Scans. Stud Health Technol Inform. 2020;272(ck1, 9214582):13–6.

77. Salmanpour MR, Shamsaei M, Saberi A, Setayeshi S, Klyuzhin IS, Sossi V, et al. Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease. Comput Biol Med. 2019;111(doc, 1250250):103347.

78. Salmanpour MR, Shamsaei M, Saberi A, Klyuzhin IS, Tang J, Sossi V, et al. Machine learning methods for optimal prediction of motor outcome in Parkinson's disease. Phys Med. 2020;69(9302888):233–40.

79. Sills MR, Ozkaynak M, Jang H. Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning. Int J Med Inf. 2021;151(ct4, 9711057):104468.

80. Stojadinovic M, Milicevic B, Jankovic S. Improved predictive performance of prostate biopsy collaborative group risk calculator when based on automated machine learning. Comput Biol Med. 2021;138(doc, 1250250):104903.

81. Su X, Chen N, Sun H, Liu Y, Yang X, Wang W, et al. Automated machine learning based on radiomics features predicts H3 K27M mutation in midline gliomas of the brain. Neuro-oncol. 2020;22(3):393–401.

82. Sun H, Qu H, Chen L, Wang W, Liao Y, Zou L, et al. Identification of suspicious invasive placentation based on clinical MRI data using textural features and automated machine learning. Eur Radiol. 2019;29(11):6152–62.

83. Tahmasebi A, Qu E, Sevrukov A, Liu JB, Wang S, Lyshchik A, et al. Assessment of Axillary Lymph Nodes for Metastasis on Ultrasound Using Artificial Intelligence. Ultrason Imaging. 2021;43(6):329–36.

84. Tan HB, Xiong F, Jiang YL, Huang WC, Wang Y, Li HH, et al. The study of automatic machine learning base on radiomics of non-focus area in the first chest CT of different clinical types of COVID-19 pneumonia. Sci rep. 2020;10(1):18926.

85. Tomic A, Tomic I, Rosenberg-Hasson Y, Dekker CL, Maecker HT, Davis MM. SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses. J Immunol. 2019;203(3):749–59.

86. Tran NK, Albahra S, Pham TN, Holmes JH 4th, Greenhalgh D, Palmieri TL, et al. Novel application of an automated-machine learning development tool for predicting burn sepsis: proof of concept. Sci rep. 2020;10(1):12354.

87. Tran NK, Howard T, Walsh R, Pepper J, Loegering J, Phinney B, et al. Novel application of automated machine learning with MALDI-TOF-MS for rapid high-throughput screening of COVID-19: a proof of concept. Sci rep. 2021;11(1):8219.

88. Vagliano I, Brinkman S, Abu-Hanna A, Arbous MS, Dongelmans DA, Elbers PWG, et al. Can we reliably automate clinical prognostic modelling? A retrospective cohort study for ICU triage prediction of in-hospital mortality of COVID-19 patients in the Netherlands. Int J Med Inf. 2022;160(ct4, 9711057):104688.

89. van Eeden WA, Luo C, van Hemert AM, Carlier IVE, Penninx BW, Wardenaar KJ, et al. Predicting the 9-year course of mood and anxiety disorders with automated machine learning: A comparison between auto-sklearn, naive Bayes classifier, and traditional logistic regression. Psychiatry Res. 2021;299(qc4, 7911385):113823.

90. Wan KW, Wong CH, Ip HF, Fan D, Yuen PL, Fong HY, et al. Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study. Quant imaging med surg. 2021;11(4):1381–93.

91. Wang G, Sun Y, Chen Y, Gao Q, Peng D, Lin H, et al. Rapid identification of human ovarian cancer in second harmonic generation images using radiomics feature analyses and tree-based pipeline optimization tool. J biophotonics. 2020;13(9):e202000050.

92. Wang G, Sun Y, Jiang S, Wu G, Liao W, Chen Y, et al. Machine learning-based rapid diagnosis of human borderline ovarian cancer on second-harmonic generation images. Biomed Opt Express. 2021;12(9):5658–69.

93. Wang HL, Hsu WY, Lee MH, Weng HH, Chang SW, Yang JT, et al. Automatic Machine-Learning-Based Outcome Prediction in Patients With Primary Intracerebral Hemorrhage. Front Neurol. 2019;10(101546899):910.

94. Wang S., Niu S., Qu E., Forsberg F., Wilkes A., Sevrukov A., et al. Characterization of indeterminate breast lesions on B-mode ultrasound using automated machine learning models. J Med Imaging. 2020;7(5):057002–1.

95. Wang S, Xu J, Tahmasebi A, Daniels K, Liu JB, Curry J, et al. Incorporation of a Machine Learning Algorithm With Object Detection Within the Thyroid Imaging Reporting and Data System Improves the Diagnosis of Genetic Risk. Front oncol. 2020;10(101568867):591846.

96. Xavier BA, Chen PH. Natural Language Processing for Imaging Protocol Assignment: Machine Learning for Multiclass Classification of Abdominal CT Protocols Using Indication Text Data. J Digit Imaging. 2022;(a19, 9100529).

97. Yang HS, Kim KR, Kim S, Park JY. Deep Learning Application in Spinal Implant Identification. Spine. 2021;46(5):E318–24.

98. Yin M, Zhang R, Zhou Z, Liu L, Gao J, Xu W, et al. Automated Machine Learning for the Early Prediction of the Severity of Acute Pancreatitis in Hospitals. Front cell infect microbiol. 2022;12(101585359):886935.

99. Zeng Y, Zhang J. A machine learning model for detecting invasive ductal carcinoma with Google Cloud AutoML Vision. Comput Biol Med. 2020;122(doc, 1250250):103861.

100. Zhang S, Sun H, Su X, Yang X, Wang W, Wan X, et al. Automated machine learning to predict the co-occurrence of isocitrate dehydrogenase mutations and O6 - methylguanine-DNA methyltransferase promoter methylation in patients with gliomas. J Magn Reson Imaging. 2021;54(1):197–205.

101. Cho BH, Kaji D, Cheung ZB, Ye IB, Tang R, Ahn A, et al. Automated Measurement of Lumbar Lordosis on Radiographs Using Machine Learning and Computer Vision. Global spine j. 2020;10(5):611–8.

102. Adaszewski S, Dukart J, Kherif F, Frackowiak R, Draganski B, Alzheimer's Disease Neuroimaging Initiative. How early can we predict Alzheimer's disease using computational anatomy?. Neurobiol Aging. 2013;34(12):2815–26.

