Information Fusion 77 (2022) 29–52

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond

Guang Yang [a,b,c,*], Qinghao Ye [d,e], Jun Xia [f,*]

[a] National Heart and Lung Institute, Imperial College London, London, UK
[b] Royal Brompton Hospital, London, UK
[c] Imperial Institute of Advanced Technology, Hangzhou, China
[d] Hangzhou Ocean's Smart Boya Co., Ltd, China
[e] University of California, San Diego, La Jolla, CA, USA
[f] Radiology Department, Shenzhen Second People's Hospital, Shenzhen, China

ARTICLE INFO

Keywords:
Explainable AI
Information fusion
Multi-domain information fusion
Weakly supervised learning
Medical image analysis

ABSTRACT

Explainable Artificial Intelligence (XAI) is an emerging research topic of machine learning aimed at *unboxing* how AI systems' *black-box* choices are made. This research field inspects the measures and models involved in decision-making and seeks solutions to explain them explicitly. Many of the machine learning algorithms cannot manifest how and why a decision has been cast. This is particularly true of the most popular deep neural network approaches currently in use. Consequently, our confidence in AI systems can be hindered by the lack of explainability in these *black-box* models. The XAI becomes more and more crucial for deep learning powered applications, especially for medical and healthcare studies, although in general these deep neural networks can return an arresting dividend in performance. The insufficient explainability and transparency in most existing AI systems can be one of the major reasons that successful implementation and integration of AI tools into routine clinical practice are uncommon. In this study, we first surveyed the current progress of XAI and in particular its advances in healthcare applications. We then introduced our solutions for XAI leveraging multi-modal and multi-centre data fusion, and subsequently validated in two showcases following real clinical scenarios. Comprehensive quantitative and qualitative analyses can prove the efficacy of our proposed XAI solutions, from which we can envisage successful applications in a broader range of clinical questions.

1. Introduction

Recent years have seen significant advances in the capacity of Artificial Intelligence (AI), which is growing in sophistication, complexity and autonomy. A continuously veritable and explosive growth of data with a rapid iteration of computing hardware advancement provides a *turbo boost* for the development of AI.

AI is a generic concept and an umbrella term that implies the use of a machine with limited human interference to model intelligent actions. It covers a broad range of research studies from machine intelligence for computer vision, robotics, natural language processing to more theoretical machine learning algorithms design and recently re-branded and thrived *deep learning* development (Fig. 1).

1.1. Born of AI

AI changes almost every sector globally, e.g., enhancing (digital) healthcare (e.g., making diagnosis more accurate, allowing improved disease prevention), accelerating drug/vaccine development and repurposing, raising agricultural productivity, leading to mitigation and adaptation in climate change, improving the efficiency of manufacturing processes by predictive maintenance, supporting the development of autonomous vehicles and programming more efficient transport networks, and in many other successful applications, which make significant positive socio-economic impact. Besides, AI systems are being deployed in highly-sensitive policy fields, such as facial recognition in the police or recidivism prediction in the criminal justice system, and in areas where diverse social and political forces are presented. Therefore, nowadays, AI systems are incorporated into a wide variety of decision-making processes. As AI systems become integrated into all kinds of decision-making processes, the degree to which people who develop AI, or are subject to an AI-enabled decision, can understand how the resulting decision-making mechanism operates and why a specific decision is reached, has been increasingly debated in science and policy communities.

* Corresponding authors.
E-mail addresses: g.yang@imperial.ac.uk (G. Yang), q7ye@ucsd.edu (Q. Ye), xiajun@email.szu.edu.cn (J. Xia).

https://doi.org/10.1016/j.inffus.2021.07.016
Received 27 January 2021; Received in revised form 25 May 2021; Accepted 25 July 2021
Available online 31 July 2021

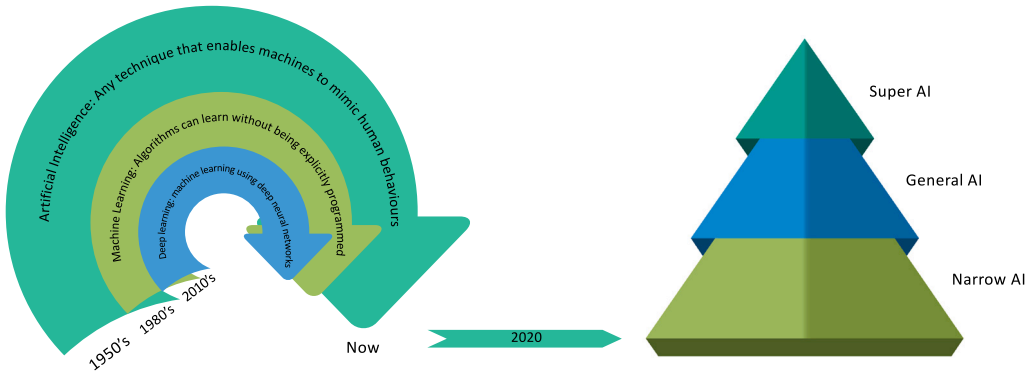


Fig. 1. Left: Terminology and historical timeline of AI, machine learning and deep learning. Right: We are still at the stage of narrow AI, a concept used to describe AI systems that are capable of handling a single or limited task. General AI is the hypothetical wisdom of AI systems capable of comprehending or learning any intelligent activity a human being might perform. Super AI is an AI that exceeds human intelligence and skills.

