Received: 8 April 2024 | Accepted: 24 June 2024

DOI: 10.1002/med4.75

REVIEW

MEDICINE ADVANCES

Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare

Han Yuan[1] | **Kunyu Yu**[1] | **Feng Xie**[1,2,3] | **Mingxuan Liu**[1] | **Shenghuan Sun**[4]

[1]Duke-NUS Medical School, Centre for Quantitative Medicine, Singapore, Singapore

[2]Department of Biomedical Data Science, Stanford University, Stanford, California, USA

[3]Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford University, Stanford, California, USA

[4]Bakar Computational Health Sciences Institute, University of California, San Francisco, California, USA

Correspondence
Han Yuan, Duke-NUS Medical School, Centre for Quantitative Medicine, 8 College Road, Singapore 169857, Singapore.
Email: yuan.han@u.duke.nus.edu

Abstract

Machine learning (ML) has achieved substantial success in performing healthcare tasks in which the configuration of every part of the ML pipeline relies heavily on technical knowledge. To help professionals with borderline expertise to better use ML techniques, Automated ML (AutoML) has emerged as a prospective solution. However, most models generated by AutoML are black boxes that are challenging to comprehend and deploy in healthcare settings. We conducted a systematic review to examine AutoML with interpretation systems for healthcare. We searched four databases (MEDLINE, EMBASE, Web of Science, and Scopus) complemented with seven prestigious ML conferences (AAAI, ACL, ICLR, ICML, IJCAI, KDD, and NeurIPS) that reported AutoML with interpretation for healthcare before September 1, 2023. We included 118 articles related to AutoML with interpretation in healthcare. First, we illustrated AutoML techniques used in the included publications, including automated data preparation, automated feature engineering, and automated model development, accompanied by a real-world case study to

Abbreviations: AAAI, annual AAAI conference on artificial intelligence; ACL, annual meeting of the association for computational linguistics; AutoML, automated machine learning; ANN, artificial neural networks; AUPRC, area under the precision recall curve; AUROC, area under the receiver operating characteristic curve; BMI, brain machine interfaces; CNN, convolutional neural networks; ChIP-seq, chromatin immunoprecipitation sequences; DNA-seq, DNA sequences; DNase-seq, DNase I hypersensitive site sequences; DSC, dice similarity coefficient; EMG, electromyogram; ECG, electrocardiogram; FE, feature engineering; fMRI, functional magnetic resonance imaging; GMM, Gaussian mixture model; GAN, generative adversarial network; GAT, graph attention network; GBM, gradient boosting machine; GNB, Gaussian Naive Bayes; GRU, gated recurrent unit; HD, hausdorff distance; HLAN, hierarchical label-wise attention network; ICD, international classification of diseases; IB, information bottleneck; ICLR, international conference on learning representations; ICML, international conference on machine learning; IJCAI, international joint conference on artificial intelligence; IOU, intersection over union; KDD, ACM SIGKDD conference on knowledge discovery & data mining; LDA, linear discriminant analysis; LSTM, long short-term memory; LR, logistic regression; LASSO, least absolute shrinkage and selection operator; MDL, minimum description length; ML, machine learning; MLP, multilayer perceptron; MSE, mean squared error; MNase-seq, micrococcal nuclease digestion with deep sequencing; NeurIPS, annual conference on neural information processing systems; OCT, optical coherence tomography; PRISMA, preferred reporting items for systematic reviews and meta-analyses; PSD, predictive sparse decomposition; PSNR, peak signal-to-noise ratio; PNN, probabilistic neural networks; R-CNN, region convolutional neural networks; RF, random forest; RMSE, root mean squared error; RNA-seq, RNA sequences; RNN, recurrent neural networks; SVM, support vector machine; SHAP, shapley additive explanations; TPOT, tree-based pipeline optimization tool; VAE, variational autoencoder.

Han Yuan, Kunyu Yu and Feng Xie are contributed equally.

demonstrate the advantages of AutoML over classic ML. Then, we summarized interpretation methods: feature interaction and importance, data dimensionality reduction, intrinsically interpretable models, and knowledge distillation and rule extraction. Finally, we detailed how AutoML with interpretation has been used for six major data types: image, free text, tabular data, signal, genomic sequences, and multi-modality. To some extent, AutoML with interpretation provides effortless development and improves users' trust in ML in healthcare settings. In future studies, researchers should explore automated data preparation, seamless integration of automation and interpretation, compatibility with multi-modality, and utilization of foundation models.

KEYWORDS
automated machine learning, explainable machine learning, health informatics

1 | INTRODUCTION

The rapid growth of biomedical big data has led to greater opportunities for the deployment of modern data-driven technologies such as machine learning (ML) [1]. ML techniques have achieved substantial success in processing various types of data and performing diverse tasks in the context of healthcare [2–4]. However, the effective exploitation of healthcare data by ML models necessitates the rigorous configuration of every part of the ML pipeline, which relies heavily on specialized technical knowledge and extensive effort.

To help professionals with borderline expertise in data science to better use ML techniques, Automated ML (AutoML) has emerged as a prospective solution. The objective of AutoML, as defined by Yao et al. [5], is to allow computer programs to replace human tuning in the process of determining all or a part of model configurations while maintaining good performance and high computational efficiency. Configurations in this context refer to all factors that are specified prior to model training and affect the final performance, including input data, feature sets, hyperparameters, and model architectures. Therefore, a complete AutoML pipeline encompasses the automation of data preparation, feature engineering, and model development [6].

AutoML techniques in the ML pipeline cater to various levels of coding proficiency. For example, sophisticated AutoML methods such as NASLib [7], which require advanced programming knowledge, aim to provide greater flexibility for experienced ML engineers. AutoML software packages such as auto-sklearn [8] focus primarily on model development, that is, algorithm selection and hyperparameter optimization, targeting users with mediate coding skills. Additionally, commercial AutoML platforms such as Google Cloud's AutoML system and H2O Driverless artificial intelligence (AI) offer no-coding solutions, featuring user-friendly interfaces and rapid convergence capabilities. Table 1 provides an overview of the toolkits developed by leading companies.

In healthcare, the extensive application of ML significantly amplifies the advantages of implementing AutoML approaches. It enables healthcare professionals with borderline ML knowledge to build high-quality models using a fully automated pipeline [9] and further addresses privacy concerns without sharing data with external ML engineers. AutoML systems effectively fill the gap between the lack of ML expertise among healthcare practitioners and the need for data analytics based on ML models [10]. AutoPrognosis [11] describes an end-to-end diagnosis and prognosis modeling framework that helps healthcare professionals leverage clinical data for risk prediction across diverse clinical settings. Additionally, AutoML improves the efficiency of ML engineers by automating tedious and time-consuming tasks such as data preprocessing [12]. For example, nnU-Net [13] introduces a self-configured biomedical image segmentation method that automates the conversion of raw image data into representative structured features.

Although AutoML systems help both healthcare professionals and ML engineers to process medical data effortlessly, the interpretability of these systems should be improved to boost confidence in the reliability of the generated ML models [14]. Given the potentially serious consequences of medical AI failures, greater demands are being placed on the interpretation of ML models in clinical decision-making to fulfill both medical validation and regulatory requirements. Thus, in contrast to conventional AutoML systems primarily centered on ML development, AutoML with interpretation aligns more closely with the real-world requirements in healthcare settings [15].

TABLE 1 Overview of AutoML toolkits from leading companies.

Company	Toolkit	Modality	Website
Amazon	AutoGluon	Multi-modality	https://auto.gluon.ai/stable/index.html
	SageMaker	Multi-modality	https://aws.amazon.com/sagemaker/canvas/
Apple	Create ML	Multi-modality	https://developer.apple.com/machine-learning/create-ml/
Google	Vertex AI	Multi-modality	https://cloud.google.com/vertex-ai?hl=en
IBM	AutoAI	Tabular data	https://www.ibm.com/products/watson-studio/autoai
	watsonx.ai	Multi-modality	https://www.ibm.com/products/watsonx-ai
Meta	Looper	Multi-modality	https://research.facebook.com/publications/looper-an-end-to-end-ml-platform-for-product-decisions/
Microsoft	Azure machine learning	Multi-modality	https://azure.microsoft.com/en-us/products/machine-learning/automatedml/#overview
NVIDIA	TAO	Multi-modality	https://developer.nvidia.com/tao-toolkit

Note: The companies are listed alphabetically for ease of reference.

Because AutoML systems with interpretation are fundamental to facilitating the clinical adoption of AI technologies, we conducted this review to gain insight into how they empower the health community by lowering the entry barrier and enhancing the credibility of ML algorithms. In recent years, several researchers [6, 9, 10, 16–19] have reviewed the development and application of either AutoML or ML interpretations. However, none have provided a systematic and in-depth summary of AutoML with interpretation, particularly its applications in healthcare. In our review, we aim to integrate existing research practices by categorizing data modalities, AutoML techniques, and interpretation methods to acquire a comprehensive understanding of AutoML with interpretation in healthcare and inspire future research topics. The purpose of the categorization is to provide practitioners with an insight into how various AutoML with interpretation systems have been implemented in different medical tasks.

The promising application of AutoML with interpretation in healthcare necessitates a systematic review of cutting-edge research to bridge the gap between technical innovation and practical application. We envisage that this review will empower healthcare practitioners by providing well-organized and referable information about AutoML with interpretation systems, and further facilitate the real-world deployment of ML systems in diverse healthcare settings.

2 | METHODS

2.1 | Search strategy and data sources

We conducted a systematic review that encompassed both methodology and application studies on AutoML with interpretation for healthcare. We performed a literature search on four databases: MEDLINE, EMBASE, Web of Science, and Scopus. Given that some of the latest ML research is often presented at conferences and may not be included in these four databases, we also searched for research papers in the proceedings of seven relevant and prestigious ML conferences: AAAI, ACL, ICLR, ICML, IJCAI, KDD, and NeurIPS. The searched terms in the medical domain were ("medical" OR "clinical" OR "health" OR "healthcare" OR "medicine"). We also added the terms ("ML" OR "deep learning" OR "AI") to limit the search to ML-based studies, and ("automated" OR "automatic") AND ("interpretable" OR "explainable" OR "interpretability") to include studies on AutoML with interpretation. We restricted our search to papers published before September 1, 2023.

2.2 | Inclusion and exclusion criteria

We followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses guidelines [20] to conduct the systematic review. We included all papers published in English that used AutoML with interpretation to perform healthcare tasks. We excluded review articles, workshop papers, duplicate records, and studies not relevant to AutoML with interpretation or healthcare. Each article was independently screened by at least two reviewers, and, if ambiguous, discussed with the corresponding author to reach a consensus.

2.3 | Data analysis

Table 2 presents our evaluation and summary of the papers from three aspects: AutoML techniques, interpretation methods, and target data types. For AutoML techniques, we identified three main research directions: automated data preparation, automated feature engineering, and automated model development [9]. For interpretation methods, we summarized from four angles: knowledge distillation and rule extraction, intrinsically interpretable models, data dimensionality reduction, and feature interaction and importance [18, 19]. For target data types, we classified the included articles into six categories: image, free text, tabular data, signal data, genomic sequence, and multi-modality. Additionally, Table 2 lists specific applications and performance advantages of AutoML for users focused on specific tasks.

3 | RESULTS

Figure 1 illustrates the literature selection process for this systematic review. Our initial search yielded 2730 papers. We removed 1378 duplicates; hence, we used 1352 records for title and abstract screening. We excluded 1184 records because they were either not relevant to healthcare ($n = 331$) or did not use AutoML methods ($n = 722$); were conference papers that were not from listed conferences ($n = 9$); were not research articles ($n = 121$); or were not in English ($n = 1$). As a result, we included 168 articles for full-text review. Finally, we included 118 papers for systematic review. Figure 2 shows the rising trend of publications in AutoML with interpretation for healthcare and indicates that image and tabular data constituted the major subsets for all included publications. In this section, we first summarize AutoML techniques. Then, we elaborate on the ML interpretations used in the included articles. Finally, we summarize the representative AutoML with interpretation systems for different data modalities.

3.1 | AutoML techniques

For AutoML techniques, we followed the previous classification criteria [9] based on three stages of the ML pipeline: automated data preparation ($n = 18$), automated feature engineering ($n = 95$), and automated model development ($n = 31$). Figure 3 provides a comprehensive overview and Table 3 offers a detailed description of the ML components automated by AutoML within the healthcare sector. Specifically, data preparation refers to the process of collecting and processing raw data into a suitable format for downstream ML stages. AutoML has been leveraged to deal with processes such as automatic data collection [96, 106], noise filtering [27, 28, 44, 119], missing value imputation [87, 95, 110, 126, 133], data imbalance compensation [87, 90, 102, 140], data normalization [44], redundant data removal [53], outlier removal [133], sample clustering [135], data pattern shift detection [137], and continuous variable binning [109]. Feature engineering describes the process of creating new features or modifying existing features to enhance ML performance and AutoML has been used to facilitate automatic feature generation [21, 60, 61, 63, 64, 66, 68, 71, 72, 76, 77, 79–81, 103, 120, 122, 123], selection [70, 72, 80, 98, 99, 101, 102, 105, 108, 121, 124, 127, 135, 141], and transformation [67, 78, 107, 138, 142, 143]. Model development refers to the process of creating, training, and optimizing a model based on either the formatted data or modified features. AutoML has also been used for the selection of main backbone models [65, 86, 88, 89, 91–93, 125, 144], the tuning of model-specific parameters [24, 98, 100, 119, 126, 128, 136], and the optimization of model-specific [21, 24, 40, 59, 74, 86, 89–93, 102, 110, 122, 124, 131, 138, 144] or agnostic hyperparameters [24, 62, 69, 95].

Additionally, we conducted a comparative analysis of commonly used metrics between AutoML and the most competitive baseline in the last column of Table 2, which demonstrated that AutoML outperformed conventional ML solutions across various data types. Specifically, slashes ("/") divide AutoML performance and the most competitive baseline performance. Hyphens ("-") indicate that specific results were not reported in the original papers. Ampersands ("&") separate the same evaluation metrics across different tasks or experimental settings and commas (",") separate different evaluation metrics. We retained all measurement units and decimal digits from the original papers. We did not report results from studies in which visual performance comparisons were made without quantitative data or from studies involving an excessive number of tasks because of content constraints.

TABLE 2 Information summary of the included studies on AutoML with interpretation for healthcare.