103. Smith R, Julian D, Dubin A. Deep neural networks are effective tools for assessing performance during surgical training. J robot surg. 2022;16(3):559–62.

104. Korot E, Pontikos N, Liu X, Wagner SK, Faes L, Huemer J, et al. Predicting sex from retinal fundus photographs using automated deep learning. Sci Rep. 2021 May 13;11(1):10286.

105. Ying X. An Overview of Overfitting and its Solutions. J Phys: Conf Ser. 2019 Feb;1168:022022.

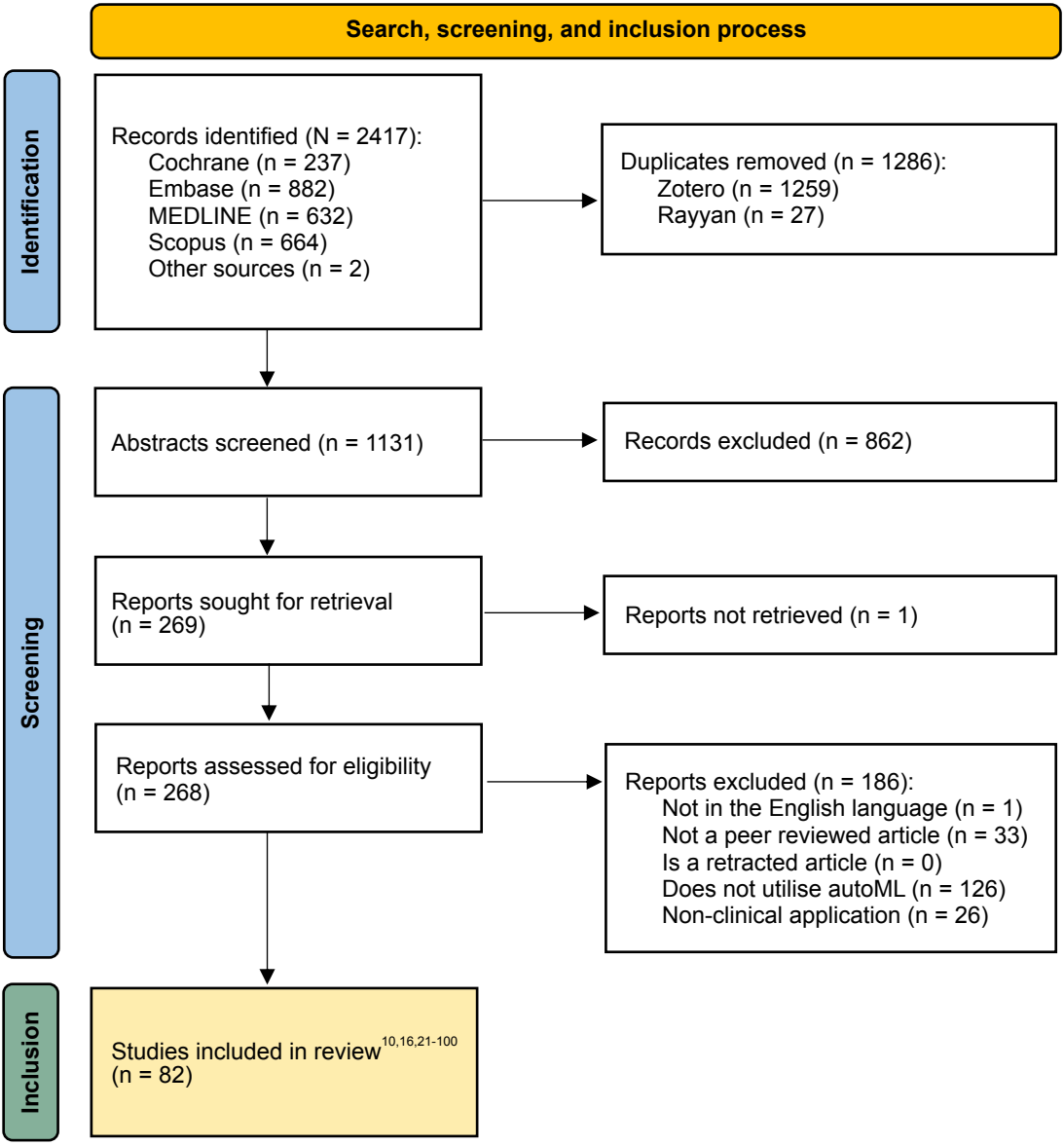
106. Cacciamani GE, Chu TN, Sanford DI, Abreu A, Duddalwar V, Oberai A, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med. 2023 Jan;29(1):14–5.

107. Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021 Jul 1;11(7):e048008.

108. Navarro CLA, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. 2021 Oct 20;375:n2281.

109. 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 Mar 25;368:m689.

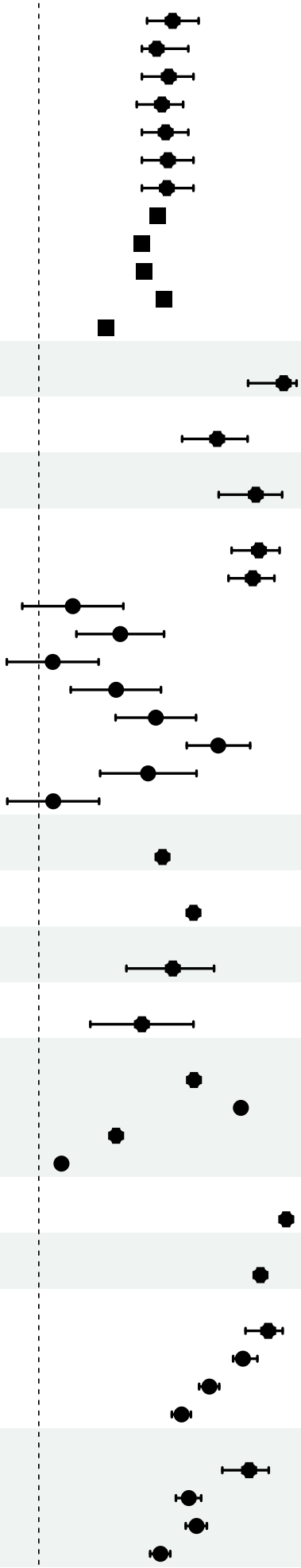
110. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Jul 17;29:1930–40.