A collection of innovations, which are typically correlated with human or animal intelligence, is defined as the term "artificial intelligence". John McCarthy, who coined this term in 1955, described it as "the scientific and technical expertise in the manufacture of intelligent machines", and since then many different definitions have been endowed.

1.2. Growth of machine learning

Machine learning is a subdivision of AI that helps computer systems to intelligently execute complex tasks. Traditional AI methods, which specify step by step how to address a problem, are normally based on hard-coded rules. Machine learning framework, by contrast, leverages the power of a large amount of data (as examples and not examples) for the identification of characteristics to accomplish a pre-defined task. The framework then learns how the target output will be better obtained. Three primary subdivisions of machine learning algorithms exist:

- A machine learning framework, which is trained using labelled data, is generally categorised as supervised machine learning. The labels of the data are grouped into one or more classes at each data point, such as "cats" or "dogs". The supervised machine learning framework exploits the nature from these labelled data (i.e., training data), and forecasts the categories of the new or so called test data.
- Learning without labels is referred to as unsupervised learning. The aim is to identify the mutual patterns among data points, such as the formation of clusters and allotting data points to these clusters.
- Reinforcement learning on the other hand is about knowledge learning, i.e., learning from experience. In standard reinforcement learning settings, an agent communicates with its environment, and is given a reward function that it tries to optimise. The purpose of the agent is to understand the effect of its decisions, and discover the best strategies for maximising its rewards during the training and learning procedure.

It is of note that some hybrid methods, e.g., semi-supervised learning (using partially labelled data) and weakly supervised (using indirect labels), are also under development.

Although not achieving the human-level intelligence often associated with the definition of the general AI, the capacity to learn from knowledge increases the amount and sophistication of tasks that can be tackled by machine learning systems (Fig. 1). A wide variety of technologies, many of which people face on a daily basis, are nowadays enabled by rapid developments in machine learning, contributing to current advancements and dispute about the influence of AI in society. Many of the concepts that frame the existing machine learning systems are not new. The mathematical underpinnings of the field date back many decades, and since the 1950s, researchers have developed machine learning algorithms with varying degrees of complexity. In order to forecast results, machine learning requires computers to process a vast volume of data. How systems equipped with machine learning can handle probabilities or uncertainty in decision-making is normally informed by statistical approaches. Statistics, however, often cover areas of research that are not associated with the development of algorithms that can learn to make forecasts or decisions from results. Although several key principles of machine learning are rooted in data science and statistical analysis, some of the complex computational models do not converge with these disciplines naturally. Symbolic approaches, compared to statistical methods, are also used for AI. In order to create interpretations of a problem and to reach a solution, these methods use logic and inference.

1.3. Boom of deep learning

Deep learning is a relatively recent congregation of approaches that have radically transformed machine learning. Deep learning is not an algorithm per se, but a range of algorithms that implements neural networks with deep layers. These neural networks are so deep that they can only be implemented on computer node clusters – modern methods of computing – such as graphics processing units (GPUs), are needed to train them successfully. Deep learning functions very well for vast quantities of data, and it is never too difficult to engineer the functionality even if a problem is complex (for example, due to the unstructured data). When it comes to image detection, natural language processing, and voice recognition, deep learning can always outperform the other types of algorithms. Deep learning assisted disease screening and clinical outcome prediction or automated driving, which were not feasible using previous methods, are well manifested now. Actually, the deeper the neural network with more data loaded for training, the higher accuracy a neural network can produce. The deep learning is very strong, but there are a few disadvantages to it. The reasoning of how deep learning algorithms reach to a certain solution is almost impossible to reveal clearly. Although several tools are now available that can increase insights into the inner workings of the deep learning model, this black-box problem still exists. Deep learning often involves long training cycles, a lot of data and complex hardware specifications, and it is not easy to obtain the specific skills necessary to create a new deep learning approach to tackle a new problem.