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Image	2023	Alkhalaf et al. [21]		√	√	Feature interaction and importance	AAOXAI-CD	CNN, RNN, GRU, LSTM	Cancer classification	Accuracy: 99.00/97.00 & 99.42/98.43
Image	2023	Berghe et al. [22]		√		Feature interaction and importance		U-net, CNN	Structural lesions detection	Accuracy: 0.89/- & 0.92/-, AUROC: 0.92/- & 0.91/-
Image	2023	Cabon et al. [23]		√		Data dimensionality reduction		LR, RF, SVM	Functional age estimation	MAE: 1.4/- & 1.6/- & 1.3/-
Image	2023	Choi et al. [24]			√	Knowledge distillation and rule extraction	SimpleMind	CNN, U-net	Endotracheal tube assessment, kidney segmentation, prostate segmentation	Accuracy: 89/-, DSC: 0.881/0.878 & 0.842/0.818, HD: 46.4/30.7
Image	2023	Custode et al. [25]		√		Intrinsically interpretable models		U-net, CNN, decision tree	Lung status evaluation	
Image	2023	Dai et al. [26]		√		Feature interaction and importance	MS-net	CNN	Lung nodules assessment	Accuracy: 92.4/90.0 & 88.5/87.3
Image	2023	Gerbasi et al. [27]	√	√		Feature interaction and importance	DeepMiCa	CNN, U-net	Microcalcifications detection	Accuracy: 0.83/- & 0.83/-, AUROC: 0.95/- & 0.89/-, AUPRC: 0.78/-, IOU: 0.74/-
Image	2023	Ghassemi et al. [28]	√			Feature interaction and importance		GAN	COVID-19 classification	Accuracy: 89.24/88.05 & 98.25/94.69 & 96.20/94.69 & 98.89/96.30 & 99.2/99.6, AUROC: 97.22/96.71 & 99.79/99.03 & 99.43/99.43 & 99.95/99.60 & 99.95/99.99
Image	2023	Jun et al. [29]		√		Feature interaction and importance		CNN, U-net	Noninvasive meningioma triaging	AUROC: 0.770/0.757, DSC: 0.910/0.907

(Continues)

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Image	2023	Leong et al. [30]		√		Feature interaction and importance		Decision tree	Lung water content evaluation	AUROC: 0.719/- & 0.756/-
Image	2023	Orton et al. [31]		√		Data dimensionality reduction		LASSO	Molecular, histopathology and clinical target prediction	
Image	2023	Pham et al. [32]		√		Feature interaction and importance		U-net, CNN	Human epidermal growth factor receptor-2 classification	F1-score: 0.80/0.81
Image	2023	Saglam et al. [33]		√		Feature interaction and importance		XGBoost, SVM	Early onset schizophrenia classification	Accuracy: 0.80/0.78, AUROC: 0.85/0.83
Image	2023	Taşcı et al. [34]		√		Feature interaction and importance	DGXAINet	CNN, SVM	Brain tumor classification	Accuracy: 98.42/95.75 & 99.96/98.91
Image	2023	Wang et al. [35]		√		Feature interaction and importance		CNN, U-net	Parkinson's disease classification	AUROC: 0.901/0.856
Image	2023	Xiang et al. [36]		√		Feature interaction and importance		CNN, GCN	Prostate cancer classification	Accuracy: 0.677/0.584, AUROC: 0.985/- & 0.986/-
Image	2023	Yoon et al. [37]		√		Feature interaction and importance		CNN	Anterior disc displacement classification	AUROC: 0.985/0.910 & 0.960/0.861
Image	2022	Yu et al. [38]		√		Feature interaction and importance		CNN, RF	Idiopathic pulmonary fibrosis prediction	AUROC: 0.987/-
Image	2022	Basso et al. [39]		√		Data dimensionality reduction		LDA, LR, RF	Glomerular disorder classification	Accuracy: 77/- & 87/-
Image	2022	Chen et al. [40]		√	√	Intrinsically interpretable models		R-CNN, U-net, LR	Blunt splenic injury triaging	Accuracy: 92/-, AUROC: 0.83/0.88

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Image	2022	Falco et al. [41]		√		Intrinsically interpretable models		Fuzzy rules	COVID-19 classification	Accuracy: 80.67/80.28
Image	2022	Kakileti et al. [42]		√		Knowledge distillation and rule extraction		V-net, RF	Early vascularity evaluation	AUROC: 0.85/0.79
Image	2022	Maqsood et al. [43]		√		Feature interaction and importance		CNN, SVM	Brain cancer prediction	Accuracy: 97.47/93.85 & 98.92/98.59
Image	2022	McCay et al. [44]	√	√		Knowledge distillation and rule extraction		LR, SVM, decision tree, LDA	Cerebral palsy prediction	Accuracy: 100/100 & 38/38 & 97.37/86.84
Image	2022	Mou et al. [45]		√		Feature interaction and importance	DeepGrading	CNN	Corneal confocal microscopy estimation	Accuracy: 84.10/82.40
Image	2022	Nafisah et al. [46]		√		Feature interaction and importance		CNN, U-net	Tuberculosis detection	Accuracy: 0.987/0.980, AUROC: 0.999/0.990
Image	2022	Nijiati et al. [47]		√		Feature interaction and importance		CNN, U-net	Active pulmonary tuberculosis classification	Accuracy: 0.910/0.895
Image	2022	Sharma et al. [48]		√		Feature interaction and importance		CNN, U-net	COVID-19 classification	Accuracy: 97.45/98.70, AUROC: 0.998/0.980
Image	2022	Park et al. [49]		√		Feature interaction and importance		U-net, LightGBM	Pilocytic astrocytomas classification	AUROC: 0.930/0.785
Image	2022	Sharma et al. [50]		√		Feature interaction and importance	COVID-MANet	CNN, U-net	COVID-19 classification	Accuracy: 97.37/97.16, IOU: 93.64/91.40, DSC: 96.70/95.49
Image	2022	Suri et al. [51]		√		Feature interaction and importance	COVLIAS 2.0-cXAI	CNN, U-net	COVID-19 localization	Accuracy: 98.5/98.2, AUROC: 0.990/0.988

(Continues)

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Image	2022	Ullah et al. [52]		√		Feature interaction and importance		GNB, SVM, decision tree, LR, KNN, RF	COVID-19 classification	Accuracy: 98.5/99.4
Image	2021	Fu et al. [53]	√	√		Feature interaction and importance		CNN, GRU	Brain disease classification	Accuracy: 0.8961/0.9458
Image	2021	Horry et al. [54]		√		Intrinsically interpretable models		CNN, decision tree	Lung cancer classification	Accuracy: 0.85/-
Image	2021	Myeongkyun et al. [55]		√		Knowledge distillation and rule extraction		R-CNN, K-means, SVM	Bacterial pneumonia classification, COVID-19 classification	Accuracy: 91.2/- & 95.0/-
Image	2021	Pietsch et al. [56]		√		Feature interaction and importance	APPLAUSE	U-net, Gaussian process regression	Placenta health prediction	AUROC 0.95/-
Image	2021	Shorfuzzaman et al. [57]		√		Feature interaction and importance		CNN	Diabetic retinopathy triaging	Accuracy: 0.962/0.986, AUROC: 0.978/0.997
Image	2021	Zhao et al. [58]		√		Feature interaction and importance		LR	COVID-19 classification	Accuracy: 0.9460/0.9249, AUROC: 0.9470/0.9797, DSC: 0.9796/0.9732, HD: 20.2249/41.0517
Image	2021	He et al. [59]			√	Feature interaction and importance	CovidNet3D	CNN	COVID-19 detection	Accuracy: 88.69/88.55 & 82.29/81.82 & 96.88/94.27
Image	2021	Boumaraf et al. [60]		√		Data dimensionality reduction		CNN	Breast cancer classification	Accuracy: 98.13/87.69 & 98.13/87.69 & 98.26/98.18
Image	2021	Cheung et al. [61]		√		Data dimensionality reduction		CNN	Retinal-vessel caliber measurement	
Image	2021	Mosquera et al. [62]			√	Data dimensionality reduction		CNN	Chest radiography diagnosis	AUROC: 0.7491/- & 0.8745/-

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Image	2021	Tamarappoo et al. [63]		√		Feature interaction and importance		XGBoost	Cardiac events prediction	AUROC: 0.81/0.75
Image	2021	Yan et al. [64]		√		Data dimensionality reduction	WBC-Profiler	PSD, RF	Leukocyte classification	Accuracy: 90.22/88.88
Image	2020	Rucco et al. [65]			√	Data dimensionality reduction	TPOT, Auto-SkLearn	CNN	Glioblastoma diagnosis	Accuracy: 0.89/0.89 & 0.92/0.89, AUROC: 0.96/0.84 & 0.91/0.84
Image	2020	Putten et al. [66]		√		Data dimensionality reduction		CNN	Early neoplasia classification	AUROC: 0.93/0.89
Image	2020	Wang et al. [67]		√		Feature interaction and importance		CNN, RNN	Congenital heart disease interpretation	Accuracy: 0.946/0.895 & 0.917/0.878, AUROC: 0.918/0.845
Image	2020	Yin et al. [68]		√		Data dimensionality reduction		PNN, SVM, LR, Adaboot, RF, MLP	Bladder cancer diagnosis	Accuracy: 96.7/84.0, AUROC: 0.990/0.926
Image	2020	Lecouat et al. [69]			√	Intrinsically interpretable models		Adaptive smoothing, game encoding	fMRI sensing	PSNR: 36.47/37.95 & 44.17/44.09
Image	2019	Wu et al. [70]			√	Intrinsically interpretable models		Decision tree	Breast cancer prediction	Accuracy: 72.43/-
Image	2019	Yamamoto et al. [71]		√		Data dimensionality reduction		Autoencoder, LASSO, ridge regression, SVM	Prostate cancer recurrence prediction	AUROC: 0.884/0.721
Image	2018	Pereira et al. [72]		√		Feature interaction and importance		Boltzmann machine, RF	Brain tumor segmentation, penumbra estimation	DSC: 0.84/0.87 & 0.75/0.82
Image	2017	Song et al. [73]		√		Data dimensionality reduction	Remurs	LASSO, elastic net	fMRI analysis	Accuracy: 78.15/75.46

(Continues)

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Free text	2021	Diao et al. [74]		√	√	Feature interaction and importance		LightGBM	ICD coding	Accuracy: 95.2/91.3
Free text	2021	Kulshrestha et al. [75]		√		Feature interaction and importance		Elastic net, XGBoost, CNN	Chest injury prediction	AUROC: 0.93/-
Free text	2021	Blanco et al. [76]		√		Feature interaction and importance		GRU	Death cause extraction	AUROC: 53.3/52.1 & 49.4/58.8 & 58.2/62.0
Free text	2021	Dong et al. [77]		√		Feature interaction and importance	HLAN	GRU	Medical coding	AUROC: 88.4/88.3 & 94.5/96.9 & 88.5/90.2
Free text	2020	Yang et al. [78]		√		Feature interaction and importance	AMFF	LSTM	Medical entity tagging	F1-score: 94.48/90.23 & 92.11/88.46 & 68.34/64.61 & 80.51/80.03
Free text	2020	Li et al. [79]		√		Feature interaction and importance	MultiResCNN	CNN	ICD coding	F1-score: 0.073/0.068 & 0.608/0.584
Free text	2019	Atutxa et al. [80]		√		Feature interaction and importance		RNN, transformer	ICD coding	F1-score: 0.838/0.786 & 0.963/0.935 & 0.952/0.895
Free text	2018	Duarte et al. [81]		√		Feature interaction and importance		GRU	ICD coding	Accuracy: 89.320/79.802 & 81.349/70.754 & 76.112/67.404
Tabular data	2023	Li et al. [82]		√		Feature interaction and importance	FETCH	MLP	Hepatitis classification	F1-score: 0.9290/0.8839
Tabular data	2023	Junaid et al. [83]		√		Feature interaction and importance		SVM, RF, LightGBM	Parkinson's disease prediction	
Tabular data	2023	Islam et al. [84]		√		Data dimensionality reduction		LR, MLP, RF, XGBoost	Hypertension prediction	AUROC: 0.894/0.829

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Tabular data	2023	Wang et al. [85]		√		Knowledge distillation and rule extraction		LR	Heart failure prediction	Accuracy: 0.999/0.995, AUROC: 0.981/0.979
Tabular data	2023	Zhang et al. [86]			√	Feature interaction and importance		LR, RF, GBM, MLP	Severe acute pancreatitis prediction	Accuracy: 0.910/0.920, AUROC: 0.907/0.849
Tabular data	2022	Agüero et al. [87]	√	√		Knowledge distillation and rule extraction		MLP, GRU, LSTM	Antimicrobial multidrug resistance prediction	Accuracy: 65.40/-, AUROC: 66.73/-
Tabular data	2022	Chou et al. [88]		√	√	Feature interaction and importance		XGBoost, RF, LR	Spinal cord injury prediction	AUROC: 0.68/-
Tabular data	2022	Cui et al. [89]			√	Feature interaction and importance		LR, RF, XGBoost, MLP, GBM	Early death prediction	Accuracy: 0.772/-, AUROC: 0.820/-
Tabular data	2022	Danilatou et al. [90]	√	√	√	Feature interaction and importance		LR, RF, SVM, decision tree	Mortality prediction	AUROC: 0.93/0.85 & 0.87/0.79
Tabular data	2022	Thongprayoon et al. [91]			√	Feature interaction and importance		RF, decision tree, XGBoost, MLP	Acute kidney injury prediction	Accuracy: 0.72/0.74, AUROC: 0.79/0.78
Tabular data	2022	Yin et al. [92]			√	Feature interaction and importance		RF, GBM, MLP, LR, XGBoost	Severe acute pancreatitis prediction	Accuracy: 0.953/0.943, AUROC: 0.945/0.898
Tabular data	2022	Yu et al. [93]			√	Feature interaction and importance		XGBoost, LR, GBM, RF, MLP	Mortality prediction	Accuracy: 0.879/0.857, AUROC: 0.888/0.782
Tabular data	2022	Zhang et al. [94]		√		Data dimensionality reduction		XGBoost, CNN	Ischemic stroke classification	Accuracy: 0.6020/0.5671, AUROC: 0.6757/0.6532
Tabular data	2021	Alaa et al. [95]	√	√	√	Knowledge distillation and rule extraction	AutoPrognosis	RF, AdaBoost, MLP	Breast cancer prediction	AUROC: 0.771/0.773 & 0.823/0.792 & 0.777/0.763 & 0.815/0.784 & 0.790/0.778 & 0.803/0.775

(Continues)

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Tabular data	2021	Chiang et al. [96]	√	√		Data dimensionality reduction		RF	Personalized lifestyle recommendations	MAE: 5.34/5.94 & 3.80/4.05, RMSE: 8.24/9.98 & 6.05/6.68
Tabular data	2021	Laria et al. [97]		√		Data dimensionality reduction		Deep LASSO	Attention-deficit hyperactivity disorder prediction	RMSE: 0.545/0.561 & 0.494/0.588
Tabular data	2021	Ikemura et al. [98]		√	√	Feature interaction and importance		GBM, XGBoost	Mortality prediction	AUPRC: 0.807/0.736
Tabular data	2021	Luo et al. [99]		√		Knowledge distillation and rule extraction		XGBoost	Asthma hospital visit prediction	
Tabular data	2020	Tong et al. [100]			√	Knowledge distillation and rule extraction		XGBoost	Asthma hospital visit prediction	
Tabular data	2020	Xie et al. [101]		√		Intrinsically interpretable models	AutoScore	RF	Mortality prediction	AUROC: 0.780/0.778
Tabular data	2020	Yang at al. [102]	√	√	√	Data dimensionality reduction	mAML	GBM, XGBoost, AdaBoost	Disease classification	
Tabular data	2019	Senderovich et al. [103]		√		Feature interaction and importance		Congestion graphs, generalized Jackson networks	Emergency department and outpatient cancer clinic time prediction	RMSE: 36/- & 104/-
Tabular data	2018	Banerjee et al. [104]			√	Knowledge distillation and rule extraction		RF, LASSO, SVM	Breast cancer prediction	MSE: 0.04/-
Tabular data	2018	Billiet et al. [105]		√		Intrinsically interpretable models	Interval coded scoring	Elastic net, linear programming	Acute inflammations diagnosis, breast cancer diagnosis, etc.	Accuracy: 0.88/-, AUROC: 0.92383/-