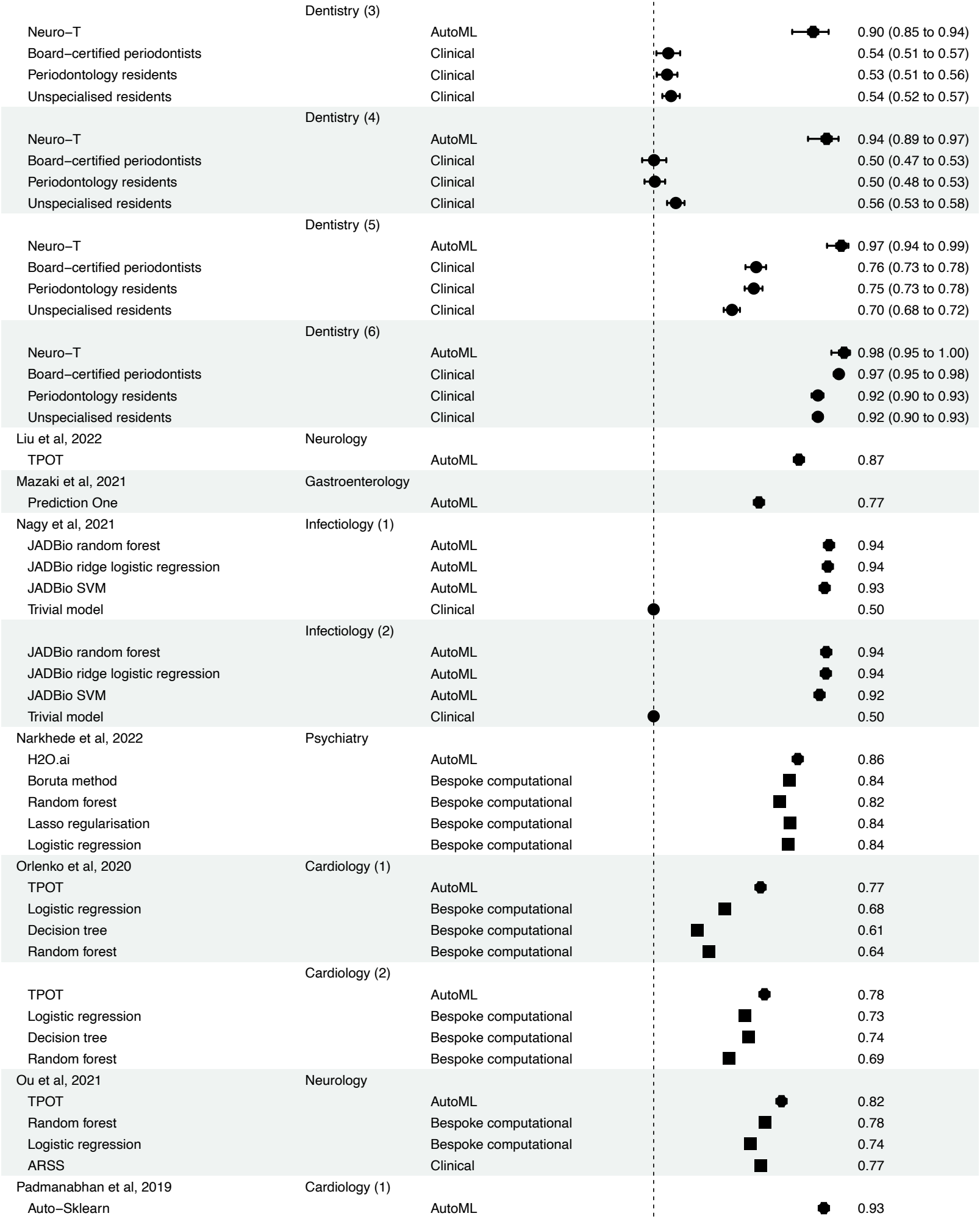

Search, screening, and inclusion process
Identification
Records identified (N = 2417):
Cochrane (n = 237)
Embase (n = 882)
MEDLINE (n = 632)
Scopus (n = 664)
Other sources (n = 2)
Duplicates removed (n = 1286):
Zotero (n = 1259)
Rayyan (n = 27)
Screening
Abstracts screened (n = 1131)
Records excluded (n = 862)
Reports sought for retrieval (n = 269)
Reports not retrieved (n = 1)
Reports assessed for eligibility (n = 268)
Reports excluded (n = 186):
Not in the English language (n = 1)
Not a peer reviewed article (n = 33)
Is a retracted article (n = 0)
Does not utilise autoML (n = 126)
Non-clinical application (n = 26)
Inclusion
Studies included in review[10,16,21-100] (n = 82)


(A)
Frequency
Year
2018
2019
2020
2021
2022
(B)
Continent
South America
North America
Europe
Asia
Frequency
(C)
Country
USA
China
Greece
Korea
Canada
Japan
UK
Other
Frequency
(D)
Funding
Academic
Commercial
Mixed
None
Not stated
Frequency
(E)
Specialty
Anaesthesiology
Breast
Cardiology
Dentistry
Dermatology
Endocrinology
Gastroenterology
Gynaecology
Infectiology
Nephrology
Neurology
Ophthalmology
Otolaryngology
Psychiatry
Pulmonology
Urology
Other
Frequency
(F)
Frequency
DECIDE-AI Score (max. 27)
(G)
PROBAST Evaluation
High
Unclear
Low
Domain
Overall CoA
CoA of Outcomes
CoA of Predictors
CoA of Participants
Overall RoB
RoB in Analysis
RoB in Outcome
RoB in Predictors
RoB in Participants
Frequency
(H)
Dataset
External Validation
Internal Validation
Training
Sample Size
Frequency
(I)
Data nature
Unstructured
Structured
Frequency
(J)
Data Source
Study
Registry
Publically available
Frequency