Although acknowledging that AI includes a wide variety of scientific areas, this paper uses the umbrella word 'AI' and much of the recent interest in AI has been motivated by developments in machine learning and deep learning. More importantly, we should realise that there is not one algorithm, though, that will adapt or solve all issues. Success

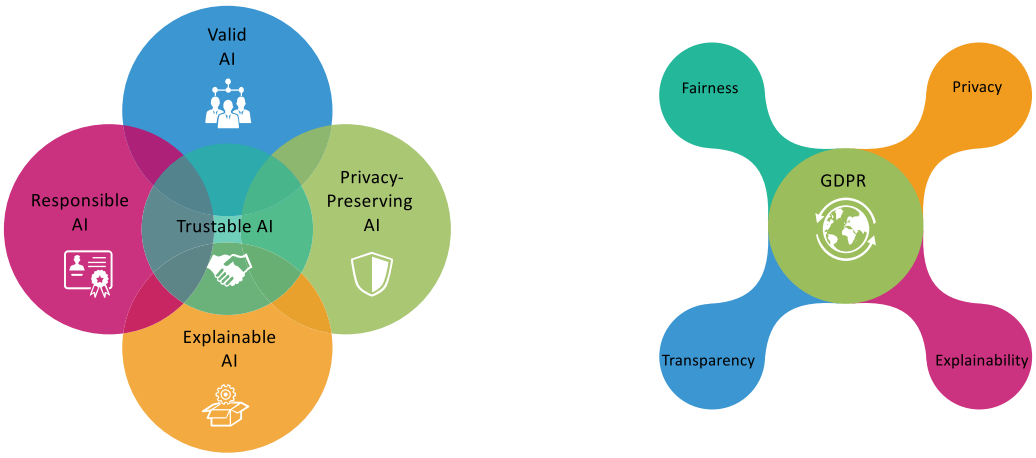


Fig. 2. Left: Trustable AI or Trustworthy AI includes Valid AI, Responsible AI, Privacy-Preserving AI, and Explainable AI (XAI). Right: EU General Data Protection Regulation (GDPR) highlights the Fairness, Privacy, Transparency and Explainability of the AI.

normally depends on the exact problem that needs to be solved and the knowledge available. A hybrid solution is often required to solve the problem, where various algorithms are combined to provide a concrete solution. Each issue involves a detailed analysis into what constitutes the best-fit algorithm. Transparency of the input size, capabilities of the deep neural network and time efficiency should also be taken into consideration, since certain algorithms take a long time to train.

1.4. Stunt by the black-box and promotion of the explainable AI

Any of today's deep learning tools are capable of generating extremely reliable outcomes, but they are often highly opaque, if not fully invisible, making it difficult to understand their behaviours. For even skilled experts to completely comprehend these so-called 'black-box' models may be still difficult. As these deep learning tools are applied on a wide scale, researchers and policymakers can challenge whether the precision of a given task outweighs more essential factors in the decision-making procedure.

As part of attempts to integrate ethical standards into the design and implementation of AI-enabled technologies, policy discussions around the world increasingly involve demands for some form of Trustable AI, which includes Valid AI, Responsible AI, Privacy-Preserving AI, and Explainable AI (XAI), in which the XAI want to address the fundamental question about the rationale of the decision making process including both human level XAI and machine level XAI (Fig. 2). For example, in the UK, such calls came from the AI Committee of the House of Lords, which argued that the development of intelligible AI systems is a fundamental requirement if AI will be integrated as a trustworthy tool for our society. In the EU, the High-Level Group on AI has initiated more studies on the pathway towards XAI (Fig. 2). Similarly, in the USA, the Defence Advanced Research Projects Agency funds a new research effort aiming at the development of AI with more explainability. These discussions will become more urgent as AI approaches are used to solve problems in a wide variety of complicated policy making areas, as experts increasingly work alongside AI-enabled decision-making tools, for example in clinical studies, and as people more regularly experience AI systems in real life when decisions have a major impact. Meanwhile, research studies in AI continue to progress at a steady pace. XAI is a vigorous area with many on-going studies emerging and several new strategies evolving that make a huge impact on AI development in various ways.

While the usage of the term is inconsistent, "XAI" refers to a class of systems that have insight into how an AI system makes decisions and predictions. XAI explores the reasoning for the decision-making process, presents the positives and drawbacks of the system, and offers a glimpse of how the system will act in the future. By offering accessible explanations of how AI systems perform their study, XAI can allow researchers to understand the insights that come from research results. For example, in Fig. 3, an additional explainable surrogate module can be added to the learnt model to achieve a more transparent and trustworthy model. In other words, for a conventional machine or deep learning model, only generalisation error has been considered while adding an explainable surrogate, both generalisation error and human experience can be considered and a verified prediction can be achieved. In contrast, a learnt black-box model without an explainable surrogate module will cause concerns for the end-users although the performance of the learnt model can be high. Such a black-box model can always cause confusions like "Why did you do that?", "Why did you not do that?", "When do you succeed or fail?", "How do I correct an error?", and "Can I trust the prediction?". The XAI powered model, on the other hand, can provide clear and transparent predictions to reassure "I understand why.", "I understand why not.", "I know why you succeed or fail.", "I know how to correct an error.", and "I understand, therefore I trust". A typical feedback loop of the XAI development can be found in Fig. 4, which includes seven steps from training, quality assurance (QA), deployment, prediction, split testing (A/B test), monitoring, and debugging.