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Tabular data	2018	Corey et al. [106]	√			Data dimensionality reduction	Pythia	RF, LASSO, GBM	Mortality prediction	AUROC: 0.836/- & 0.883/- & 0.916/- & 0.828/- & 0.820/- & 0.781/- & 0.908/- & 0.845/- & 0.890/- & 0.875/- & 0.910/- & 0.850/- & 0.924/- & 0.890/-
Tabular data	2018	Khurana et al. [107]		√		Intrinsically interpretable models		Transformation graph	Diabetes prediction, oncology prediction	F1-score 0.820/0.615 & 0.895/0.832
Tabular data	2017	Laet et al. [108]		√		Data dimensionality reduction		Naïve bayes, LR	Cerebral palsy classification	Accuracy: 91/90
Tabular data	2016	Drakakis et al. [109]	√			Intrinsically interpretable models		Decision tree	Histamine H1 receptor binding prediction	Accuracy: 73.21/68.91 & 86.35/84.50 & 70.43/65.12 & 83.60/88.12
Tabular data	2007	Keles et al. [110]	√		√	Intrinsically interpretable models	NEFCLASS	Neuro-fuzzy system	Prostate cancer classification	
Signal	2023	Donckt et al. [111]		√		Intrinsically interpretable models		LR, GBM	Sleep triaging	Accuracy: 0.866/0.864 & 0.831/0.849 & 0.836/0.815 & 0.867/0.875
Signal	2023	Heitmann et al. [112]		√		Feature interaction and importance	DeepBreath	CNN, LR	Respiratory pathology detection	AUROC: 0.887/0.703 & 0.739/0.315 & 0.743/0.614 & 0.870/0.896
Signal	2023	Raeisi et al. [113]		√		Feature interaction and importance		CNN, GAT	Neonatal seizure detection	AUROC: 0.966/0.957
Signal	2022	Han et al. [114]		√		Knowledge distillation and rule extraction		CNN, knowledge graph	Myocardial infarction prediction	Accuracy: 98.88/97.27 & 93.65/93.77 & 94.13/91.54
Signal	2022	Huang et al. [115]		√		Feature interaction and importance		Autoencoder	Epilepsy detection	Accuracy: 97.3/72
Signal	2022	Jahmunah et al. [116]		√		Feature interaction and importance		CNN	Myocardial infarction detection	Accuracy: 98.9/- & 98.5/-

(Continues)

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Signal	2022	Yang et al. [117]		√		Knowledge distillation and rule extraction		KNN, SVM, RF, MLP	Cardiac abnormalities classification	Accuracy: 99.0/98.7
Signal	2021	Lee et al. [118]		√		Feature interaction and importance		CNN	Arrhythmia classification	F1-score: 81.75/82.2
Signal	2021	Fuchs et al. [119]	√		√	Intrinsically interpretable models		Fuzzy rules	Tremor severity assessments	MAE: 1.85/6.41 & 2.30/8.65
Signal	2021	Kim et al. [120]		√		Data dimensionality reduction		CNN	BMI channel selection	Accuracy: 76.8/79.6 & 58.3/58.5 & 70.8/71.4
Signal	2019	Saboo et al. [121]		√		Data dimensionality reduction		GMM	Active electrodes selection	AUROC: 0.974/0.752
Signal	2019	Tison et al. [122]		√	√	Feature interaction and importance		CNN	Cardiac disease detection	AUROC: 0.94/- & 0.91/- & 0.86/- & 0.77/-
Genomic sequence	2021	Clauwaert et al. [123]		√		Feature interaction and importance		Transformer	Genome annotation	AUROC: 0.740/0.882 & 0.920/0.961 & 0.976/0.958 & 0.981/0.978 & 0.976/0.964, AUPRC: 0.039/0.035 & 0.057/0.132 & 0.141/0.098 & 0.128/0.128 & 0.141/0.137
Genomic sequence	2020	Le et al. [124]		√	√	Data dimensionality reduction	TPOT-FSS	TPOT, XGBoost	TPOT enhancement	
Genomic sequence	2019	Trabelsi et al. [125]			√	Feature interaction and importance	deepRAM	CNN, RNN	DNA/RNA sequence binding specificities prediction	AUROC: 0.930/- & 0.951/-
Genomic sequence	2018	Nagorski et al. [126]	√		√	Data dimensionality reduction	SpaCC	Convex optimization	Cancer epiGenomictics subtype discovery	

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Genomic sequence	2018	Shen et al. [127]		√		Data dimensionality reduction	OFSSVM	SVM	Cancer prediction	Accuracy: 82.35/79.41 & 88.24/88.24 & 97.06/97.06
Genomic sequence	2007	Yap et al. [128]			√	Knowledge distillation and rule extraction		Bayesian network	Ovarian cancer detection	
Multi-modality	2023	Roest et al. [129]		√		Feature interaction and importance		CNN, U-net, SVM	Prostate cancer detection	AUROC: 0.81/0.69
Multi-modality	2023	Wouters et al. [130]		√		Feature interaction and importance		VAE, LR, cox regression	Cardiac resynchronization therapy outcome prediction	C-statistic: 0.72/0.70 & 0.70/0.72
Multi-modality	2022	Abbas et al. [131]		√	√	Feature interaction and importance		XGBoost, CNN, U-net	Visual acuity prediction	AUROC: 0.849/0.847
Multi-modality	2022	Gerbasi et al. [132]	√	√		Data dimensionality reduction		XGBoost	Stroke prediction	Accuracy: 0.79/-, AUROC 0.85/-
Multi-modality	2022	Gutierrez et al. [133]	√	√		Data dimensionality reduction	GA-MADRID	SVM, KNN, decision tree	Alzheimer's disease classification, frontotemporal dementia classification	Accuracy: 0.849/- & 0.872/- & 0.885/- & 0.926/-
Multi-modality	2022	Zhang et al. [134]		√		Feature interaction and importance		Transformer, RF, LSTM	In-hospital mortality prediction, physiological decompensation prediction, length of stay prediction	AUROC: 0.845/0.841 & 0.845/0.826, AUPRC: 0.464/0.453 & 0.180/0.125
Multi-modality	2021	Ferté et al. [135]	√	√		Intrinsically interpretable models	PheVis	LR	Medical condition prediction	AUROC: 0.957/0.994 & 0.987/0.910, AUPRC: 0.798/0.975 & 0.299/0.262
Multi-modality	2019	Li et al. [136]			√	Feature interaction and importance	KERP	Graph transformer	Medical image report generation	AUROC: 0.686/0.612 & 0.726/0.646 & 0.760/0.689 & 0.862/0.800

(Continues)

TABLE 2 (Continued)

Data type	Year	Paper	Automated data preparation	Automated feature engineering	Automated model development	Interpretability methods	Model name	Main ML architectures	Healthcare applications	Performance comparison
Multi-modality	2018	Chen et al. [137]	√			Data dimensionality reduction	DASSA	IB, MDL	Disease propagation pattern detection	
Multi-modality	2017	Guo et al. [138]		√	√	Knowledge distillation and rule extraction		Hierarchical probabilistic framework	Dermatology image analysis	Accuracy: 75.3/62.9, AUROC: 0.78/0.67

Abbreviations: AUPRC, area under the precision recall curve; AUROC, area under the receiver operating characteristic curve; C-statistic, concordance statistic; DSC, dice similarity coefficient; HD, hausdorff distance; IOU, intersection over union; MAE, mean absolute error; MSE, mean squared error; PSNR, peak signal-to-noise ratio; RMSE, root mean squared error.

Furthermore, we implemented a toy example to compare AutoML solutions with classic ML models based on 44 918 de-identified patients from BIDMC critical care units [145]. The prediction target was in-hospital mortality (8.81% across all patients) and the candidate variables were age, temperature, platelet, glucose, sodium, lactate, potassium, bicarbonate, heart rate, respiration rate, hematocrit, creatinine, hemoglobin, chloride, anion gap, white blood cells, blood urea nitrogen, systolic blood pressure, diastolic blood pressure, mean arterial pressure, and peripheral capillary oxygen saturation [146]. We randomly divided the entire dataset using the ratio 6:2:2 for model training, validation, and testing. For traditional ML models, we optimized the hyperparameters using grid search based on the area under the receiver operating characteristic curve (AUROC) evaluated on the validation set. For AutoML solutions, we automatically determined the hyperparameters using their inherent algorithms; therefore, their training data included both the training and validation sets. Figure 4 presents the AUROC results on the unseen test set, which demonstrates that the two AutoML solutions of AutoGluon [147] and TPOT [124] statistically significantly outperformed the conventional ML models random forest [148], gradient boosting machine (GBM) [149], and *K*-nearest neighbor [150]. We made the code open access to enable reproducibility and serve as an exemplary case study [151].

3.2 | Interpretation methods

Regarding the ML interpretations, we grouped them into four categories based on the commonly adopted criteria [18, 19]: feature interaction and importance ($n = 63$), data dimensionality reduction ($n = 27$), intrinsically interpretable models ($n = 14$), and knowledge distillation and rule extraction ($n = 14$).

Feature interaction entails quantifying the effect of one feature on another, considering their mutual influence, whereas feature importance involves discerning the significance of input features in shaping the output targets of ML models [61, 63, 67, 72, 76, 78–81, 98, 103, 108, 122, 125, 136]. In the healthcare domain, the alignment of feature interaction and importance with clinical expertise enhances healthcare professionals' trust in ML outputs [152]. However, when feature interaction and importance diverge from established knowledge, ML models may encounter overfitting issues. Remarkably, such disparities occasionally reveal previously unidentified biomarkers [153].

Data dimension reduction refers to the use of a subset of the most informative raw inputs or modified

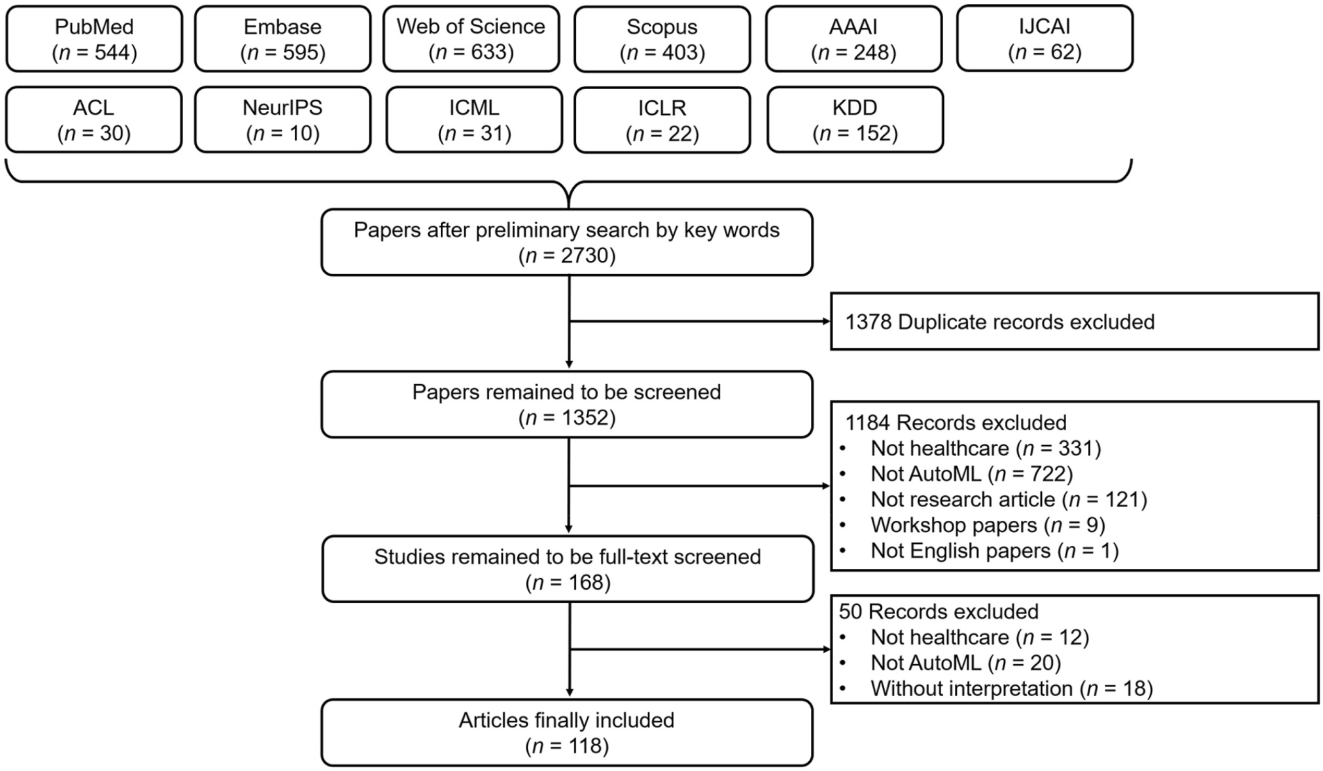

FIGURE 1 Literature selection flow for automated machine learning with interpretation in healthcare.

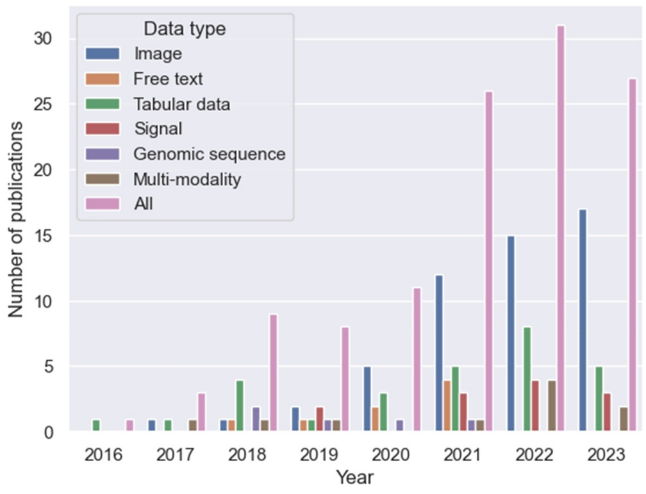

FIGURE 2 Timeline of publications on automated machine learning with interpretation for healthcare since 2016. Our search concluded on September 1, 2023, which accounts for the lower number of included publications published in 2023 compared with those published in 2022.

features in model development and subsequent analyses [61, 62, 64–66, 68, 71, 73, 76, 102, 106, 108, 120, 121, 124, 126, 127, 137]. In the context of high-dimensional samples, data dimension reduction helps the model to focus on salient features, thereby simplifying model complexity and enhancing its interpretability [154]. Additionally, data dimension reduction enables the effective graphical visualization of data distributions within a low-dimensional space [155]. This visualization reveals latent data patterns that can be integrated into subsequent model development, thereby enhancing both model performance and interpretability [18, 19].

Intrinsically interpretable models represent the application of transparent models to solve prediction problems [18] such as logistic regression [101, 111, 135, 140, 141, 156, 157], decision tree [25, 54, 70, 105, 137], fuzzy rules [41, 110, 119], and mathematical solid decision functions [40, 69, 107]. Intrinsically interpretable models feature simple architectures or algorithms, thereby fostering a clear understanding of the relationship between inputs and outputs [18, 19]. These models may not consistently achieve predictive performance comparable with that of their black box counterparts, but within high-stakes tasks that impact lives, model transparency is substantially more important than marginal performance superiority [158].

Knowledge distillation and rule extraction refer to the processes of simplifying intricate ML models into either streamlined models or human-comprehensible rules, respectively [99, 100, 104, 128, 138]. Knowledge distillation is a technique designed to train simple student models by mirroring the behavior of complex teacher models while preserving model performance [159]. Post distillation, student models demonstrate reduced

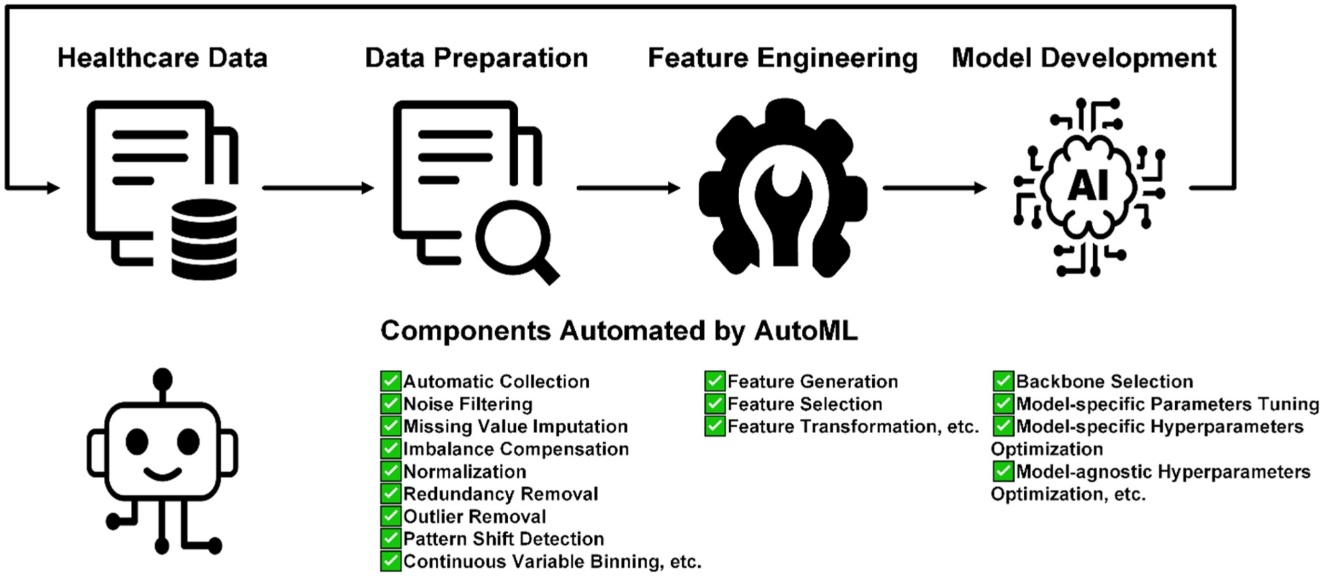

FIGURE 3 Overview of the ML components automated by automated ML within the healthcare sector. This figure is reproduced from [139] with permission. ML, machine learning.