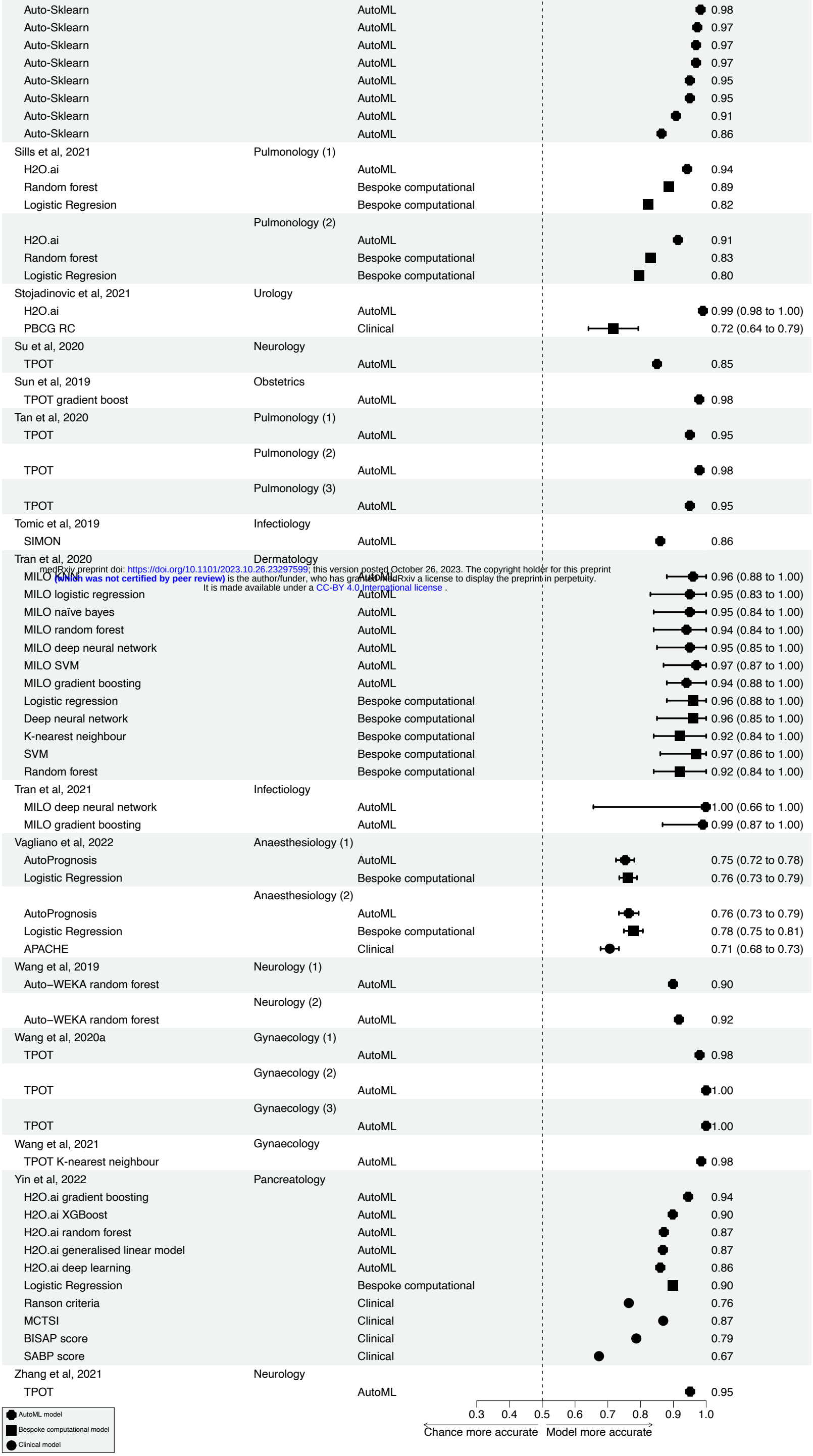
Platform	Specialty	Modality	Area under the receiver operating characteristic curve (95% CI)
Abbas et al, 2022	Ophthalmology		
Cloud AutoML		AutoML	0.85
XGBoost		Bespoke computational	0.85
Alaa and van der Schaar, 2018	Pulmonology		
AutoPrognosis		AutoML	0.89 (0.88 to 0.90)
TPOT		AutoML	0.84 (0.83 to 0.85)
SVM		Bespoke computational	0.84 (0.81 to 0.87)
Gradient boosting		Bespoke computational	0.87 (0.85 to 0.89)
Bagging		Bespoke computational	0.83 (0.80 to 0.86)
Nkam method		Clinical	0.86 (0.85 to 0.87)
Buzzetti method		Clinical	0.83 (0.82 to 0.84)
CF–ABLE–UK method		Clinical	0.77 (0.76 to 0.78)
FEV1% predicted criterion		Clinical	0.70 (0.69 to 0.71)
Alaa et al, 2019	Cardiology		
AutoPrognosis		AutoML	0.77 (0.77 to 0.78)
Framingham Score		Clinical	0.72 (0.72 to 0.73)
Cox PH Model		Clinical	0.76 (0.75 to 0.76)
SVM		Bespoke computational	0.71 (0.65 to 0.77)
Random forest		Bespoke computational	0.73 (0.73 to 0.73)
AdaBoost		Bespoke computational	0.76 (0.76 to 0.76)
Gradient boosting		Bespoke computational	0.77 (0.76 to 0.77)
Neural network		Bespoke computational	0.76 (0.76 to 0.75)
An et al, 2021	Pulmonology		
KNIME		AutoML	0.97
Antaki et al, 2020	Ophthalmology (1)		
MATLAB quadratic SVM		AutoML	0.90
MATLAB optimised naïve Bayes		AutoML	0.86
	Ophthalmology (2)		
MATLAB optimised SVM		AutoML	0.81
MATLAB optimised naïve Bayes		AutoML	0.81
Chou et al, 2022	Neurology		
DataRobot linear regression		AutoML	0.68 (0.66 to 0.70)
DataRobot parsimonious linear regression		AutoML	0.84 (0.82 to 0.86)
DataRobot eXtreme gradient boost		AutoML	0.87 (0.86 to 0.88)
Danilatou et al, 2022	Anaesthesiology		
JADBio		AutoML	0.89 (0.87 to 0.91)
XGBoost		Bespoke computational	0.84 (0.83 to 0.85)
Feretzakis et al, 2021	Infectiology		
Azure StackEnsemble		AutoML	0.85
Azure VotingEnsemble		AutoML	0.85
Azure LightGBM		AutoML	0.84
Azure XGBoostClassifier		AutoML	0.84
Hasimbegovic et al, 2021	Cardiology		
Dedicaid		AutoML	0.91
Hu et al, 2022	Hepatology (1)		
TPOT		AutoML	0.76
Radiomics pipeline		Bespoke computational	0.80
	Hepatology (2)		
TPOT		AutoML	0.79
Radiomics pipeline		Bespoke computational	0.79
Ikemura et al, 2021	Pulmonology		
H20.ai stacked ensemble		AutoML	0.92
GBM		Bespoke computational	0.91
DRF		Bespoke computational	0.91
XGBoost		Bespoke computational	0.91
XRT		Bespoke computational	0.90