A variety of terms are used to define certain desired characteristics of an XAI system in research, public, and policy debates, including:

- Interpretability: it means a sense of knowing how the AI technology functions.
- Explainability: it provides an explanation for a wider range of users that how a decision has been drawn.
- Transparency: it measures the level of accessibility to the data or model.
- Justifiability: it indicates an understanding of the case to support a particular outcome.
- Contestability: it implies how the users can argue against a decision.

Comprehensive surveys on general XAI can be found elsewhere, e.g., [1–4]; therefore, here we provide an overview of most important concepts of the XAI. Broadly speaking, XAI can be categorised into model-specific or model-agnostic based approaches. Besides, these methods can be classified into local or global methods that can be either intrinsic or post-hoc [1]. Essentially, there are many machine learning models that are intrinsically explainable, e.g., linear models, rule-based models and decision trees, which are also known as transparent models or white-box models. However, these relatively simple models may have a relatively lower performance (Fig. 5). For more complex models, e.g., support vector machines (SVM), convolutional neural networks

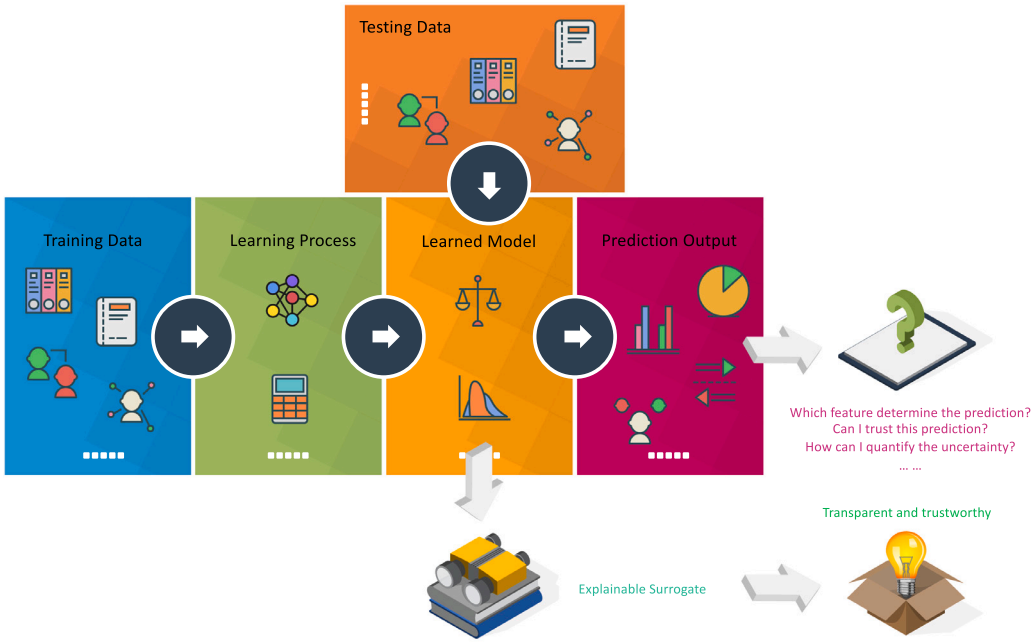


Fig. 3. Schema of the added explainable surrogate module for the normal machine or deep learning procedure that can achieve a more transparent and trustworthy model.

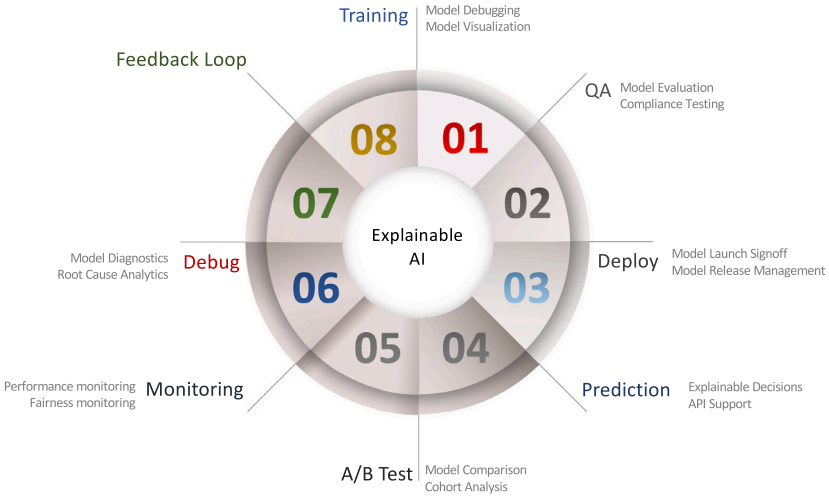


Fig. 4. A typical feedback loop of the XAI development that includes seven steps from training, quality assurance (QA), deployment, prediction, split testing (A/B test), monitoring, and debugging.