TABLE 3 Description of ML components automated by AutoML in the healthcare sector.

Stages	Operations	Description
Automated data preparation	Automatic data collection	Collecting raw data in an automated manner.
	Noise filtering	Removing inherent noise from the data.
	Missing value imputation	Filling in missing values in the dataset.
	Data imbalance compensation	Addressing and compensating for imbalanced classes in the data.
	Data normalization	Scaling data to a standard range.
	Redundant data removal	Eliminating duplicate or unnecessary data entries.
	Outlier removal	Identifying and removing anomalous data points.
	Samples clustering	Grouping similar data samples together.
	Data pattern shift detection	Detecting changes in data patterns over time.
	Continuous variable binning	Converting continuous variables into discrete bins.
Automated feature engineering	Automatic feature generation	Creating features automatically using algorithms.
	Feature selection	Choosing the most relevant features for modeling.
	Feature transformation	Transforming features to a more suitable form for modeling.
Automated model development	Backbone model selection	Choosing the main model architecture.
	Model tuning	Adjusting model-specific parameters for better performance.
	Hyperparameter optimization	Finding the best hyperparameters for better performance.

complexity, which renders them more comprehensible to humans and potentially bolsters transferability [160]. Rule extraction yields human-understandable rules because each rule inherently provides a logical explanation for its decision [161]. Based on these interpretation methods discussed above, healthcare practitioners can discern potential errors and ascertain the reliability of ML models [162].

3.3 | Data modalities

In this section, we discuss AutoML with interpretation for different types of healthcare data: image ($n = 53$), free text ($n = 8$), tabular data ($n = 29$), signal ($n = 12$), genomic sequence ($n = 6$), and multi-modality ($n = 10$). Figures 5 and 6 present the summary statistics of AutoML and interpretation techniques in the included

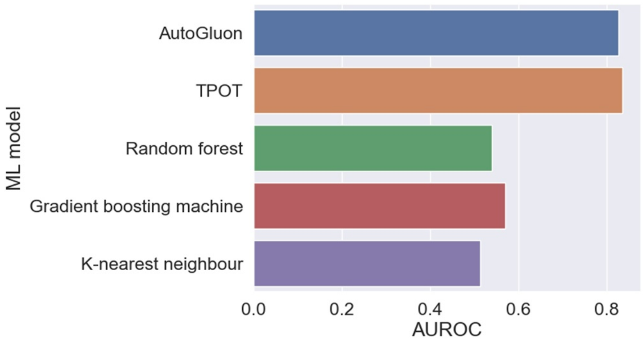

FIGURE 4 AUROC comparison of automated ML solutions versus conventional ML methods for real-world in-hospital mortality prediction. AUROC, area under the receiver operating characteristic curve; ML, machine learning.

publications. For AutoML techniques, automated feature engineering dominated in five out of the six modalities; for the genomic sequence, automated model development was more prevalent. Regarding interpretation methods, feature interaction and importance were widely used in all modalities except the genomic sequence, where data dimensionality reduction was the preferred approach. For each data modality, we focused on the principal tasks addressed by AutoML with interpretation systems and elaborated on them using representative studies.

3.3.1 | Image

Medical images are essential diagnostic tools for a spectrum of diseases [66, 163]. AutoML with interpretation enables clinicians with little coding experience [164] to perform a spectrum of healthcare tasks, such as retinal-vessel caliber measurement [61], breast cancer classification [60], and thoracopathy lesion localization [165]. Based on whether they transform raw pixels into useable features, current systems can be classified into two categories: (1) two-step systems that consist of feature extraction and subsequent modeling [64, 66, 68, 70, 71, 166]; and (2) end-to-end systems without the explicit extraction of intermediate features [67, 69].

Two-step systems first extract image features from raw pixels and build up the subsequent analysis based on the extracted features. Various methods have been proposed to automate the extraction of image features, including both commercial software and homemade models. Yin et al. [68] applied the commercial software CellProfiler [167] and ImageJ [168] to extract individual and textual features, and then integrated domain knowledge from pathologists to shortlist useful features. PDE [66] has also demonstrated its effectiveness in automatic feature extraction. By contrast, Yan et al. [64] developed a feature extraction tool and demonstrated the effectiveness of their homemade model through a comparison with human clinicians. With diverse off-the-shelf solutions, multiple tools have been combined to improve the robustness of extracted features [169]. In these systems, the most common interpretation is feature interaction and importance that results from mapping the extracted features back to the original images and highlighting relevant pixels or patches [64, 71]. Additionally, in some systems, inherently interpretable models are applied based on the extracted features to improve model interpretations [70, 166]. For instance, Wu et al. [70] implemented a decision tree to mimic how radiologists interpret the extracted features. Moreover, knowledge distilled from an inherently interpretable model, such as a decision tree, can serve as diagnostic guidelines in the future [166].

Different from two-step methods, end-to-end systems process image inputs without the implicit extraction of intermediate features and output predictions of interest in addition to useful interpretations [170]. In the task of compressed sensing for functional magnetic resonance imaging (fMRI), Lecouat et al. [69] automated the architecture design and parameter training of artificial neural networks (ANN) based on convex optimization and non-cooperative games [171]. To enhance interpretability, they introduced a decision function with sparse parameters and clear mathematical formulas. Wang et al. [67] developed a classic end-to-end system for congenital heart disease classification, including automatic data clustering and model parameter tuning. Similar to two-step systems, their system highlighted important areas on the input image toward ML predictions and used these sub-areas as an interpretation. Although end-to-end systems provide more ceaseless automation and are thus more user-friendly, users should choose the appropriate systems based on whether they need the intermediate features for further modeling and interpretation [68].

3.3.2 | Free text

Medical text records various patients' information, such as hospitalization descriptions, diagnoses, and treatments [172]. Accurate mining of such information can summarize patients' former health conditions and guide subsequent interventions [173]. A fundamental task addressed by AutoML with interpretation is the coding of unstructured raw clinical notes into structured medical codes, such as the international classification of diseases (ICD). Similar to the two-step systems adopted in medical image analysis, this process extracts standard intermediate features from text records, and these intermediate

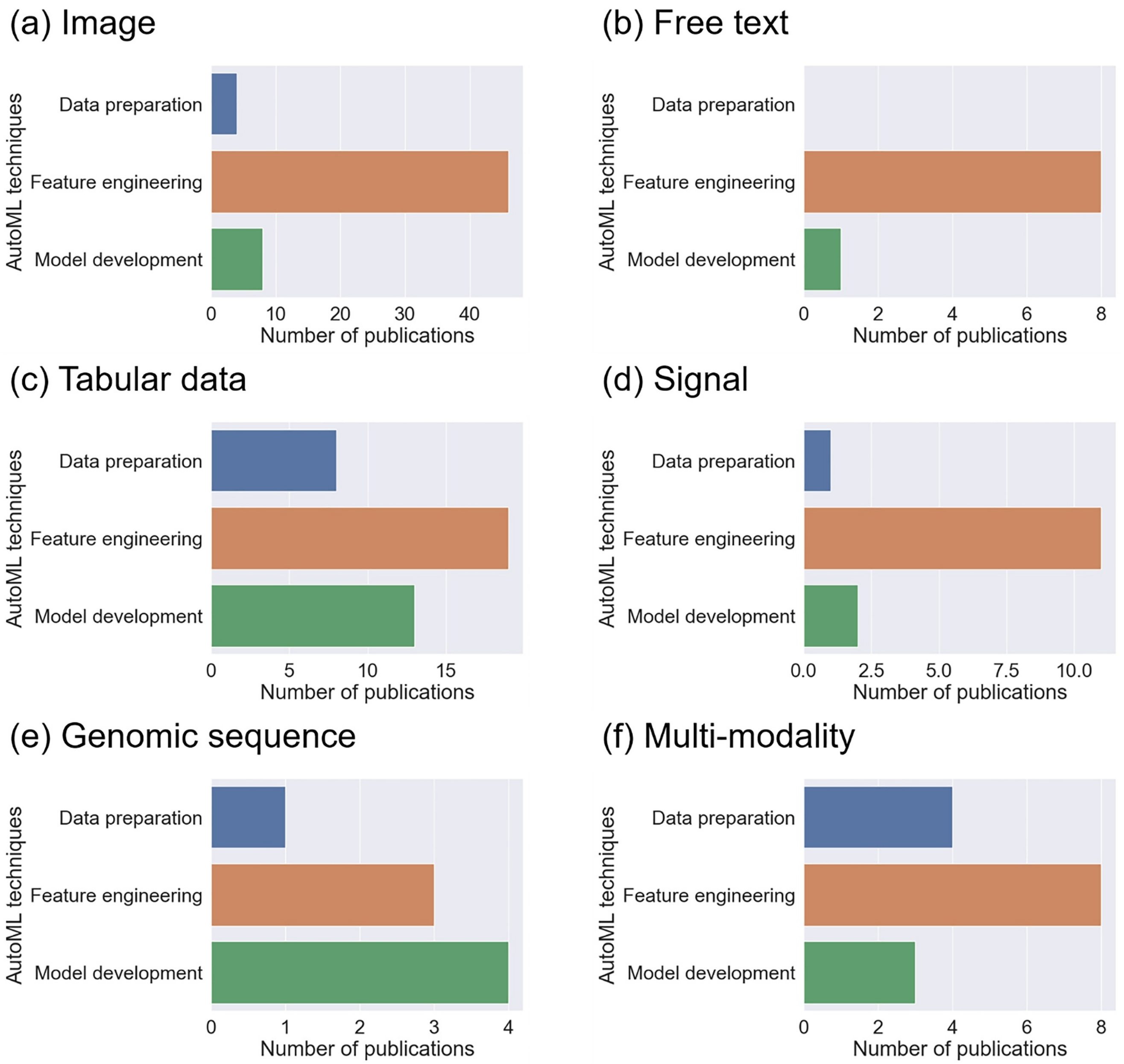

FIGURE 5 Summary statistics of automated machine learning techniques for the included publications targeting each data modality: (a) image; (b) free text; (c) tabular data; (d) signal; (e) genomic sequence; and (f) multi-modality.

features facilitate various subsequent analyses [80]. Conventionally, such a transformation was conducted manually [79, 80], but has been gradually replaced by either commercial software or home-made models to save time and eliminate errors. For example, commercial software called clinical text analysis and knowledge extraction system has been demonstrated to be an effective method for mapping trauma encounter text to structured medical concepts [75]. Additionally, researchers have demonstrated that homemade models are useful for generating informative feature vectors from free text and subsequently projecting these vectors to medical codes [74, 76, 77, 80, 81]. Duarte et al. proposed a framework similar to residual learning, wherein word embeddings are processed using a gated recurrent unit (GRU) to generate representations [81]. These representations are then concatenated with the initial embeddings to prevent information loss and enhance model accuracy. Additionally, Atutxa et al. demonstrated that beyond classic recurrent neural networks (RNN) such as GRU, convolutional neural networks (CNN) and transformers are also effective for mapping diagnostic text to ICD codes [80].

Across all analytical tasks that use medical text, the attention mechanism is the most important backbone. It is valued not only because of its superior performance in

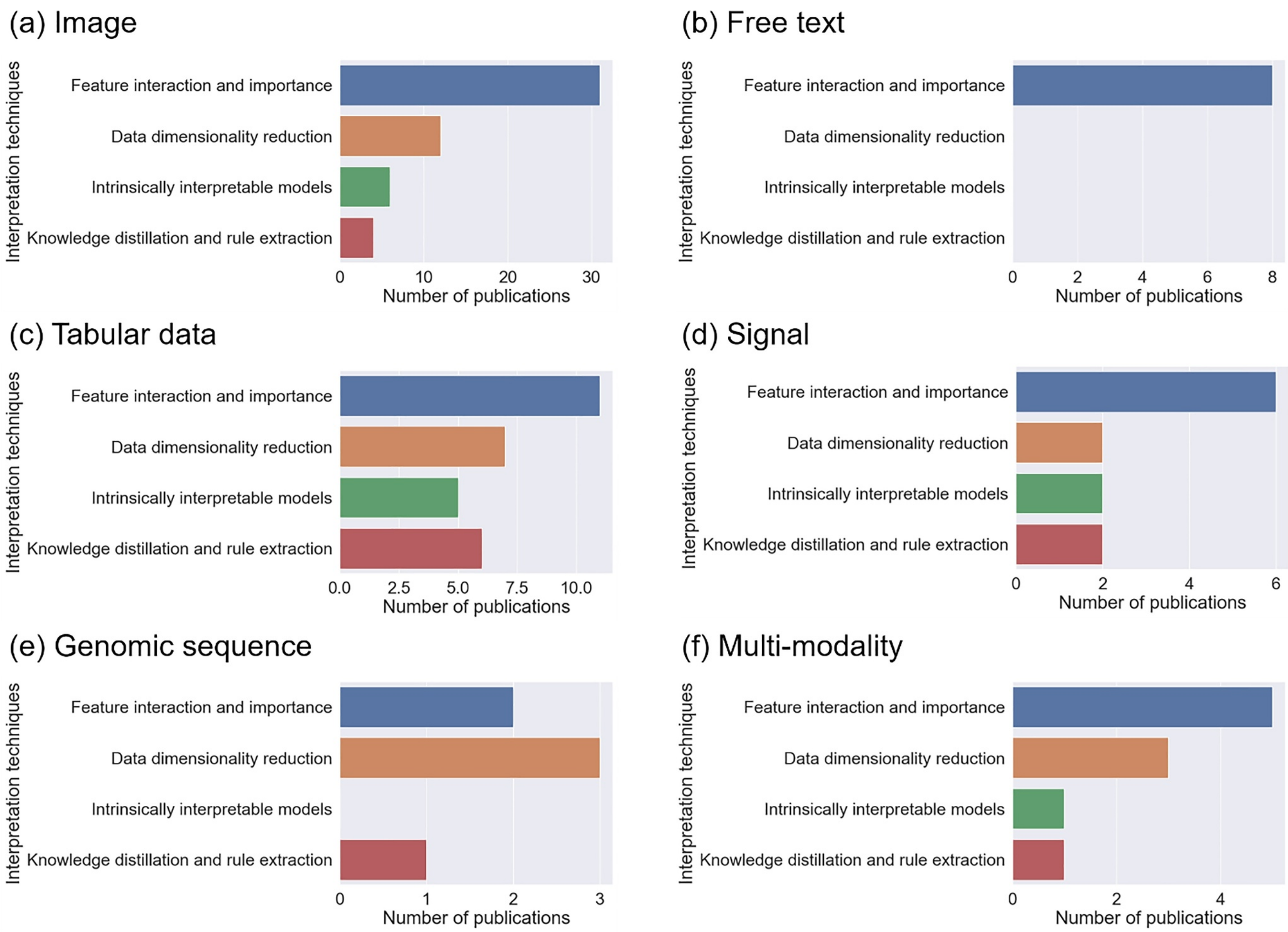

FIGURE 6 Summary statistics of interpretation techniques for the included publications targeting each data modality: (a) image; (b) free text; (c) tabular data; (d) signal; (e) genomic sequence; and (f) multi-modality.

the attention-based transformer [174] but also because of its inherent weights that provide feature interaction and the importance of each part in the input text [76, 78, 81]. For instance, in the sentence "He should undergo chemotherapy when he is diagnosed before his cancer cells metastasize," attention detects that "his cancer cells metastasize" is a crucial component in the automatic determination of the patient's cause of death [76]. In recent studies, researchers introduced the hierarchical attention mechanism, which uses the various types of attention and interprets feature representations on different levels. The hierarchical label-wise attention network [77] applies two-level attention mechanisms at the word-level and sentence-level for selecting important words and sentences in each paragraph, respectively.