Jen et al, 2021	Nephrology		
MILO neural network		AutoML	0.76 (0.71 to 0.81)
MILO GBM		AutoML	0.73 (0.70 to 0.79)
MILO KNN		AutoML	0.75 (0.70 to 0.80)
MILO naïve bayes		AutoML	0.74 (0.69 to 0.78)
MILO random forest		AutoML	0.75 (0.70 to 0.79)
MILO SVM		AutoML	0.75 (0.70 to 0.80)
MILO logistic regresion		AutoML	0.75 (0.70 to 0.80)
Chapal method		Bespoke computational	0.73
Irish method 1		Bespoke computational	0.70
Irish method 2		Bespoke computational	0.70
Jeldres method		Bespoke computational	0.74
Zaza method		Bespoke computational	0.63
Karaglani et al, 2020	Neurology (1)		
JADBio SVM		AutoML	0.97 (0.91 to 1.00)
	Neurology (2)		
JADBio Random forest		AutoML	0.85 (0.78 to 0.91)
	Neurology (3)		
JADBio Ridge logistic regression		AutoML	0.92 (0.85 to 0.97)
Karaglani et al, 2022	Endocrinology		
JADBio random forest		AutoML	0.93 (0.87 to 0.97)
JADBio ridge logistic regression		AutoML	0.92 (0.87 to 0.96)
Age		Clinical	0.57 (0.47 to 0.66)
BMI		Clinical	0.66 (0.57 to 0.74)
ccfDNA		Clinical	0.53 (0.44 to 0.62)
INS		Clinical	0.65 (0.56 to 0.74)
IAPP		Clinical	0.73 (0.65 to 0.81)
GCK		Clinical	0.85 (0.79 to 0.91)
KCNJ11		Clinical	0.71 (0.62 to 0.81)
ABCC8		Clinical	0.53 (0.44 to 0.62)
Karhade et al, 2021	Dentistry (1)		
Cloud AutoML		AutoML	0.74
	Dentistry (2)		
Cloud AutoML		AutoML	0.80
Karstoft et al, 2020	Psychiatry (1)		
JADBio random forest		AutoML	0.76 (0.67 to 0.84)
	Psychiatry (2)		
JADBio random forest		AutoML	0.70 (0.60 to 0.80)
Katsuki et al, 2021	Neurology		
Prediction One		AutoML	0.80
SAFIRE score		Clinical	0.89
Prediction One		AutoML	0.65
Fisher CT scale		Clinical	0.54
Katsuki and Matsuo, 2021	Pulmonology		
Prediction One		AutoML	0.98
Kumar et al, 2022	Endocrinology		
TPOT		AutoML	0.93
Lee et al, 2020	Dentistry (1)		
Neuro–T		AutoML	0.94 (0.90 to 0.97)
Board–certified periodontists		Clinical	0.90 (0.88 to 0.92)
Periodontology residents		Clinical	0.83 (0.81 to 0.85)
Unspecialised residents		Clinical	0.78 (0.76 to 0.80)
	Dentistry (2)		
Neuro–T		AutoML	0.91 (0.86 to 0.95)
Board–certified periodontists		Clinical	0.79 (0.77 to 0.81)
Periodontology residents		Clinical	0.81 (0.79 to 0.83)
Unspecialised residents		Clinical	0.74 (0.72 to 0.76)



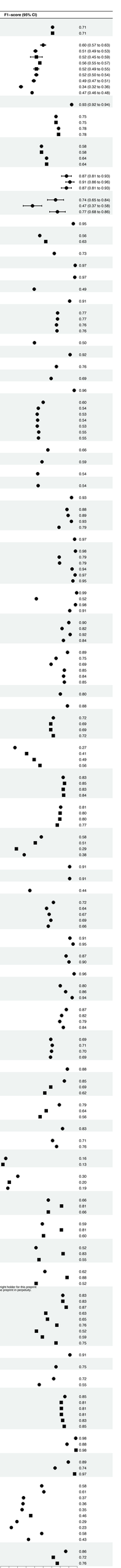

Dentistry (3)
Neuro−T AutoML 0.90 (0.85 to 0.94)
Board−certified periodontists Clinical 0.54 (0.51 to 0.57)
Periodontology residents Clinical 0.53 (0.51 to 0.56)
Unspecialised residents Clinical 0.54 (0.52 to 0.57)
Dentistry (4)
Neuro−T AutoML 0.94 (0.89 to 0.97)
Board−certified periodontists Clinical 0.50 (0.47 to 0.53)
Periodontology residents Clinical 0.50 (0.48 to 0.53)
Unspecialised residents Clinical 0.56 (0.53 to 0.58)
Dentistry (5)
Neuro−T AutoML 0.97 (0.94 to 0.99)
Board−certified periodontists Clinical 0.76 (0.73 to 0.78)
Periodontology residents Clinical 0.75 (0.73 to 0.78)
Unspecialised residents Clinical 0.70 (0.68 to 0.72)
Dentistry (6)
Neuro−T AutoML 0.98 (0.95 to 1.00)
Board−certified periodontists Clinical 0.97 (0.95 to 0.98)
Periodontology residents Clinical 0.92 (0.90 to 0.93)
Unspecialised residents Clinical 0.92 (0.90 to 0.93)
Liu et al, 2022 Neurology
TPOT AutoML 0.87
Mazaki et al, 2021 Gastroenterology
Prediction One AutoML 0.77
Nagy et al, 2021 Infectiology (1)
JADBio random forest AutoML 0.94
JADBio ridge logistic regression AutoML 0.94
JADBio SVM AutoML 0.93
Trivial model Clinical 0.50
Infectiology (2)
JADBio random forest AutoML 0.94
JADBio ridge logistic regression AutoML 0.94
JADBio SVM AutoML 0.92
Trivial model Clinical 0.50
Narkhede et al, 2022 Psychiatry
H2O.ai AutoML 0.86
Boruta method Bespoke computational 0.84
Random forest Bespoke computational 0.82
Lasso regularisation Bespoke computational 0.84
Logistic regression Bespoke computational 0.84
Orlenko et al, 2020 Cardiology (1)
TPOT AutoML 0.77
Logistic regression Bespoke computational 0.68
Decision tree Bespoke computational 0.61
Random forest Bespoke computational 0.64
Cardiology (2)
TPOT AutoML 0.78
Logistic regression Bespoke computational 0.73
Decision tree Bespoke computational 0.74
Random forest Bespoke computational 0.69
Ou et al, 2021 Neurology
TPOT AutoML 0.82
Random forest Bespoke computational 0.78
Logistic regression Bespoke computational 0.74
ARSS Clinical 0.77
Padmanabhan et al, 2019 Cardiology (1)
Auto−Sklearn AutoML 0.93