(CNN), recurrent neural networks (RNN) and ensemble models, we can design model-specific and post-hoc XAI strategies for each of them. For example, commonly used strategies include explanation by simplification, architecture modification, feature relevance explanation, and visual explanation [1]. In general, these more complex models may achieve better performance while the explainability becomes lower (Fig. 5). It is of note that due to the nature of the problems, less complicated models may still perform well compared to deep learning based models while preserving the explainability. Therefore, Fig. 5 just depicts a preliminary representation inspired by [1], in which XAI demonstrates its ability to enhance the trade-off between model interpretability and efficiency.

Recently, model-agnostic based approaches attract great attention that rely on a simplified surrogate function to explain the predictions [3]. Model-agnostic approaches are not attached to a specific machine learning model. This class of techniques, in other words, distinguishes prediction from explanation. Model-agnostic representations are usually post-hoc that are generally used to explain deep neural networks with interpretable surrogates that can be local or global [2]. Below is some summary for XAI in more complex deep learning based models.

1.4.1. *Model-specific global XAI*

By integrating interpretability constraints into the procedure of deep learning, these model-specific global XAI strategies can improve the understandability of the models. Structural restrictions may include sparsity and monotonicity, where fewer input features are leveraged or the correlation between features and predictions is confined as monotonic). Semantic prior knowledge can also be impelled to restrict the higher-level abstractions derived from the data. For instance, in a CNN based brain tumour detection/classification model using multimodal MRI data fusion, constraints can be imposed by forcing disengaged representations that are recognisable to each MRI modality (e.g., T1, T1 post-contrast and FLAIR), respectively. In doing so, the model can identify crucial information from each MRI modality and distinguish brain tumours and sub-regions into necrotic, more or less infiltrative

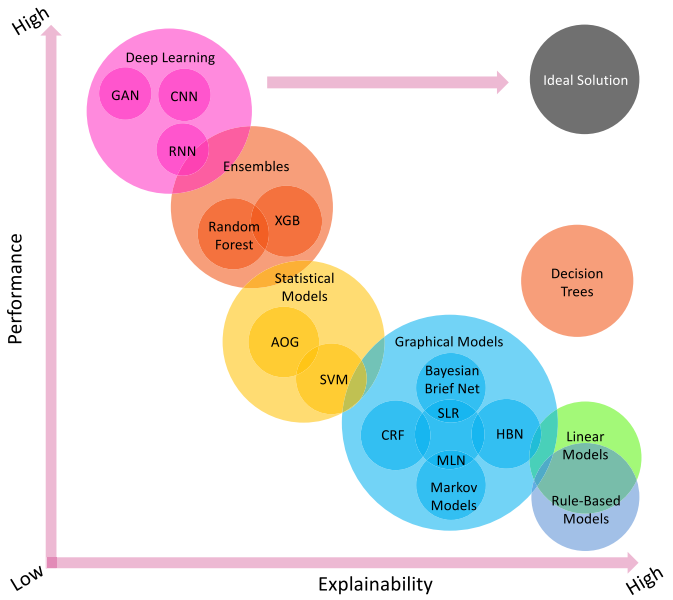


Fig. 5. Model explainability vs. model performance for widely used machine learning and deep learning algorithms. The ideal solution should have both high explainability and high performance. However, existing linear models, rule-based models and decision trees are more transparent, but with lower performance in general. In contrast, complex models, e.g., deep learning and ensembles, manifest higher performance while less explainability can be obtained. HBN: Hierarchical Bayesian Networks; SLR: Simple Linear Regression; CRF: Conditional Random Fields; MLN: Markov Logic Network; SVM: Support Vector Machine; AOG: Stochastic And-Or-Graphs; XGB: XGBoost; CNN: Convolutional Neural Network; RNN: Recurrent Neural Network; and GAN: Generative Adversarial Network.

that can provide vital diagnosis and prognosis information. On the contrary, simple aggregation based information fusion (combining all the multimodal MRI data like a sandwich) would not provide such explainability.

1.4.2. Model-specific local XAI

In a deep learning model, a model-specific local XAI technique offers an interpretation for a particular instance. Recently, novel attention mechanisms have been proposed to emphasise the importance of different features of the high-dimensional input data to provide an explanation of a representative instance. Consider a deep learning algorithm that encodes an X-ray image into a vector using a CNN and then use an RNN to produce a clinical description for the X-ray image by using the encoded vector. For the RNN, an attention module can be applied to explain to the user what image fragments the model focuses on to produce each substantive term for the clinical description. For example, the attention mechanism will represent the appropriate segments of the image corresponding to the clinical key words derived by the deep learning model when a clinician is baffled to link the clinical key words to the regions of interest in the X-ray image.