3.3.3 | Tabular data

Tabular data, the most common data format in healthcare, includes structured demographic data, vital signs, lab tests, diagnoses, treatments, and procedures [1]. Unlike pixels in images and words in free text, raw features, such as gender in tabular data, typically have clinically explainable meanings, therefore feature engineering becomes the focal point of automation and interpretation for AutoML systems. It should be noted that the proposed methods for tabular data in the included studies can also be applied to structured information derived from unstructured healthcare data, as illustrated in the two-step methods above. In this section, we focus on studies in which researchers explored raw inputs in a structured tabular format.

Traditional feature engineering for tabular data is labor-consuming and costly. It requires ML engineers' intuition and domain knowledge [107]. By contrast, automatic and interpretable feature engineering automatically performs transformation and aggregation across candidate features in a transparent manner. For example, Khurana et al. [107] proposed automatic feature selection and transformation based on intrinsically interpretable transformation graphs, and found that the modified features reduced ML errors. Their work demonstrated the utility of intrinsically interpretable models in feature engineering. AutoScore [141, 156] further exploits the full potential of the intrinsically interpretable clinical score as the backbone for predicting parameters such as the in-hospital mortality rate [141, 156], survival time [157], and rare event occurrence [140]. Although complicated ML models have dominated the analysis of high-dimensional data, for tabular data with a limited number of features, transparent features and intrinsically interpretable models are still preferred in practice [175, 176].

In addition to feature engineering, data preparation (pre-processing) [106] and model development [98] have been automated using AutoML with interpretation systems. Ikemura et al. [98] automated the entire ML lifecycle using commercial software [177] and interpreted models through feature interaction and importance generated by Shapley additive explanations (SHAP) [178]. In addition to commercial software such as H2O.ai, researchers have also developed comprehensive homemade systems for mining clinical tabular data. mAML [102] is an example that includes automated imbalance compensation [179], feature selection [180], and hyperparameter optimization [181]. Specifically, imbalance compensation is addressed using RandomOverSampler [179], SMOTE [182], and ADASYN [183]. Feature selection methods include the distal DBA method [184], HFE [180], and mRMR [185]. Hyperparameter optimization is performed using a grid search [181].

3.3.4 | Signal data

Signal data refers to electrical or mechanical signals collected from physiological sensors to monitor the functioning of the human body and make informed intervention decisions [186]. ML has been applied to identify the sophisticated relationships between various signal inputs and clinical events. AutoML with interpretation further automates and improves the reliability of this analytical process. A promising research direction involves transforming signal data into two-dimensional representations and subsequently applying image-related methods [118]. However, in this section, we focus on these techniques specifically designed for signal data to avoid confusion. Specifically, Fuchs et al. [119] used an intrinsically interpretable fuzzy model to analyze tremor signals, in which the wrapper approach [187] and pyFUME [188] automate feature selection and model development, respectively. Kim et al. [120] proposed an automated channel selection method based on CNN for analyzing electroencephalograms. They further elucidated neurophysiological feature interaction and importance by correlating the selected channels with specific brain regions. In addition to the end-to-end architecture, Tison et al. [122] devised a two-step framework for predicting distinct heart diseases. Initially, the system autonomously generated features using a CNN-hidden Markov model from electrocardiograms (ECG). Subsequently, these features were input into a GBM for predicting the target diseases. Finally, the system calculated the interaction and importance of segments within ECG as the model interpretation. A similar strategy was implemented by Jahmunah et al. [116] in which ECG beats were first extracted using an off-the-shelf algorithm and then input into the downstream DenseNet [189] for myocardial infarction detection. Han et al. [114] conducted an extensive investigation into the use of AutoML for diagnosing myocardial infarction. On top of clinical standards, diagnostic guidelines, and DenseNet-based signal morphology, they developed an interpretable diagnostic system based on production rules.

3.3.5 | Genomic sequence

Genomic sequence data [190, 191] indicate the precise order and arrangement of fundamental genetic elements, such as nucleotides (adenine, thymine, cytosine, and guanine), within DNA sequences (DNA-seq). In addition to DNA-seq, other common genomic sequences include RNA sequences (RNA-seq), Deoxyribonuclease I hypersensitive site sequences (DNase-seq), micrococcal nuclease digestion with deep sequencing (MNase-seq), and chromatin immunoprecipitation sequences (ChIP-seq). These sequences encapsulate the detailed composition of genetic material, thereby offering fundamental information that is essential for comprehending potential associations between genetic patterns and diseases [192]. The principal application of AutoML with interpretation in genomic sequence data mining is to identify genomic sites of interest from the entire genomic sequence. Trabelsi et al. [125] proposed deepRAM for identifying protein binding sites in DNA and RNA-seq based on a hybrid architecture of CNN and RNN. The hyperparameters were automatically tuned through a combination of random search and cross-validation. Sequence motifs, which represent patterns with biological significance, were extracted from the initial CNN layer to improve interpretability [125]. In addition to genomic sites, AutoML has been applied to the data mining of gene expression data. Shen et al. [127] introduced elastic net-based [193] automatic feature selection to a support vector machine (SVM), which demonstrated that feature selection boosted both model performance and interpretability. In addition to classic ML models such as SVM, the transformer has gradually gained popularity in genomic sequence analyses, such as automatic prokaryotic genome annotation [123]. In addition to inherent attention in the transformer for acquiring feature interaction and importance, data dimensionality reduction [124] and rule extraction [128] are used to improve model interpretability.

3.3.6 | Multi-modality

Multi-modality refers to the simultaneous use of more than one data type discussed above to gain a comprehensive understanding of a patient's condition [194]. The integration of these complementary modalities enhances the overall diagnostic accuracy of ML models [195]. AutoML with interpretation is highly valued for processing complex data that involve multiple modalities [196]. PheVis [135] uses a dictionary-based named entity recognition tool to extract medical concepts from free text and then fuses these features with diagnosis codes to predict rheumatoid arthritis and tuberculosis. The SAFE algorithm [197] is used for automatic feature selection, and logistic regression is used for the transparent modeling of the relationship between shortlisted features and medical conditions of interest. Similarly, Zhang et al. [134] combined phenotypical features from free text and clinical features from tabular data to predict in-hospital mortality, physiological decompensation, and length of stay in intensive care units. Compared with features from a single modality of either free text or tabular data, multimodal features have led to statistically significant improvements in performance across most evaluated settings. For analogous frameworks within the field of image modality and signal modality, readers can refer to Abbas et al. [131] and Wouters et al. [130], respectively. They used different tools to extract features from image or signal data and combined them with tabular features, which achieved state-of-the-art performance. In addition to integrating different data modalities for predicting events of clinical interest, the aligned data of different modalities facilitates the translation of high-dimensional data into human-understandable formats, such as human language. KERP [136] was proposed to automatically generate free text reports for medical images, where feature interaction and importance, derived from attention weights, are leveraged to connect generated reports with original image regions, mimicking the inference process of a human radiologist.

In addition to the six detailed data categories above, healthcare data can also be generally classified as spatial or sequential data. Image data primarily encompasses spatial information, whereas temporal tabular data, free text, signal data, and genomic sequence data fall into the sequential data category. Medical videos represent an integration of both spatial and sequential data. The shared characteristics across different modalities pave the way for a unified architecture that is capable of handling various data types. Chen et al. [137] designed DASSA for automatic pattern change detection within any sequential data and demonstrated its potential for analyzing the aforementioned sequential data within a unified framework.

4 | DISCUSSION

As a fundamental component for the successful implementation of ML in healthcare, AutoML with interpretation reduces the barriers to the full lifecycle of ML analyses and provides interpretations for healthcare professionals [198]. Through a systematic literature review, we discussed the methodologies and applications of AutoML with interpretation for six data types: image, free text, tabular data, signals, genomic sequence, and multimodality. We identified three components that have been automated in ML analyses: data preparation, feature engineering, and model development. We summarized four major interpretation methods: feature interaction and importance, data dimensionality reduction, intrinsically interpretable models, and knowledge distillation and rule extraction. Using Table 2, readers can easily identify papers in which AutoML with interpretation and model performance are discussed for their tasks of interest. Despite the promising performance achieved by AutoML with interpretation systems, several challenges persist, including the absence of automatic data preparation, the loose integration of automation and interpretation, and the unmet compatibility with multi-modality. Additionally, the latest advancements in foundation models have the potential to revolutionize AutoML with interpretation.

The first challenge of current AutoML with interpretation systems is the absence of automatic data preparation, as highlighted by the finding that automatic data preparation was integrated into AutoML with interpretation systems in only 18 out of 118 studies [199]. Real-world healthcare records contain issues such as missing values, outliers, inconsistencies, duplicates, and nonstandardization [200]. These issues constitute almost 50%–80% of the overall workload in the complete lifecycle of ML analyses, underscoring the necessity for automated data preparation within the infrastructure of future AutoML systems [201]. Additionally, we suggest that ML engineers should frequently communicate with healthcare professionals during the system design phase to align their work with real-world demands [202]. For instance, although complex ANNs have become the primary choice in some application domains, such as reinforcement learning [203], intrinsically interpretable models are favored in healthcare settings, such as emergency departments [204]. Hence, ML engineers should ensure the inclusion of common intrinsically interpretable models in their systems rather than exclusively incorporating various ANN architectures.

The second challenge identified in the included papers is the loose integration of automation and interpretation. In all the included studies, the researchers addressed interpretation issues to some extent. Researchers should leverage the insights gained from interpretation to enhance their model automation rather than merely adding post hoc explanations as the last module in their frameworks. A good demonstration was provided by Ikemura et al. [98]. They applied SHAP and PD plots to analyze the decision processes of their AutoML models, indicated potential medical knowledge from their studies, and further reused these findings to enhance their models. The interaction between model development and model interpretation can be achieved by automated feature selection, which reveals feature importance, offers model interpretation, simplifies model structure, and potentially enhances model performance [205]. In addition to automated feature selection, for future AutoML with interpretation, researchers should explore the research direction of developing the tightly knit integration of AutoML and ML interpretations.

Furthermore, the expanding collection of multimodalities presents an opportunity for ML engineers to develop an AutoML with interpretation system that emulates a human clinician's inference process based on various types of healthcare data [177]. Specifically, when patients visit a hospital, clinicians and nurses investigate their former medical records, which are in the form of text and tabular data. Then, some tests may be conducted on the feedback image and signal data. Some advanced treatments involve genome sequencing, which introduces genetic data into the consultation and diagnosis. Handling such abundant and complex information requires a great deal of domain knowledge. The scenario becomes even more intricate when healthcare professionals seek to leverage ML, and this is an exact application scenario for AutoML with interpretation systems. Given the recent versatile application of the transformer for the data types image [206], free text [207], signal data [208], and genomic sequence [209], future researchers can explore the development of comprehensive AI doctors that use multi-modal healthcare data as inputs, automate the entire pipeline of data analyses, and generate results along with interpretations based on a unified backbone architecture.

Recent advancements in foundation models for text, image, and multi-modality have the potential to significantly enhance all three stages of ML: data preparation, feature engineering, and model development [210]. These models excel in zero-shot learning, which enables them to perform tasks without additional training on specific datasets. For example, large language models, such as ChatGPT, can perform a range of tasks from ICD code extraction [211] to risk triage prediction [212] based on prompts provided by healthcare professionals. This zero-shot capability elevates ML to an unprecedented level of automation, potentially obviating the need for tedious data preparation and computationally intensive model development in certain tasks [213]. By contrast, in tasks in which foundation models exhibit suboptimal performance, they can serve as effective tools for feature engineering. The representations within their architectures can be extracted to enhance downstream models [214]; in previous studies, researchers validated that downstream models embedded with these representations outperformed powerful baseline models [215].

Our study had certain limitations that warrant refinement in future work. First, we sought to provide an overview of current AutoML with interpretation systems in healthcare settings. Hence, we did not consider the technical details of AutoML and interpretation techniques. For readers interested in these technical intricacies, we recommend referring to the original papers for a more in-depth exploration. In future work, we may conduct a detailed review of areas such as the underlying algorithms, methodologies, and implementation frameworks. Second, for a given data modality, various commercial software and homemade solutions are readily available, as illustrated above. Although Figure 4 exemplifies the effectiveness of AutoML in predicting in-hospital mortality for a real-world application, we refrained from suggesting a one-size-fits-all solution because of the heterogeneous properties of datasets across different scenarios. In future endeavors, we could undertake a thorough benchmarking analysis to delineate guidelines. An exemplary precedent is in the investigation conducted by Gijsbers et al. [216], wherein they meticulously scrutinized 9 AutoML frameworks across 71 classification and 33 regression tasks. Finally, to ensure that all the reviewed papers underwent peer review, we excluded preprints, which may have resulted in the latest developments in the field being overlooked. In future studies, we could explore the integration of bibliometric methodologies to discern high-quality preprints from a broader pool, thereby enhancing the comprehensiveness of paper inclusion [217].

5 | CONCLUSION

AutoML with interpretation is essential for the successful uptake of ML by healthcare professionals. This review provides a comprehensive summary of the current state of AutoML with interpretation systems in the context of healthcare. To some extent, the proposed systems facilitate effortless development and improve users' trust in ML in healthcare settings. In future studies, researchers should focus on automated data preparation, the seamless integration of automation and interpretation, compatibility with multi-modalities, and the utilization of foundation models to expedite clinical implementation.

AUTHOR CONTRIBUTIONS

Han Yuan: Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); visualization (equal); writing—original draft (lead); writing—review & editing (lead). **Kunyu Yu**: Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); writing—original draft (equal); writing—review & editing (equal). **Feng Xie**: Conceptualization (equal); data curation (equal); investigation (equal); methodology (equal); writing—original draft (equal); writing—review & editing (equal). **Mingxuan Liu**: Formal analysis (supporting); investigation (supporting); writing—original draft (supporting); writing—review & editing (supporting), **Shenghuan Sun**: Formal analysis; investigation; writing—original draft.

ACKNOWLEDGMENTS

We would like to acknowledge Prof. Nan Liu at Duke-NUS Medical School for his invaluable support. Additionally, we thank Dr. Ahmed Allam and Prof. Michael Krauthammer at the University of Zurich for their helpful discussions. The funder played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ETHICS STATEMENT

This study is exempt from review by the ethics committee because it does not involve human participants, animal subjects, or sensitive data collection.