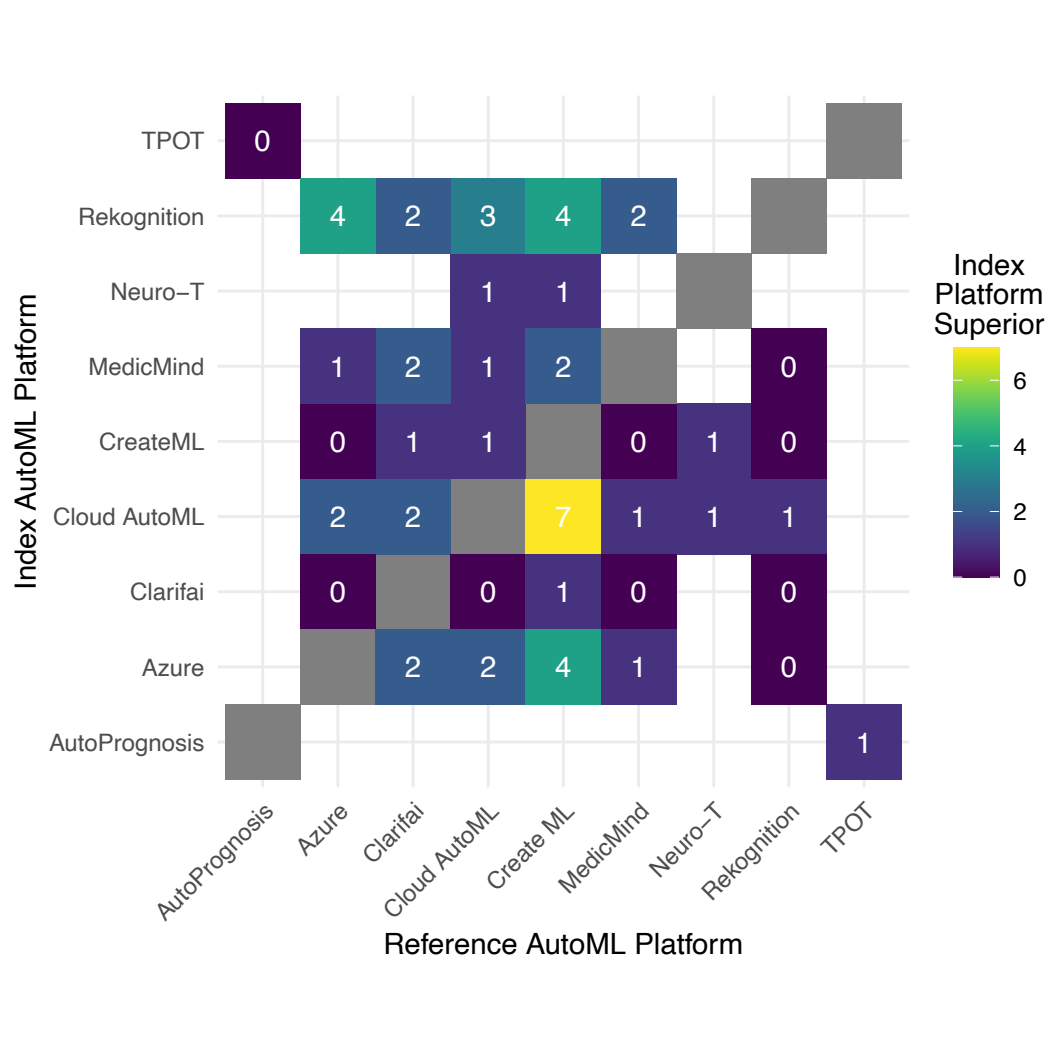
Scikit learn		Bespoke computational	0.82
	Cardiology (2)		
Auto−Sklearn		AutoML	0.80
Scikit learn		Bespoke computational	0.73
Panagopoulou et al, 2021a	Gastrointestinal		
JADBio Logistic regression		AutoML	0.93 (0.89 to 0.97)
JADBio random forest		AutoML	0.93 (0.87 to 0.97)
Panagopoulou et al, 2021b	Breast (1)		
JADBio		AutoML	0.99 (0.98 to 1.00)
	Breast (2)		
JADBio		AutoML	0.99 (0.92 to 1.00)
	Breast (3)		
JADBio		AutoML	0.97 (0.92 to 1.00)
Papoutsoglou et al, 2021	Pulmonology (1)		
JADBio ridge logistic regression		AutoML	0.88
JADBio SVM		AutoML	0.92
Shen method		Bespoke computational	0.88
	Pulmonology (2)		
JADBio Random forest		AutoML	0.94
Shen method		Bespoke computational	0.94
	Pulmonology (3)		
JADBio random forest		AutoML	0.98
Peng et al, 2022	Cardiology (1)		
TPOT		AutoML	0.94
	Cardiology (2)		
TPOT		AutoML	0.96
Purkayastha et al, 2020	Nephrology		
TPOT		AutoML	0.60 (0.50 to 0.69)
Bayesian Classifier		Bespoke computational	0.59 (0.49 to 0.68)
Radzi et al, 2021	Breast		
TPOT		AutoML	0.94
SVM		Bespoke computational	0.50
Multi−layer perceptron		Bespoke computational	0.50
Naïve Bayes		Bespoke computational	0.72
Rallabandi et al, 2020	Neurology		
Auto−WEKA SVM		AutoML	0.76
Auto−WEKA Naïve Bayes		AutoML	0.68
Auto−WEKA K−nearest neighbour		AutoML	0.70
Auto−WEKA Random forest		AutoML	0.71
Auto−WEKA Decision tree		AutoML	0.70
Rashidi et al, 2021a	Infectiology (1)		
MILO neural network		AutoML	0.94
MILO logistic Regression		AutoML	0.97
Rashidi et al, 2021b	Nephrology		
MILO logistic regresion		AutoML	0.96
Rashidi et al, 2022	Infectiology		
MILO gradient boosting		AutoML	0.95 (0.87 to 1.00)
MILO random forest		AutoML	0.96 (0.82 to 1.00)
Random forest		Bespoke computational	0.97 (0.94 to 1.00)
Real et al, 2022	Dentistry (1)		
Auto−WEKA bagging		AutoML	0.86
Auto−WEKA random committee		AutoML	0.90
Auto−WEKA multilayer perceptron		AutoML	0.92
	Dentistry (2)		
Auto−WEKA logistic model tree		AutoML	0.91
Auto−WEKA reduced error pruning tree		AutoML	0.79
Auto−WEKA J48		AutoML	0.79
Auto−WEKA random tree		AutoML	0.90
	Dentistry (3)		
Auto−WEKA sequential minimal optimisation		AutoML	0.74
Auto−WEKA multilayer perceptron		AutoML	0.35
Auto−WEKA adaboost		AutoML	0.72
Auto−WEKA bagging		AutoML	0.74
Ritter et al, 2022	Haematology		
Dedicaid		AutoML	0.85
Shen et al, 2020	Pathology		
Auto-Sklearn		AutoML	0.99