1.4.3. Model-agnostic global XAI

In model-agnostic global XAI, a surrogate representation is developed to approximate an interpretable module for the black-box model. For instance, an interpretable decision tree based model can be used to approximate a more complex deep learning model on how clinical symptoms impact treatment response. A clarification of the relative importance of variables in affecting treatment response to clinical symptoms can be given by the IF-THEN logic of the decision tree. Clinical experts can analyse these variables and are likely to believe the model to the extent that particular symptomatic factors are known to be rational and confounding noises can be accurately removed. Diagnostic methods can also be useful to produce insights into the significance of individual characteristics in the predictions of the model. Partial dependence plots can be leveraged to determine the marginal effects of the chosen characteristics vs. the performance of the forecast, whereas individual conditional expectation can be employed to obtain a granular explanation of how a specific feature affects particular instances and to explore variation in impacts throughout instances. For example, a partial dependency plot can elucidate the role of clinical symptoms in reacting favourably to a particular treatment strategy, as observed by a computer-aided diagnosis system. On the other hand, individual conditional expectation can reveal variability in the treatment response among subgroups of patients.

1.4.4. Model-agnostic local XAI

For this type of XAI approaches, the aim is to produce model-agnostic explanations for a particular instance or the vicinity of a particular instance. Local Interpretable Model-Agnostic Explanation (LIME) [5], a well-validated tool, can provide an explanation for a complex deep learning model in the neighbourhood of an instance. Consider a deep learning algorithm that classifies a physiological attribute as a high-risk factor for certain diseases or cause of death, for which the clinician requires a post-hoc clarification. The interpretable modules are perturbed to determine how the predictions made by the change of those physiological attributes. For this perturbed dataset, a linear model is learnt with higher weights given to the perturbed instances in the vicinity of the physiological attribute. The most important components of the linear model can indicate the influence of a particular physiological attribute that can suggest a high-risk factor or the contrary can be implied. This can provide comprehensible means for the clinicians to interpret the classifier.

2. Related studies in AI for healthcare and XAI for healthcare

2.1. AI in healthcare

AI attempts to emulate the neural processes of humans, and it introduces a paradigm change to healthcare, driven by growing healthcare data access and rapid development in analytical techniques. We survey briefly the present state of healthcare AI applications and explore their prospects. For a detailed up to date review, the readers can refer to Jiang et al. [6], Panch et al. [7], and Yu et al. [8] on general AI techniques for healthcare and Shen et al. [9], Litjens et al. [10] and Ker et al. [11] on medical image analysis.

In the medical literature, the effects of AI have been widely debated [12–14]. Sophisticated algorithms can be developed using AI to 'read' features from a vast amount of healthcare data and then use the knowledge learnt to help clinical practice. To increase its accuracy based on feedback, AI can also be fitted with learning and self-correcting capabilities. By presenting up-to-date medical knowledge from journals, manuals and professional procedures to advise effective patient care, an AI-powered device [15] will support clinical decision making. Besides, in human clinical practice, an AI system may help to reduce medical and therapeutic mistakes that are unavoidable (i.e., more objective and reproducible) [12,13,15–19]. In addition, to help render real-time inferences for health risk warning and health outcome estimation, an AI system can handle valuable knowledge collected from a large patient population [20].

As AI has recently re-emerged into the scientific and public consciousness, AI in healthcare has new breakthroughs and clinical environments are imbued with novel AI-powered technologies at a breakneck pace. Nevertheless, healthcare was described as one of the most exciting application fields for AI. Researchers have suggested and built several systems for clinical decision support since the mid-twentieth century [21,22]. Since the 1970s, rule-based methods had many achievements and have been seen to interpret ECGs [23], identify diseases [24], choose optimal therapies [25], offer scientific logic explanations [26] and assist doctors in developing diagnostic hypotheses

and theories in challenging cases of patients [27]. Rule-based systems, however, are expensive to develop and can be unstable, since they require clear expressions of decision rules and, like any textbook, require human-authored modifications. Besides, higher-order interactions between various pieces of information written by different specialists are difficult to encode and the efficiency of the structures is constrained by the comprehensiveness of prior medical knowledge [28]. To narrow down the appropriate psychological context, prioritise medical theories, and prescribe treatment, it was also difficult to incorporate a method that combines deterministic and probabilistic reasoning procedures [29,30].

Recent AI research has leveraged machine learning approaches, which can account for complicated interactions [31], to recognise patterns from the clinical results, in comparison to the first generation of AI programmes, which focused only on the curation of medical information by experts and the formulation of rigorous decision laws. The machine learning algorithm learns to create the correct output for a given input in new instances by evaluating the patterns extracted from all the labelled input–output pairs [32]. Supervised machine learning algorithms are programmed to determine the optimal parameters in the models in order to minimise the differences between their training case predictions and the effects observed in these cases, with the hope that the correlations found are generalisable to cases not included in the dataset of training. The model generalisability can be then calculated using the test dataset. For supervised machine learning models, grouping, regression and characterisation of the similarity between instances with similar outcome labels are among the most commonly used tasks. For the unlabelled dataset, unsupervised learning infers the underlying patterns for discovering sub-clusters of the original dataset, for detecting outliers in the data, or for generating low-dimensional data representations. However, it is of note that in a supervised manner, the recognition of low-dimensional representations for labelled dataset may be done more effectively. Machine-learning approaches allow the development of AI applications that promote the exploration of previously unrecognised data patterns without the need to define decision-making rules for each particular task or to account for complicated interactions between input features. Machine learning has therefore been the preferred method for developing AI utilities [31,33,34].