ORCID

Han Yuan https://orcid.org/0000-0002-2674-6068

REFERENCES

[1] Xie F, Yuan H, Ning Y, Ong MEH, Feng M, Hsu W, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. J Biomed Inform. 2022;126:103980. https://doi.org/10.1016/j.jbi.2021.103980

[2] Ting DSW, Cheung CYL, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. J Am Med Assoc. 2017;318(22):2211. https://doi.org/10.1001/jama.2017.18152

[3] Blomberg SN, Folke F, Ersbøll AK, Christensen HC, Torp-Pedersen C, Sayre MR, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation. 2019;138:322–9. https://doi.org/10.1016/j.resuscitation.2019.01.015

[4] Zhao YX, Yuan H, Wu Y. Prediction of adverse drug reaction using machine learning and deep learning based on an imbalanced electronic medical records dataset. In: Proceedings of the international conference on medical and health informatics. Kyoto; 2021. p. 17–21. https://doi.org/10.1145/3472813.3472817

[5] Yao Q, Wang M, Chen Y, Dai W, Hu Y, Li Y, et al. Taking human out of learning applications: a survey on automated machine learning. arXiv. 2018.

[6] Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. Artif Intell Med. 2020;104:101822. https://doi.org/10.1016/j.artmed.2020.101822

[7] Mehta Y, White C, Zela A, Krishnakumar A, Zabergja G, Moradian S, et al. NAS-Bench-Suite: NAS evaluation is (now) surprisingly easy. In: Proceedings of the international conference on learning representations; 2022.

[8] Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. Proceedings of the Advances in Neural Information Processing Systems. 2015.

[9] He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. Knowl Based Syst. 2021;212:106622. https://doi.org/10.1016/j.knosys.2020.106622

[10] Quinn TP, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inf Assoc. 2021;28(4):890–4. https://doi.org/10.1093/jamia/ocaa268

[11] Alaa A, Schaar M. Autoprognosis: automated clinical prognostic modeling via bayesian optimization with structured kernel learning. In: Proceedings of the international conference on machine learning; 2018.

[12] Zöller MA, Huber MF. Benchmark and survey of automated machine learning frameworks. Jair. 2021;70:409–72. https://doi.org/10.1613/jair.1.11854

[13] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat

Methods. 2021;18(2):203–11. https://doi.org/10.1038/s41592-020-01008-z

[14] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI):toward medical XAI. IEEE Transact Neural Networks Learn Syst. 2021;32(11):4793–813. https://doi.org/10.1109/TNNLS.2020.3027314

[15] Dey S, Chakraborty P, Kwon BC, Dhurandhar A, Ghalwash M, Suarez Saiz FJ, et al. Human-centered explainability for life sciences, healthcare, and medical informatics. Patterns. 2022;3(5):100493. https://doi.org/10.1016/j.patter.2022.100493

[16] ElShawi R, Sherif Y, Al-Mallah M, Sakr S. Interpretability in healthcare: a comparative study of local machine learning interpretability techniques. Comput Intell. 2021;37(4):1633–50. https://doi.org/10.1111/coin.12410

[17] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. Jair. 2021;70:245–317. https://doi.org/10.1613/jair.1.12228

[18] Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic healthrecord data: a systematic scoping review. J Am Med Inf Assoc. 2020;27(7):1173–85. https://doi.org/10.1093/jamia/ocaa053

[19] Du M, Liu N, Hu X. Techniques for interpretable machine learning. Commun ACM. 2019;63(1):68–77. https://doi.org/10.1145/3359786

[20] Liberati A. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med. 2009;151(4):W. https://doi.org/10.7326/0003-4819-151-4-200908180-00136

[21] Alkhalaf S, Alturise F, Bahaddad AA, Elnaim BME, Shabana S, Abdel-Khalek S, et al. Adaptive *Aquila* optimizer with explainable artificial intelligence-enabled cancer diagnosis on medical imaging. Cancers. 2023;15(5):1492. https://doi.org/10.3390/cancers15051492

[22] Van Den Berghe T, Babin D, Chen M, Callens M, Brack D, Maes H, et al. Neural network algorithm for detection of erosions and ankylosis on CT of the sacroiliac joints: multicentre development and validation of diagnostic accuracy. Eur Radiol. 2023;33(11):8310–23. https://doi.org/10.1007/s00330-023-09704-y

[23] Cabon S, Weber R, Simon A, Pladys P, Porée F, Carrault G. Functional age estimation through neonatal motion characterization using continuous video recordings. IEEE J Biomed Health Inform. 2023;27(3):1500–11. https://doi.org/10.1109/JBHI.2022.3230061

[24] Choi Y, Wahi-Anwar MW, Brown MS. SimpleMind: an open-source software environment that adds thinking to deep neural networks. PLoS One. 2023;18(4):e0283587. https://doi.org/10.1371/journal.pone.0283587

[25] Custode LL, Mento F, Tursi F, Smargiassi A, Inchingolo R, Perrone T, et al. Multi-objective automatic analysis of lung ultrasound data from COVID-19 patients by means of deep learning and decision trees. Appl Soft Comput. 2023;133:109926. https://doi.org/10.1016/j.asoc.2022.109926

[26] Dai D, Dong C, Li Z, Xu S. MS-Net: learning to assess the malignant status of a lung nodule by a radiologist and her peers. J Applied Clin Med Phys. 2023;24(7):e13964. https://doi.org/10.1002/acm2.13964

[27] Gerbasi A, Clementi G, Corsi F, Albasini S, Malovini A, Quaglini S, et al. DeepMiCa: automatic segmentation and classification of breast MIcroCAlcifications from mammograms. Comput Methods Progr Biomed. 2023;235:107483. https://doi.org/10.1016/j.cmpb.2023.107483

[28] Ghassemi N, Shoeibi A, Khodatars M, Heras J, Rahimi A, Zare A, et al. Automatic diagnosis of COVID-19 from CT images using CycleGAN and transfer learning. Appl Soft Comput. 2023;144:110511. https://doi.org/10.1016/j.asoc.2023.110511

[29] Jun Y, Park YW, Shin H, Shin Y, Lee JR, Han K, et al. Intelligent noninvasive meningioma grading with a fully automatic segmentation using interpretable multiparametric deep learning. Eur Radiol. 2023;33(9):6124–33. https://doi.org/10.1007/s00330-023-09590-4

[30] Leong C, Xiao Y, Yun Z, Iskander MF. Non-invasive assessment of lung water content using chest patch RF sensors: a computer study using NIH patients CT scan database and AI classification algorithms. IEEE Access. 2023;11:13058–66. https://doi.org/10.1109/ACCESS.2023.3238969

[31] Orton MR, Hann E, Doran SJ, Shepherd STC, Ap Dafydd D, Spencer CE, et al. Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the TRACERx Renal study. Cancer Imag. 2023;23(1):76. https://doi.org/10.1186/s40644-023-00594-3

[32] Pham MD, Balezo G, Tilmant C, Petit S, Salmon I, Ben HS, et al. Interpretable HER2 scoring by evaluating clinical guidelines through a weakly supervised, constrained deep learning approach. Comput Med Imag Graph. 2023;108:102261. https://doi.org/10.1016/j.compmedimag.2023.102261

[33] Saglam Y, Oz A, Yildiz G, Ermis C, Kargin OA, Arslan S, et al. Can diffusion tensor imaging have a diagnostic utility to differentiate early-onset forms of bipolar disorder and schizophrenia: a neuroimaging study with explainable machine learning algorithms. Psychiatry Res Neuroimaging. 2023;335:111696. https://doi.org/10.1016/j.pscychresns.2023.111696

[34] Taşcı B. Attention deep feature extraction from brain MRIs in explainable mode: DGXAINet. Diagnostics. 2023;13(5):859. https://doi.org/10.3390/diagnostics13050859

[35] Wang Y, He N, Zhang C, Zhang Y, Wang C, Huang P, et al. An automatic interpretable deep learning pipeline for accurate Parkinson's disease diagnosis using quantitative susceptibility mapping and T1-weighted images. Hum Brain Mapp. 2023;44(12):4426–38. https://doi.org/10.1002/hbm.26399

[36] Xiang J, Wang X, Wang X, Zhang J, Yang S, Yang W, et al. Automatic diagnosis and grading of Prostate Cancer with weakly supervised learning on whole slide images. Comput Biol Med. 2023;152:106340. https://doi.org/10.1016/j.compbiomed.2022.106340

[37] Yoon K, Kim JY, Kim SJ, Huh JK, Kim JW, Choi J. Explainable deep learning-based clinical decision support

engine for MRI-based automated diagnosis of temporomandibular joint anterior disk displacement. Comput Methods Progr Biomed. 2023;233:107465. https://doi.org/10.1016/j.cmpb.2023.107465

[38] Yu W, Zhou H, Choi Y, Goldin JG, Teng P, Wong WK, et al. Multi-scale, domain knowledge-guided attention + random forest: a two-stage deep learning-based multi-scale guided attention models to diagnose idiopathic pulmonary fibrosis from computed tomography images. Med Phys. 2023;50(2):894–905. https://doi.org/10.1002/mp.16053

[39] Basso MN, Barua M, John R, Khademi A. Explainable biomarkers for automated glomerular and patient-level disease classification. Kidney360. 2022;3(3):534–45. https://doi.org/10.34067/kid.0005102021

[40] Chen H, Unberath M, Dreizin D. Toward automated interpretable AAST grading for blunt splenic injury. Emerg Radiol. 2023;30(1):41–50. https://doi.org/10.1007/s10140-022-02099-1

[41] De Falco I, De Pietro G, Sannino G. Classification of Covid-19 chest X-ray images by means of an interpretable evolutionary rule-based approach. Neural Comput Appl. 2023;35(22):16061–71. https://doi.org/10.1007/s00521-021-06806-w

[42] Kakileti ST, Shrivastava R, Manjunath G, Vidyasagar M, Graewingholt A. Automated vascular analysis of breast thermograms with interpretable features. J Med Imag. 2022;9(4):044502. https://doi.org/10.1117/1.jmi.9.4.044502

[43] Maqsood S, Damaševičius R, Maskeliūnas R. Multi-modal brain tumor detection using deep neural network and multiclass SVM. Medicina. 2022;58(8):1090. https://doi.org/10.3390/medicina58081090

[44] McCay KD, Hu P, Shum HPH, Woo WL, Marcroft C, Embleton ND, et al. A pose-based feature fusion and classification framework for the early prediction of cerebral palsy in infants. IEEE Trans Neural Syst Rehabil Eng. 2022;30:8–19. https://doi.org/10.1109/tnsre.2021.3138185

[45] Mou L, Qi H, Liu Y, Zheng Y, Matthew P, Su P, et al. DeepGrading: deep learning grading of corneal nerve tortuosity. IEEE Trans Med Imag. 2022;41(8):2079–91. https://doi.org/10.1109/tmi.2022.3156906

[46] Nafisah SI, Muhammad G. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence. Neural Comput Appl. 2024;36(1):111–31. https://doi.org/10.1007/s00521-022-07258-6

[47] Nijiati M, Zhou R, Damaola M, Hu C, Li L, Qian B, et al. Deep learning based CT images automatic analysis model for active/non-active pulmonary tuberculosis differential diagnosis. Front Mol Biosci. 2022;9:1086047. https://doi.org/10.3389/fmolb.2022.1086047

[48] Nillmani, Saba L, Khanna N, Kalra M, Fouda M, Suri JS. Segmentation-based classification deep learning model embedded with explainable AI for COVID-19 detection in chest X-ray scans. Diagnostics. 2022;12(9):2132. https://doi.org/10.3390/diagnostics12092132

[49] Park YW, Eom J, Kim D, Ahn SS, Kim EH, Kang SG, et al. A fully automatic multiparametric radiomics model for differentiation of adult pilocytic astrocytomas from high-grade gliomas. Eur Radiol. 2022;32(7):4500–9. https://doi.org/10.1007/s00330-022-08575-z

[50] Sharma A, Mishra PK. Covid-MANet: multi-task attention network for explainable diagnosis and severity assessment of COVID-19 from CXR images. Pattern Recogn. 2022;131:108826. https://doi.org/10.1016/j.patcog.2022.108826

[51] Suri J, Agarwal S, Chabert G, Carriero A, Paschè A, Danna P, et al. COVLIAS 2.0-cXAI: cloud-based explainable deep learning system for COVID-19 lesion localization in computed tomography scans. Diagnostics. 2022;12(6):1482. https://doi.org/10.3390/diagnostics12061482

[52] Ullah F, Moon J, Naeem H, Jabbar S. Explainable artificial intelligence approach in combating real-time surveillance of COVID19 pandemic from CT scan and X-ray images using ensemble model. J Supercomput. 2022;78(17):19246–71. https://doi.org/10.1007/s11227-022-04631-z

[53] Fu G, Li J, Wang R, Ma Y, Chen Y. Attention-based full slice brain CT image diagnosis with explanations. Neurocomputing. 2021;452:263–74. https://doi.org/10.1016/j.neucom.2021.04.044

[54] Horry M, Chakraborty S, Pradhan B, Paul M, Gomes D, Ul-Haq A, et al. Deep mining generation of lung cancer malignancy models from chest X-ray images. Sensors. 2021;21(19):6655. https://doi.org/10.3390/s21196655

[55] Kang M, Hong KS, Chikontwe P, Luna M, Jang JG, Park J, et al. Quantitative assessment of chest CT patterns in COVID-19 and bacterial pneumonia patients: a deep learning perspective. J Kor Med Sci. 2021;36(5):1–14. https://doi.org/10.3346/jkms.2021.36.e46

[56] Pietsch M, Ho A, Bardanzellu A, Ahmad Zeidan AM, Chappell LC, Hajnal JV, et al. APPLAUSE: automatic Prediction of PLAcental health via U-net Segmentation and statistical Evaluation. Med Image Anal. 2021;72:102145. https://doi.org/10.1016/j.media.2021.102145

[57] Shorfuzzaman M, Hossain MS, El Saddik A. An explainable deep learning ensemble model for robust diagnosis of diabetic retinopathy grading. ACM Trans Multimed Comput Commun Appl. 2021;17(3s):1–24. https://doi.org/10.1145/3469841

[58] Zhao C, Xu Y, He Z, Tang J, Zhang Y, Han J, et al. Lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images. Pattern Recogn. 2021;119:108071. https://doi.org/10.1016/j.patcog.2021.108071

[59] He X, Wang S, Chu X, Shi S, Tang J, Liu X, et al. Automated model design and benchmarking of deep learning models for COVID-19 detection with chest CT scans. Proc AAAI Conf Artif Intell. 2021;35(6):4821–9. https://doi.org/10.1609/aaai.v35i6.16614

[60] Boumaraf S, Liu X, Wan Y, Zheng Z, Ferkous C, Ma X, et al. Conventional machine learning versus deep learning for magnification dependent histopathological breast cancer image classification: a comparative study with visual explanation. Diagnostics. 2021;11(3):528. https://doi.org/10.3390/diagnostics11030528

[61] Cheung CY, Xu D, Cheng CY, Sabanayagam C, Tham YC, Yu M, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. Nat Biomed Eng. 2020;5(6):498–508. https://doi.org/10.1038/s41551-020-00626-4

[62] Mosquera C, Diaz FN, Binder F, Rabellino JM, Benitez SE, Beresñak AD, et al. Chest X-ray automated triage: a semiologic approach designed for clinical implementation, exploiting different types of labels through a combination of four Deep Learning architectures. Comput Methods Progr Biomed. 2021;206:106130. https://doi.org/10.1016/j.cmpb.2021.106130

[63] Tamarappoo BK, Lin A, Commandeur F, McElhinney PA, Cadet S, Goeller M, et al. Machine learning integration of circulating and imaging biomarkers for explainable patient-specific prediction of cardiac events: a prospective study. Atherosclerosis. 2021;318:76–82. https://doi.org/10.1016/j.atherosclerosis.2020.11.008

[64] Yan H, Mao X, Yang X, Xia Y, Wang C, Wang J, et al. Development and validation of an unsupervised feature learning system for leukocyte characterization and classification: a multi-hospital study. Int J Comput Vis. 2021;129(6):1837–56. https://doi.org/10.1007/s11263-021-01449-9

[65] Rucco M, Viticchi G, Falsetti L. Towards personalized diagnosis of glioblastoma in fluid-attenuated inversion recovery (FLAIR) by topological interpretable machine learning. Mathematics. 2020;8(5):770. https://doi.org/10.3390/math8050770

[66] van der Putten JV, Struyvenberg M, de Groof JD, Scheeve T, Curvers W, Schoon E, et al. Deep principal dimension encoding for the classification of early neoplasia in Barrett's Esophagus with volumetric laser endomicroscopy. Comput Med Imag Graph. 2020;80:101701. https://doi.org/10.1016/j.compmedimag.2020.101701

[67] Wang J, Liu X, Wang F, Zheng L, Gao F, Zhang H, et al. Automated interpretation of congenital heart disease from multi-view echocardiograms. Med Image Anal. 2021;69:101942. https://doi.org/10.1016/j.media.2020.101942

[68] Yin PN, Kc K, Wei S, Yu Q, Li R, Haake AR, et al. Histopathological distinction of non-invasive and invasive bladder cancers using machine learning approaches. BMC Med Inf Decis Making. 2020;20(1):162. https://doi.org/10.1186/s12911-020-01185-z

[69] Lecouat B, Ponce J, Mairal J. A flexible framework for designing trainable priors with adaptive smoothing and game encoding. In: Proceedings of the advances in neural information processing systems; 2020.