medRxiv preprint doi: https://doi.org/10.1101/2023.10.26.23297599; this version posted October 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Platform	Specialty	Modality	F1–score (95% CI)
Abbas et al, 2022	Ophthalmology		
Cloud AutoML		AutoML	0.71
XGBoost		Bespoke computational	0.71
Alaa and van der Schaar, 2018	Pulmonology		
AutoPrognosis		AutoML	0.60 (0.57 to 0.63)
TPOT		AutoML	0.51 (0.49 to 0.53)
SVM		Bespoke computational	0.52 (0.45 to 0.59)
Gradient boosting		Bespoke computational	0.56 (0.55 to 0.57)
Bagging		Bespoke computational	0.52 (0.49 to 0.55)
Naum method		Clinical	0.52 (0.50 to 0.54)
Buzzetti method		Clinical	0.49 (0.47 to 0.51)
CF–ABLE–UK method		Clinical	0.34 (0.32 to 0.36)
FEV1% predicted criterion		Clinical	0.47 (0.46 to 0.48)
An et al, 2021	Pulmonology		
KNIME		AutoML	0.93 (0.92 to 0.94)
Antaki et al, 2020	Ophthalmology (1)		
MATLAB quadratic SVM		AutoML	0.75
Manual quadratic SVM		Bespoke computational	0.75
MATLAB optimised naïve Bayes		AutoML	0.78
Manual optimised naïve Bayes		Bespoke computational	0.78
	Ophthalmology (2)		
MATLAB optimised SVM		AutoML	0.58
Manual optimised SVM		Bespoke computational	0.58
MATLAB optimised naïve Bayes		AutoML	0.64
Manual optimised naïve Bayes		Bespoke computational	0.64
Bang et al, 2021	Gastrointestinal (1)		
Cloud AutoML		AutoML	0.87 (0.81 to 0.93)
Neuro–T		AutoML	0.91 (0.86 to 0.96)
Create ML Image Classifier		AutoML	0.87 (0.81 to 0.93)
	Gastrointestinal (2)		
Cloud AutoML		AutoML	0.74 (0.65 to 0.84)
Neuro–T		AutoML	0.47 (0.37 to 0.58)
Create ML Image Classifier		AutoML	0.77 (0.68 to 0.86)
Borkowski et al, 2020	Pulmonology		
Azure		AutoML	0.95
Danilatou et al, 2022	Anesthesiology		
JADBio		AutoML	0.56
XGBoost		Bespoke computational	0.63
Faes et al, 2019	Ophthalmology (1)		
Cloud AutoML		AutoML	0.73
	Ophthalmology (2)		
Cloud AutoML		AutoML	0.97
	Pulmonology (1)		
Cloud AutoML		AutoML	0.97
	Pulmonology (2)		
Cloud AutoML		AutoML	0.49
	Dermatology		
Cloud AutoML		AutoML	0.91
Feretzakis et al, 2021	Infectiology		
Azure StackEnsemble		AutoML	0.77
Azure VotingEnsemble		AutoML	0.77
Azure LightGBM		AutoML	0.76
Azure XGBoostClassifier		AutoML	0.76
Ghosh et al, 2021	Pulmonology		
Cloud AutoML		AutoML	0.50
Hasimbegovic et al, 2021	Cardiology		
Dedicaid AutoMl		AutoML	0.92
Ito et al, 2022	Dermatology		
Cloud AutoML		AutoML	0.76
Ito et al, 2021	Urology (1)		
Cloud AutoML		AutoML	0.69
	Urology (2)		
Cloud AutoML		AutoML	0.96
Jen et al, 2021	Nephrology		
MILO neural network		AutoML	0.60
MILO GBM		AutoML	0.54
MILO KNN		AutoML	0.53
MILO naïve bayes		AutoML	0.54
MILO random forest		AutoML	0.53
MILO SVM		AutoML	0.55
MILO logistic regresion		AutoML	0.55
Karhade et al, 2021	Dentistry (1)		
Cloud AutoML		AutoML	0.66
	Dentistry (2)		
Cloud AutoML		AutoML	0.59
Karstoft et al, 2020	Psychiatry (1)		
JADBio random forest		AutoML	0.54
	Psychiatry (2)		
JADBio random forest		AutoML	0.54
Katsuki and Matsuo, 2021	Pulmonology		
Prediction One		AutoML	0.93
Kim et al, 2021	Ophthalmology		
Cloud AutoML 1		AutoML	0.88
Cloud AutoML 2		AutoML	0.89
Retina specialist		Clinical	0.93
Ophthalmology residents		Clinical	0.79
Koga et al, 2021	Neurology		
Cloud AutoML		AutoML	0.97
Korot et al, 2021	Ophthalmology (1)		
Rekognition		AutoML	0.98
Create ML		AutoML	0.79
Clarifai		AutoML	0.79
Cloud AutoML		AutoML	0.94
MedicMind		AutoML	0.97
Azure		AutoML	0.95
	Ophthalmology (2)		
Rekognition		AutoML	0.99
Create ML		AutoML	0.52
Cloud AutoML		AutoML	0.98
Azure		AutoML	0.91
	Ophthalmology (3)		
Rekognition		AutoML	0.90
Create ML		AutoML	0.82
Cloud AutoML		AutoML	0.92
Azure		AutoML	0.84
	Ophthalmology (4)		
Rekognition		AutoML	0.89
Create ML		AutoML	0.75
Clarifai		AutoML	0.69
Cloud AutoML		AutoML	0.85
MedicMind		AutoML	0.84
Azure		AutoML	0.85
Liu et al, 2022	Neurology		
TPOT		AutoML	0.80
Livingstone and Chau, 2020	Otolaryngology		
Cloud AutoML		AutoML	0.88
Mohsen et al, 2022	Endocrinology		
H2O.ai		AutoML	0.72
Random forest		Bespoke computational	0.69
AdaBoost		Bespoke computational	0.69
Support Vector Classifier		Bespoke computational	0.72
Nero et al, 2020	Gynaecology		
TPOT		AutoML	0.27
XGBoost		Bespoke computational	0.41
Logistic regression		Bespoke computational	0.49
SVM		Bespoke computational	0.56
Orlenko et al, 2020	Cardiology (1)		
TPOT		AutoML	0.83
Logistic regression		Bespoke computational	0.85