The recent rebirth of AI has primarily been motivated by the active implementation of deep learning – which includes training a multi-layer artificial neural network (i.e., a deep neural network) on massive datasets – to wide sources of labelled data [35]. Existing neural networks are getting deeper and typically have > 100 layers. Multi-layer neural networks may model complex interactions between input and output, but may also require more data, processing time, or advanced architecture designs to achieve better performance. Modern neural networks commonly have tens of millions to hundreds of millions of parameters and require significant computing resources to perform the model training [8]. Fortunately, recent developments in computer-processor architecture have empowered the computing resources required for deep learning [36]. However, in labelled instances, deep-learning algorithms are incredibly 'data hungry.' Huge repositories of medical databases that can be integrated into these algorithms have only recently become readily available, due to the establishment of a range of large-scale research (in particular the Cancer Genome Atlas [37] and the UK Biobank [38]), data collection platforms (e.g., Broad Bioimage Benchmark Collection [39] and the Image Data Resources [40]) and the Health Information Technology for Economic and Clinical Health (HITECH) Act, which has promised to provide financial incentives for the use of electronic health records (EHRs) [41,42]. In general, deep learning based AI algorithms have been developed for image-based classification [43], diagnosis [44–46] and prognosis [47,48], genome interpretation [49], biomarker discovery [50,51], monitoring by wearable life-logging devices [52], and automated robotic surgery [53] to enhance the digital healthcare [54].

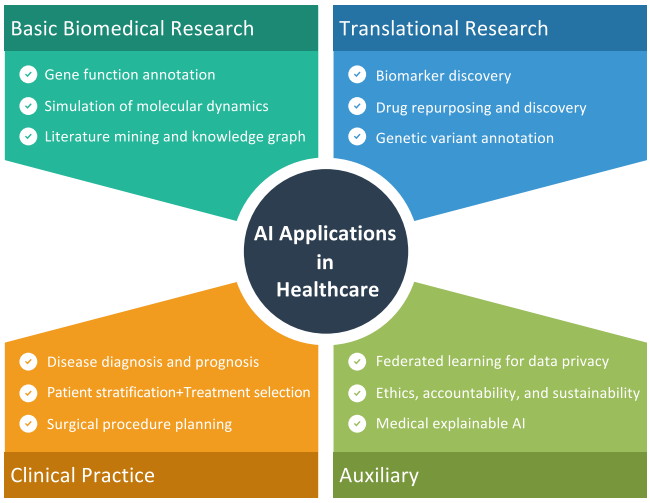


Fig. 6. A non-exhaustive map of the AI in healthcare applications.

The rapid explosion of AI has given rise to the possibilities of using aggregated health data to generate powerful models that can automate diagnosis and also allow an increasingly precise approach to medicine by tailoring therapies and targeting services with optimal efficacy in a timely and dynamic manner. A non-exhaustive map of possible applications is showing in Fig. 6.

While AI is promising to revolutionise medical practice, several technological obstacles lie ahead. Because deep learning based approaches rely heavily on the availability of vast volumes of high-quality training data, caution must be taken to collect data that is representative of the target patient population. For example, data from various healthcare settings, which include different forms of bias and noise, may cause a model trained in the data of one hospital to fail to generalise to another [55]. Where the diagnostic role has an incomplete inter-expert agreement, it has been shown that consensus diagnostics could greatly boost the efficiency of the training of the deep learning based models [56]. In order to manage heterogeneous data, adequate data curation is important. However, achieving a good quality gold standard for identifying the clinical status of the patients requires physicians to review their clinical results independently and maybe repeatedly, which is prohibitively costly at a population scale. A silver standard [57] that used natural-language processing methods and diagnostic codes to determine the true status of patients has recently been proposed [58]. Sophisticated algorithms that can handle the idiosyncrasies and noises of different datasets can improve the efficiency and safety of prediction models in life-and-death decisions.

Most of the recent advancement in neural networks has been limited to well-defined activities that do not require data integration across several modalities. Approaches for the application of deep neural networks to general diagnostics (such as analysis of signs and symptoms, prior medical history, laboratory findings and clinical course) and treatment planning are less simple. While deep learning has been effective in image detection [59], translation [60], speech recognition [61,62], sound synthesis [63] and even automated neural architecture search [64], clinical diagnosis and treatment tasks often need more care (e.g., patient interests, beliefs, social support and medical history) than the limited tasks that deep learning can be normally adept. Moreover, it is unknown if transfer learning approaches will be able to translate models learnt from broad non-medical datasets into algorithms for the study of multi-modality clinical datasets. This suggests that more comprehensive data-collection and data-annotation activities are needed to build end-to-end clinical AI programmes. It is of note that one most recent study advocated for the use of Graph Neural Networks as a tool of choice for multi-modal causability knowledge fusion [65].