[70] Wu M, Zhong X, Peng Q, Xu M, Huang S, Yuan J, et al. Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" machine learning approach in a multi-modal imaging setting. Eur J Radiol. 2019;114:175–84. https://doi.org/10.1016/j.ejrad.2019.03.015

[71] Yamamoto Y, Tsuzuki T, Akatsuka J, Ueki M, Morikawa H, Numata Y, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. Nat Commun. 2019;10(1):5642. https://doi.org/10.1038/s41467-019-13647-8

[72] Pereira S, Meier R, McKinley R, Wiest R, Alves V, Silva CA, et al. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation. Med Image Anal. 2018;44:228–44. https://doi.org/10.1016/j.media.2017.12.009

[73] Song X, Lu H. Multilinear regression for embedded feature selection with application to fMRI analysis. Proc AAAI Conf Artif Intell. 2017;31(1):2562–8. https://doi.org/10.1609/aaai.v31i1.10871

[74] Diao X, Huo Y, Zhao S, Yuan J, Cui M, Wang Y, et al. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. Int J Med Inf. 2021;153:104543. https://doi.org/10.1016/j.ijmedinf.2021.104543

[75] Kulshrestha S, Dligach D, Joyce C, Gonzalez R, O'Rourke AP, Glazer JM, et al. Comparison and interpretability of machine learning models to predict severity of chest injury. JAMIA Open. 2021;4(1):1–8. https://doi.org/10.1093/jamiaopen/ooab015

[76] Blanco A, Pérez A, Casillas A, Cobos D. Extracting cause of death from verbal autopsy with deep learning interpretable methods. IEEE J Biomed Health Inform. 2021;25(4):1315–25. https://doi.org/10.1109/JBHI.2020.3005769

[77] Dong H, Suárez-Paniagua V, Whiteley W, Wu H. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. J Biomed Inform. 2021;116:103728. https://doi.org/10.1016/j.jbi.2021.103728

[78] Yang Z, Chen H, Zhang J, Ma J, Chang Y. Attention-based multi-level feature fusion for named entity recognition. In: Proceedings of the twenty-ninth international Joint conference on artificial intelligence; 2020. p. 3594–600. https://doi.org/10.24963/ijcai.2020/497

[79] Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network. Proc AAAI Conf Artif Intell. 2020;34(5):8180–7. https://doi.org/10.1609/aaai.v34i05.6331

[80] Atutxa A, de Ilarraza AD, Gojenola K, Oronoz M, Perez-de-Viñaspre O. Interpretable deep learning to map diagnostic texts to ICD-10 codes. Int J Med Inf. 2019;129:49–59. https://doi.org/10.1016/j.ijmedinf.2019.05.015

[81] Duarte F, Martins B, Pinto CS, Silva MJ. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. J Biomed Inform. 2018;80:64–77. https://doi.org/10.1016/j.jbi.2018.02.011

[82] Li L, Wang H, Zha L, Huan Q, Wu S, Chen G, et al. Learning a data-driven policy network for pre-training automated feature engineering. In: Proceedings of the international conference on learning representations; 2022.

[83] Junaid M, Ali S, Eid F, El-Sappagh S, Abuhmed T. Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease. Comput Methods Progr Biomed. 2023;234:107495. https://doi.org/10.1016/j.cmpb.2023.107495

[84] Islam MM, Alam MJ, Maniruzzaman M, Ahmed NAMF, Ali MS, Rahman MJ, et al. Predicting the risk of hypertension using machine learning algorithms: a cross sectional study in Ethiopia. PLoS One. 2023;18(8):e0289613. https://doi.org/10.1371/journal.pone.0289613

[85] Wang S, Du X, Liu G, Xing H, Jiao Z, Yan J, et al. An interpretable data-driven medical knowledge discovery pipeline based on artificial intelligence. IEEE J Biomed Health Inform. 2023;27(10):5099–109. https://doi.org/10.1109/jbhi.2023.3299339

[86] Zhang R, Yin M, Jiang A, Zhang S, Liu L, Xu X. Application value of the automated machine learning model based on modified computed tomography severity index combined with serological indicators in the early prediction of severe acute pancreatitis. J Clin Gastroenterol. 2023;58(7):692–701. https://doi.org/10.1097/mcg.0000000000001909

[87] Martínez-Agüero S, Soguero-Ruiz C, Alonso-Moral JM, Mora-Jiménez I, Álvarez-Rodríguez J, Marques AG. Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance. Future Generat Comput Syst. 2022;133:68–83. https://doi.org/10.1016/j.future.2022.02.021

[88] 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. https://doi.org/10.1371/journal.pone.0265254

[89] Cui Y, Shi X, Wang S, Qin Y, Wang B, Che X, et al. Machine learning approaches for prediction of early death among lung cancer patients with bone metastases using routine clinical characteristics: an analysis of 19, 887 patients. Front Public Health. 2022;10:1019168. https://doi.org/10.3389/fpubh.2022.1019168

[90] 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. Int J Mol Sci. 2022;23(13):7132. https://doi.org/10.3390/ijms23137132

[91] Thongprayoon C, Pattharanitima P, Kattah AG, Mao MA, Keddis MT, Dillon JJ, et al. Explainable preoperative automated machine learning prediction model for cardiac surgery-associated acute kidney injury. J Clin Med. 2022;11(21):6264. https://doi.org/10.3390/jcm11216264

[92] 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;12:886935. https://doi.org/10.3389/fcimb.2022.886935

[93] Yu C, Li Y, Yin M, Gao J, Xi L, Lin J, et al. Automated machine learning in predicting 30-day mortality in patients with non-cholestatic cirrhosis. J Personalized Med. 2022;12(11):1930. https://doi.org/10.3390/jpm12111930

[94] Zhang S, Wang J, Pei L, Liu K, Gao Y, Fang H, et al. Interpretable CNN for ischemic stroke subtype classification with active model adaptation. BMC Med Inf Decis Making. 2022;22(1):3. https://doi.org/10.1186/s12911-021-01721-5

[95] Alaa AM, Gurdasani D, Harris AL, Rashbass J, van der Schaar M. Machine learning to guide the use of adjuvant therapies for breast cancer. Nat Mach Intell. 2021;3(8):716–26. https://doi.org/10.1038/s42256-021-00353-8

[96] Chiang PH, Wong M, Dey S. Using wearables and machine learning to enable personalized lifestyle recommendations to improve blood pressure. IEEE J Transl Eng Health Med. 2021;9:2700513. https://doi.org/10.1109/JTEHM.2021.3098173

[97] Laria JC, Delgado-Gómez D, Peñuelas-Calvo I, Baca-García E, Lillo RE. Accurate prediction of children's ADHD severity using family burden information: a neural lasso approach. Front Comput Neurosci. 2021;15:674028. https://doi.org/10.3389/fncom.2021.674028

[98] 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. https://doi.org/10.2196/23458

[99] Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Koebnick C. Generalizability of an automatic explanation method for machine learning prediction results on asthma-related hospital visits in patients with asthma: quantitative analysis. J Med Internet Res. 2021;23(4):e24153. https://doi.org/10.2196/24153

[100] Tong Y, Messinger AI, Luo G. Testing the generalizability of an automated method for explaining machine learning predictions on asthma patients' asthma hospital visits to an academic healthcare system. IEEE Access. 2020;8:195971–9. https://doi.org/10.1109/ACCESS.2020.3032683

[101] Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N. AutoScore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. JMIR Med Inform. 2020;8(10):e21798. https://doi.org/10.2196/21798

[102] Yang F, Zou Q. mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. Database. 2020;2020:baaa050. https://doi.org/10.1093/database/baaa050

[103] Senderovich A, Beck JC, Gal A, Weidlich M. Congestion graphs for automated time predictions. Proc AAAI Conf Artif Intell. 2019;33(1):4854–61. https://doi.org/10.1609/aaai.v33i01.33014854

[104] Banerjee S. Automated interpretable computational biology in the clinic: a framework to predict disease severity and stratify patients from clinical data. Interdiscip Descr Complex Syst. 2017;15(3):199–208. https://doi.org/10.7906/indecs.15.3.4

[105] Billiet L, Van Huffel S, Van Belle V. Interval Coded Scoring: a toolbox for interpretable scoring systems. Peerj Comput Sci. 2018;4:e150. https://doi.org/10.7717/peerj-cs.150

[106] Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (*Pythia*):a retrospective, single-site study. PLoS Med. 2018;15(11):e1002701. https://doi.org/10.1371/journal.pmed.1002701

[107] Khurana U, Samulowitz H, Turaga D. Feature engineering for predictive modeling using reinforcement learning. Proc AAAI Conf Artif Intell. 2018;32(1). https://doi.org/10.1609/aaai.v32i1.11678

[108] De Laet T, Papageorgiou E, Nieuwenhuys A, Desloovere K. Does expert knowledge improve automatic probabilistic classification of gait joint motion patterns in children with cerebral palsy? PLoS One. 2017;12(6):e0178378. https://doi.org/10.1371/journal.pone.0178378

[109] Drakakis G, Moledina S, Chomenidis C, Doganis P, Sarimveis H. Decision trees for continuous data and conditional

mutual information as a criterion for splitting instances. Comb Chem High Throughput Screen. 2016;19(5):423–8. https://doi.org/10.2174/1386207319666160414105217

[110] Keles A, Samet Hasiloglu A, Keles A, Aksoy Y. Neuro-fuzzy classification of prostate cancer using NEFCLASS-J. Comput Biol Med. 2007;37(11):1617–28. https://doi.org/10.1016/j.compbiomed.2007.03.006

[111] Van Der Donckt J, Van Der Donckt J, Deprost E, Vandenbussche N, Rademaker M, Vandewiele G, et al. Do not sleep on traditional machine learning. Biomed Signal Process Control. 2023;81:104429. https://doi.org/10.1016/j.bspc.2022.104429

[112] Heitmann J, Glangetas A, Doenz J, Dervaux J, Shama DM, Garcia DH, et al. DeepBreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries. NPJ Digit Med. 2023;6(1):104. https://doi.org/10.1038/s41746-023-00838-3

[113] Raeisi K, Khazaei M, Tamburro G, Croce P, Comani S, Zappasodi F. A class-imbalance aware and explainable spatio-temporal graph attention network for neonatal seizure detection. Int J Neural Syst. 2023;33(9):2350046. https://doi.org/10.1142/s0129065723500466

[114] Han C, Pan S, Que W, Wang Z, Zhai Y, Shi L. Automated localization and severity period prediction of myocardial infarction with clinical interpretability based on deep learning and knowledge graph. Expert Syst Appl. 2022;209:118398. https://doi.org/10.1016/j.eswa.2022.118398

[115] Huang X, Sun X, Zhang L, Zhu T, Yang H, Xiong Q, et al. A novel epilepsy detection method based on feature extraction by deep autoencoder on EEG signal. Int J Environ Res Public Health. 2022;19(22):15110. https://doi.org/10.3390/ijerph192215110

[116] Jahmunah V, Ng EYK, Tan RS, Oh SL, Acharya UR. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. Comput Biol Med. 2022;146:105550. https://doi.org/10.1016/j.compbiomed.2022.105550

[117] Yang W, Xu J, Xiang J, Yan Z, Zhou H, Wen B, et al. Diagnosis of cardiac abnormalities based on phonocardiogram using a novel fuzzy matching feature extraction method. BMC Med Inf Decis Making. 2022;22(1):230. https://doi.org/10.1186/s12911-022-01976-6

[118] Lee H, Shin M. Learning explainable time-morphology patterns for automatic arrhythmia classification from short single-lead ECGs. Sensors. 2021;21(13):4331. https://doi.org/10.3390/s21134331

[119] Fuchs C, Nobile MS, Zamora G, Degeneffe A, Kubben P, Kaymak U. Tremor assessment using smartphone sensor data and fuzzy reasoning. BMC Bioinform. 2021;22(2):57. https://doi.org/10.1186/s12859-021-03961-8

[120] Kim HS, Ahn MH, Min BK. Deep-learning-based automatic selection of fewest channels for brain–machine interfaces. IEEE Trans Cybern. 2022;52(9):8668–80. https://doi.org/10.1109/TCYB.2021.3052813

[121] Saboo KV, Varatharajah Y, Berry BM, Kremen V, Sperling MR, Davis KA, et al. Unsupervised machine-learning classification of electrophysiologically active electrodes during human cognitive task performance. Sci Rep. 2019;9(1):17390. https://doi.org/10.1038/s41598-019-53925-5

[122] Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. Circ Cardiovasc Qual Outcomes. 2019;12(9). https://doi.org/10.1161/circoutcomes.118.005289

[123] Clauwaert J, Menschaert G, Waegeman W. Explainability in transformer models for functional genomics. Briefings Bioinf. 2021;22(5):1–11. https://doi.org/10.1093/bib/bbab060

[124] Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics. 2020;36(1):250–6. https://doi.org/10.1093/bioinformatics/btz470

[125] Trabelsi A, Chaabane M, Ben-Hur A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. Bioinformatics. 2019;35(14):i269–77. https://doi.org/10.1093/bioinformatics/btz339

[126] Nagorski J, Allen GI. Genomic region detection via spatial convex clustering. PLoS One. 2018;13(9):e0203007. https://doi.org/10.1371/journal.pone.0203007

[127] Shen Y, Wu C, Liu C, Wu Y, Xiong N. Oriented feature selection SVM applied to cancer prediction in precision medicine. IEEE Access. 2018;6:48510–21. https://doi.org/10.1109/ACCESS.2018.2868098

[128] Yap G, Tan A, Pang H. Learning causal models for noisy biological data mining: an application to ovarian cancer detection. Proc AAAI Conf Artif Intell. 2007;22:354–9.