Decision tree		Bespoke computational	0.83
Random forest		Bespoke computational	0.84
	Cardiology (2)		
TPOT		AutoML	0.81
Logistic regression		Bespoke computational	0.80
Decision tree		Bespoke computational	0.80
Random forest		Bespoke computational	0.77
Ou et al, 2021	Neurology		
TPOT		AutoML	0.58
Random forest		Bespoke computational	0.51
Logistic regression		Bespoke computational	0.29
ARSS		Clinical	0.38
Peng et al, 2022	Cardiology (1)		
TPOT		AutoML	0.91
	Cardiology (2)		
TPOT		AutoML	0.91
Purkayastha et al, 2020	Nephrology		
TPOT		AutoML	0.44
Rallabandi et al, 2020	Neurology		
Auto–WEKA SVM		AutoML	0.72
Auto–WEKA Naïve Bayes		AutoML	0.64
Auto–WEKA K–nearest neighbour		AutoML	0.67
Auto–WEKA Random forest		AutoML	0.69
Auto–WEKA Decision tree		AutoML	0.66
Rashidi et al, 2021a	Infectiology (1)		
MILO neural network		AutoML	0.91
MILO logistic Regression		AutoML	0.95
	Infectiology (2)		
MILO neural network		AutoML	0.87
MILO logistic Regression		AutoML	0.90
Rashidi et al, 2021b	Nephrology		
MILO logistic regresion		AutoML	0.96
Real et al, 2022	Dentistry (1)		
Auto–WEKA bagging		AutoML	0.80
Auto–WEKA random committee		AutoML	0.86
Auto–WEKA multilayer perceptron		AutoML	0.94
	Dentistry (2)		
Auto–WEKA logistic model tree		AutoML	0.87
Auto–WEKA reduced-error pruning tree		AutoML	0.82
Auto–WEKA J48		AutoML	0.79
Auto–WEKA random tree		AutoML	0.84
	Dentistry (3)		
Auto–WEKA sequential minimal optimisation		AutoML	0.69
Auto–WEKA multilayer perceptron		AutoML	0.71
Auto–WEKA adaboost		AutoML	0.70
Auto–WEKA bagging		AutoML	0.69
Sakaguanni et al, 2020	Pulmonology		
Cloud AutoML		AutoML	0.88
Sills et al, 2021	Pulmonology (1)		
H2O.ai		AutoML	0.85
Random forest		Bespoke computational	0.69
Logistic Regresion		Bespoke computational	0.62
	Pulmonology (2)		
H2O.ai		AutoML	0.79
Random forest		Bespoke computational	0.64
Logistic Regresion		Bespoke computational	0.56
Su et al, 2020	Neurology		
TPOT		AutoML	0.83
Tahmasebi et al, 2021	Breast		
Cloud AutoML		AutoML	0.71
Radiologist		Clinical	0.76
Vagliano et al, 2022	Anaesthesiology (1)		
AutoPrognosis		AutoML	0.16
Logistic Regression		Bespoke computational	0.13
	Anaesthesiology (2)		
AutoPrognosis		AutoML	0.30
Logistic Regression		Bespoke computational	0.20
APACHE		Clinical	0.19
van Eeden et al, 2021	Psychiatry (1)		
Auto–Sklearn		AutoML	0.66
Logistic regression		Bespoke computational	0.81
Naïve Bayes		Bespoke computational	0.66
	Psychiatry (2)		
Auto–Sklearn		AutoML	0.59
Logistic regression		Bespoke computational	0.81
Naïve Bayes		Bespoke computational	0.60
	Psychiatry (3)		
Auto–Sklearn		AutoML	0.52
Logistic regression		Bespoke computational	0.83
Naïve Bayes		Bespoke computational	0.55
	Psychiatry (4)		
Auto–Sklearn		AutoML	0.62
Logistic regression		Bespoke computational	0.88
Naïve Bayes		Bespoke computational	0.52
Wang et al, 2020a	Breast (1)		
Cloud AutoML		AutoML	0.91
	Breast (2)		
Cloud AutoML		AutoML	0.75
Wang et al, 2020c	Otolaryngology		
Cloud AutoML		AutoML	0.72
TI–RADS classification		Clinical	0.55
Xavier and Chen, 2022	Radiology		
Cloud AutoML		AutoML	0.85
Random forest		Bespoke computational	0.81
Tree ensemble		Bespoke computational	0.81
Gradient boosting		Bespoke computational	0.81
Multi–layer perceptron		Bespoke computational	0.83
Universal language model fine–tuning		Bespoke computational	0.85
Yang et al, 2021	Neurology (1)		
Cloud AutoML		AutoML	0.98
Create ML		AutoML	0.88
Convolutional neural network		Bespoke computational	0.98
	Neurology (2)		
Cloud AutoML		AutoML	0.89
Create ML		AutoML	0.74
Convolutional neural network		Bespoke computational	0.97
Yin et al, 2022	Pancreatology		
H2O ai gradient boosting		AutoML	0.58
H2O ai XGBoost		AutoML	0.61
H2O ai random forest		AutoML	0.37
H2O ai generalised linear model		AutoML	0.36
H2O.ai deep learning		AutoML	0.35
Logistic Regression		Bespoke computational	0.46
Ranson criteria		Clinical	0.29
MCTSI		Clinical	0.23
BISAP score		Clinical	0.58
SABP score		Clinical	0.43
Zeng and Zhang, 2020	Breast		
Cloud AutoML		AutoML	0.86
Cruz–Roa method		Bespoke computational	0.72
Janowczyk method		Bespoke computational	0.76




TPOT
Rekognition
Neuro−T
MedicMind
CreateML
Cloud AutoML
Clarifai
Azure
AutoPrognosis
Index AutoML Platform
AutoPrognosis
Azure
Clarifai
Cloud AutoML
Create ML
MedicMind
Neuro−T
Rekognition
TPOT
Reference AutoML Platform
Index Platform Superior
6
4
2
0