[129] Roest C, Kwee TC, Saha A, Fütterer JJ, Yakar D, Huisman H. AI-assisted biparametric MRI surveillance of prostate cancer: feasibility study. Eur Radiol. 2023;33(1):89–96. https://doi.org/10.1007/s00330-022-09032-7

[130] Wouters PC, van de Leur RR, Vessies MB, van Stipdonk AMW, Ghossein MA, Hassink RJ, et al. Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. Eur Heart J. 2023;44(8):680–92. https://doi.org/10.1093/eurheartj/ehac617

[131] 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. Graefe's Arch Clin Exp Ophthalmol. 2022;260(8):2461–73. https://doi.org/10.1007/s00417-021-05544-y

[132] Gerbasi A, Konduri P, Tolhuisen M, Cavalcante F, Rinkel L, Kappelhof M, et al. Prognostic value of combined radiomic features from follow-up DWI and T2-FLAIR in acute ischemic stroke. Journal of Cardiovascular Development and Disease. 2022;9(12):468. https://doi.org/10.3390/jcdd9120468

[133] García-Gutierrez F, Díaz-Álvarez J, Matias-Guiu JA, Pytel V, Matías-Guiu J, Cabrera-Martín MN, et al. GA-MADRID: design and validation of a machine learning tool for the diagnosis of Alzheimer's disease and frontotemporal dementia using genetic algorithms. Med Biol Eng Comput. 2022;60(9):2737–56. https://doi.org/10.1007/s11517-022-02630-z

[134] Zhang J, Bolanos Trujillo LD, Tanwar A, Ive J, Gupta V, Guo Y. Clinical utility of automatic phenotype annotation in unstructured clinical notes: intensive care unit use. BMJ Health Care Inform. 2022;29(1):e100519. https://doi.org/10.1136/bmjhci-2021-100519

[135] Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronical health record: PheVis algorithm. J Biomed Inform. 2021;117: 103746. https://doi.org/10.1016/j.jbi.2021.103746
[136] Li CY, Liang X, Hu Z, Xing EP. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. Proc AAAI Conf Artif Intell. 2019;33(1):6666–73. https://doi.org/10.1609/aaai.v33i01.33016666
[137] Chen L, Amiri SE, Prakash BA. Automatic segmentation of data sequences. Proc AAAI Conf Artif Intell. 2018;32(1): 2844–51. https://doi.org/10.1609/aaai.v32i1.11815
[138] Guo X, Li R, Yu Q, Haake A. Modeling physicians' utterances to explore diagnostic decision-making. In: Proceedings of the twenty-sixth international Joint conference on artificial intelligence. Melbourne; 2017. p. 3700–6. https://doi.org/10.24963/ijcai.2017/517
[139] Yuan H, Kang L, Li Y, Fan Z. Human-in-the-loop machine learning for healthcare: current progress and future opportunities in electronic health records. Medicine Advances. 2024;2(2):1–5. https://doi.org/10.1002/med4.70
[140] Yuan H, Xie F, Ong MEH, Ning Y, Chee ML, Saffari SE, et al. AutoScore-Imbalance: an interpretable machine learning tool for development of clinical scores with rare events data. J Biomed Inform. 2022;129:104072. https://doi.org/10.1016/j.jbi.2022.104072
[141] Xie F, Ning Y, Liu M, Li S, Saffari SE, Yuan H, et al. A universal AutoScore framework to develop interpretable scoring systems for predicting common types of clinical outcomes. STAR Protoc. 2023;4(2):102302. https://doi.org/10.1016/j.xpro.2023.102302
[142] Wu A, Pashkovski S, Datta S, Pillow J. Learning a latent manifold of odor representations from neural responses in piriform cortex. Proceedings of the Advances in Neural Information Processing Systems. 2018.
[143] Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga D. Learning feature engineering for classification. In: Proceedings of the international Joint conference on artificial intelligence. Melbourne; 2017. p. 2529–35. https://doi.org/10.24963/ijcai.2017/352
[144] Duong-Trung N, Tang NQT, Ha XS. Interpretation of machine learning models for medical diagnosis. Adv Sci Technol Eng Syst J. 2020;5(5):469–77. https://doi.org/10.25046/aj050558
[145] Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):160035. https://doi.org/10.1038/sdata.2016.35
[146] Yuan H, Lee J, Liu M, Li S, Niu C, Wen J, et al. Interpretable machine learning-based risk scoring with individual and ensemble model selection for clinical decision making. In: Proceedings of the international conference on learning representations; 2023.
[147] Erickson N, Mueller J, Shirkow A, Zhang H, Larroy P, Li M, et al. Autogluon-tabular: robust and accurate automl for structured data. In: Proceedings of the ICML workshop on automated machine learning; 2020.
[148] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324
[149] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist. 2001;29(5):1189–232. https://doi.org/10.1214/aos/1013203451
[150] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory. 1967;13(1):21–7. https://doi.org/10.1109/tit.1967.1053964
[151] Yuan H. Toy example of AutoML. 2024. https://github.com/Han-Yuan-Med/toy-example-of-AutoML
[152] Schwartz JM, George M, Rossetti SC, Dykes PC, Minshall SR, Lucas E, et al. Factors influencing clinician trust in predictive clinical decision support systems for In-hospital deterioration: qualitative descriptive study. JMIR Hum Factors. 2022;9(2):e33960. https://doi.org/10.2196/33960
[153] Gamble P, Jaroensri R, Wang H, Tan F, Moran M, Brown T, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. Commun Med. 2021;1:14. https://doi.org/10.1038/s43856-021-00013-3
[154] Wang H, Doumard E, Soulé-Dupuy C, Kémoun P, Aligon J, Monsarrat P. Explanations as a new metric for feature selection: a systematic approach. IEEE J Biomed Health Inform. 2023;27(8):4131–42. https://doi.org/10.1109/JBHI.2023.3279340
[155] Yuan J, Chan GYY, Barr B, Overton K, Rees K, Nonato LG, et al. *SUBPLEX*: a visual analytics approach to understand local model explanations at the subpopulation level. IEEE Comput Grap Appl. 2022;42(6):24–36. https://doi.org/10.1109/mcg.2022.3199727
[156] Xie F, Ning Y, Yuan H, Saffari E, Chakraborty B, Liu N. Package 'AutoScore': an interpretable machine learning-based automatic clinical score generator. R Package. 2022.
[157] Xie F, Ning Y, Yuan H, Goldstein BA, Ong MEH, Liu N, et al. AutoScore-Survival: developing interpretable machine learning-based time-to-event scores with right-censored survival data. J Biomed Inform. 2022;125:103959. https://doi.org/10.1016/j.jbi.2021.103959
[158] Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. Comput Biol Med. 2022;149:106043. https://doi.org/10.1016/j.compbiomed.2022.106043
[159] Wang L, Yoon KJ. Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans Pattern Anal Mach Intell. 2022;44(6):3048–68. https://doi.org/10.1109/TPAMI.2021.3055564
[160] Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: a survey. Int J Comput Vis. 2021;129(6):1789–819. https://doi.org/10.1007/s11263-021-01453-z
[161] Barakat N, Bradley AP. Rule extraction from support vector machines: a review. Neurocomputing. 2010;74(1–3):178–90. https://doi.org/10.1016/j.neucom.2010.02.016
[162] Giuste F, Shi W, Zhu Y, Naren T, Isgut M, Sha Y, et al. Explainable artificial intelligence methods in combating pandemics: a systematic review. IEEE Rev Biomed Eng. 2023;16:5–21. https://doi.org/10.1109/rbme.2022.3185953
[163] Henglin M, Stein G, Hushcha PV, Snoek J, Wiltschko AB, Cheng S. Machine learning approaches in cardiovascular

imaging. Circ Cardiovasc Imag. 2017;10(10):e005614. https://doi.org/10.1161/circimaging.117.005614
[164] 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):e232–42. https://doi.org/10.1016/s2589-7500(19)30108-6
[165] Yuan H, Jiang P, Zhao G. Human-Guided design to explain deep learning-based pneumothorax classifier. In: Proceedings of the medical imaging with deep learning; 2023.
[166] Behadada O, Chikh MA. An interpretable classifier for detection of cardiac arrhythmias by using the fuzzy decision tree. Artif Intell Res. 2013;2(3):45–58. https://doi.org/10.5430/air.v2n3p45
[167] Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006;7(10):R100. https://doi.org/10.1186/gb-2006-7-10-r100
[168] Rueden CT, Eliceiri KW. ImageJ for the next generation of scientific image data. Microsc Microanal. 2019;25(S2):142–3. https://doi.org/10.1017/s1431927619001442
[169] Wang Y, Herrington DM. Machine intelligence enabled radiomics. Nat Mach Intell. 2021;3(10):838–9. https://doi.org/10.1038/s42256-021-00404-0
[170] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005
[171] Nikaidô H, Isoda K. Note on non-cooperative convex game. Pac J Math. 1955;5(5):807–15. https://doi.org/10.2140/pjm.1955.5.807
[172] Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. Int J Med Inf. 2018;114:57–65. https://doi.org/10.1016/j.ijmedinf.2018.03.013
[173] Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med. 2019;2(1):130. https://doi.org/10.1038/s41746-019-0208-8
[174] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: Proceedings of the advances in neural information processing systems; 2017.
[175] Gardner-Thorpe J, Love N, Wrightson J, Walsh S, Keeling N. The value of modified early warning score (MEWS) in surgical In-patients: a prospective observational study. Annals. 2006;88(6):571–5. https://doi.org/10.1308/003588406x130615
[176] Li S, Ning Y, Ong MEH, Chakraborty B, Hong C, Xie F, et al. FedScore: a privacy-preserving framework for federated scoring system development. J Biomed Inform. 2023;146:104485. https://doi.org/10.1016/j.jbi.2023.104485
[177] LeDell E, Poirier S. H2o automl: scalable automatic machine learning. In: Proceedings of the international conference on machine learning AutoML workshop; 2020.
[178] Lundberg S, Lee S. A unified approach to interpreting model predictions. In: Proceedings of the advances in neural information processing systems; 2017.
[179] Lemaître G, Nogueira F, Aridas C. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(1):559–63.
[180] Oudah M, Henschel A. Taxonomy-aware feature engineering for microbiome classification. BMC Bioinform. 2018;19(1):227. https://doi.org/10.1186/s12859-018-2205-3
[181] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(2):281–305.
[182] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Jair. 2002;16:321–57. https://doi.org/10.1613/jair.953
[183] He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international Joint conference on neural networks. Hong Kong: IEEE World Congress on Computational Intelligence); 2008. p. 1322–8.
[184] Quinn TP, Erb I. Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection. mSystems. 2020;5(2):1–11. https://doi.org/10.1128/msystems.00230-19
[185] Brown G, Pocock A, Zhao M, Luján M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res. 2012;13(1):27–66.
[186] Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. Deep learning for healthcare applications based on physiological signals: a review. Comput Methods Progr Biomed. 2018;161:1–13. https://doi.org/10.1016/j.cmpb.2018.04.005
[187] Vieira SM, Sousa JMC, Kaymak U. Fuzzy criteria for feature selection. Fuzzy Set Syst. 2012;189(1):1–18. https://doi.org/10.1016/j.fss.2011.09.009
[188] Fuchs C, Spolaor S, Nobile MS, Kaymak U. pyFUME: a python package for fuzzy model estimation. In: 2020 IEEE international conference on fuzzy systems. Glasgow: FUZZ-IEEE; 2020. p. 1–8.
[189] Huang G, Liu Z, Pleiss G, van der Maaten L, Weinberger KQ. Convolutional networks with dense connectivity. IEEE Trans Pattern Anal Mach Intell. 2022;44(12):8704–16. https://doi.org/10.1109/TPAMI.2019.2918284
[190] Liu B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Briefings Bioinf. 2019;20(4):1280–94. https://doi.org/10.1093/bib/bbx165
[191] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–32. https://doi.org/10.1038/nrg3920
[192] Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a review of computational problems and data sets. Proc IEEE. 2016;104(1):176–97. https://doi.org/10.1109/JPROC.2015.2494198
[193] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x
[194] Shade JK, Prakosa A, Popescu DM, Yu R, Okada DR, Chrispin J, et al. Predicting risk of sudden cardiac death in patients with cardiac sarcoidosis using multimodality imaging and personalized heart modeling in a multivariable classifier. Sci Adv. 2021;7(31):eabi8020. https://doi.org/10.1126/sciadv.abi8020
[195] Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision

oncology. Nat Rev Cancer. 2022;22(2):114–26. https://doi.org/10.1038/s41568-021-00408-3

[196] Cai L, Wang Z, Gao H, Shen D, Ji S. Deep adversarial learning for multi-modality missing data completion. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge Discovery & data mining. London; 2018. p. 1158–66. https://doi.org/10.1145/3219819.3219963

[197] Yu S, Chakrabortty A, Liao KP, Cai T, Ananthakrishnan AN, Gainer VS, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inf Assoc. 2017;24(e1):e143–9. https://doi.org/10.1093/jamia/ocw135

[198] van der Schaar M. AutoML and interpretability: powering the machine learning revolution. In: HealthcareProceedings of the 2020 ACM-IMS on foundations of data science conference; 2020. p. 1. https://doi.org/10.1145/3412815.3416879

[199] Benhar H, Idri A, Fernández-Alemán JL. A systematic mapping study of data preparation in heart disease knowledge discovery. J Med Syst. 2018;43(1):17. https://doi.org/10.1007/s10916-018-1134-z

[200] Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, et al. Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques. Artif Intell Med. 2023;142:102587. https://doi.org/10.1016/j.artmed.2023.102587

[201] Chai CP. The importance of data cleaning: three visualization examples. Chance. 2020;33(1):4–9. https://doi.org/10.1080/09332480.2020.1726112

[202] Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. 2022;6(12):1330–45. https://doi.org/10.1038/s41551-022-00898-y

[203] Kang L, Liu Y, Luo Y, Yang JZ, Yuan H, Zhu C. Approximate policy iteration with deep minimax average bellman error minimization. IEEE Transact Neural Networks Learn Syst. 2024:1–12. https://doi.org/10.1109/TNNLS.2023.3346992

[204] Liu N, Ng JCJ, Ting CE, Sakamoto JT, Ho AFW, Koh ZX, et al. Clinical scores for risk stratification of chest pain patients in the emergency department: an updated systematic review. J Emerg Crit Care Med. 2018;2:16. https://doi.org/10.21037/jeccm.2018.01.10

[205] Yuan H, Liu M, Kang L, Miao C, Wu Y. An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. In: Proceedings of the international conference on learning representations; 2023.

[206] Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):87–110. https://doi.org/10.1109/tpami.2022.3152247

[207] Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. J Biomed Inform. 2022;126:103982. https://doi.org/10.1016/j.jbi.2021.103982

[208] Zhang C, He Y, Du B, Yuan L, Li B, Jiang S. Transformer fault diagnosis method using IoT based monitoring system and ensemble machine learning. Future Gener Comput Syst. 2020;108:533–45. https://doi.org/10.1016/j.future.2020.03.008

[209] Chen K, Zhao H, Yang Y. Capturing large genomic contexts for accurately predicting enhancer-promoter interactions. Briefings Bioinf. 2022;23(2):bbab577. https://doi.org/10.1093/bib/bbab577

[210] 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):259–65. https://doi.org/10.1038/s41586-023-05881-4

[211] Ong J, Kedia N, Harihar S, Vupparaboina SC, Singh SR, Venkatesh R, et al. Applying large language model artificial intelligence for retina International Classification of Diseases (ICD) coding. J Med Artif Intell. 2023;6:21. https://doi.org/10.21037/jmai-23-106

[212] Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroğlu M, İmamoğlu M. Assessing the precision of artificial intelligence in ED triage decisions: insights from a study with ChatGPT. Am J Emerg Med. 2024;78:170–5. https://doi.org/10.1016/j.ajem.2024.01.037

[213] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Proceedings of the Advances in Neural Information Processing Systems. 2022.

[214] Pai S, Bontempi D, Hadzic I, Prudente V, Sokač M, Chaunzwa T, et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell. 2024;6:1–14. https://doi.org/10.1038/s42256-024-00807-9

[215] Azizi S, Culp L, Freyberg J, Mustafa B, Baur S, Kornblith S, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat Biomed Eng. 2023;7(6):756–79. https://doi.org/10.1038/s41551-023-01049-7

[216] Gijsbers P, Bueno M, Coors S, LeDell E, Poirier S, Thomas J, et al. Amlb: an automl benchmark. J Mach Learn Res. 2024;25(101):1–65.

[217] Gordon M, Bishop M, Chen Y, Dreber A, Goldfedder B, Twardy C, et al. Forecasting the publication and citation outcomes of COVID-19 preprints. R Soc Open Sci. 2022;9(9):220440. https://doi.org/10.1098/rsos.220440

How to cite this article: Yuan H, Yu K, Xie F, Liu M, Sun S. Automated machine learning with interpretation: a systematic review of methodologies and applications in healthcare. Med Adv. 2024;1–33. https://doi.org/10.1002/med4.75