Table 1
Summary of various XAI methods in digital healthcare and medicine including their category (XAI via dimension reduction, feature importance, attention mechanism, knowledge distillation, and surrogate representations), reference, key idea, type (Intrinsic or Post-hoc, Local or Global, and Model-specific or Model-agnostic) and specific clinical applications.

XAI Category	Reference	Method	Intrinsic/ Post-hoc	Local/Global	Model-specific/ Model-agnostic	Application
Dimension Reduction	Zhang et al. [66]	Optimal feature selection	Intrinsic	Global	Model-specific	Drug side effect estimation
	Yang et al. [67]	Laplacian Eigenmaps	Intrinsic	Global	Model-specific	Brain tumour classification using MRS
	Zhao and Bolouri [68]	Cluster analysis and LASSO	Intrinsic	Global	Model-agnostic	Lung cancer patients stratification
	Kim et al. [69]	Optimal feature selection	Intrinsic	Global	Model-agnostic	Cell-type specific enhancers prediction
	Hao et al. [70]	Sparse deep learning	Intrinsic	Global	Model-agnostic	Long-term survival prediction for glioblastoma multiforme
	Bernardini et al. [71]	Sparse-balanced SVM	Intrinsic	Global	Model-agnostic	Early diagnosis of type 2 diabetes
Feature Importance	Eck et al. [72]	Feature marginalisation	Post-hoc	Global, Local	Model-agnostic	Gut and skin microbiota/inflammatory bowel diseases diagnosis
	Ge et al. [73]	Feature weighting	Post-hoc	Global	Model-agnostic	ICU mortality prediction (all-cause)
	Zuallaert et al. [74]	DeepLIFT	Post-hoc	Global	Model-agnostic	Splice site detection
	Suh et al. [75]	Shapley value	Post-hoc	Global, Local	Model-agnostic	Decision-supporting for prostate cancer
	Singh et al. [76]	DeepLIFT and others	Post-hoc	Global, Local	Model-agnostic	Ophthalmic diagnosis
Attention Mechanism	Kwon et al. [77]	Attention	Intrinsic	Global, Local	Model-specific	Clinical risk prediction (cardiac failure/cataract)
	Zhang et al. [78]	Attention	Intrinsic	Local	Model-specific	EHR based future hospitalisation prediction
	Choi et al. [79]	Attention	Intrinsic	Local	Model-specific	Heart failure prediction
	Kaji et al. [80]	Attention	Intrinsic	Global, Local	Model-specific	Predictions of clinical events in ICU
	Shickel et al. [81]	Attention	Intrinsic	Global, Local	Model-specific	Sequential organ failure assessment/in-hospital mortality
	Hu et al. [82]	Attention	Intrinsic	Local	Model-specific	Prediction of HIV genome integration site
	Izadyyazdanabadi et al. [83]	MLCAM	Intrinsic	Local	Model-specific	Brain tumour localisation
	Zhao et al. [84]	Respond-CAM	Intrinsic	Local	Model-specific	Macromolecular complexes
	Couture et al. [85]	Super-pixel maps	Intrinsic	Local	Model-specific	Histologic tumour subtype classification
	Lee et al. [86]	CAM	Intrinsic	Local	Model-specific	Acute intracranial haemorrhage detection
	Kim et al. [87]	CAM	Intrinsic	Local	Model-specific	Breast neoplasm ultrasonography analysis
	Rajpurkar et al. [88]	Grad-CAM	Intrinsic	Local	Model-specific	Diagnosis of appendicitis
	Porumb et al. [89]	Grad-CAM	Intrinsic	Local	Model-specific	ECG based hypoglycaemia detection
	Hu et al. [43]	Multiscale CAM	Intrinsic	Local	Model-specific	COVID-19 classification
Knowledge Distillation	Caruana et al. [90]	Rule-based system	Intrinsic	Global	Model-specific	Prediction of pneumonia risk and 30-day readmission forecast
	Letham et al. [91]	Bayesian rule lists	Intrinsic	Global	Model-specific	Stroke prediction
	Che et al. [92]	Mimic learning	Post-hoc	Global, Local	Model-specific	ICU outcome prediction (acute lung injury)
	Ming et al. [93]	Visualisation of rules	Post-hoc	Global	Model-specific	Clinical diagnosis and classification (breast cancer, diabetes)
	Xiao et al. [94]	Complex relationships distilling	Post-hoc	Global	Model-specific	Prediction of the heart failure caused hospital readmission
	Davoodi and Moradi [95]	Fuzzy rules	Intrinsic	Global	Model-specific	In-hospital mortality prediction (all-cause)
	Lee et al. [96]	Visual/textual justification	Post-hoc	Global, Local	Model-specific	Breast mass classification
	Prentzas et al. [97]	Decision rules	Intrinsic	Global	Model-specific	Stroke Prediction
Surrogate Models	Pan et al. [98]	LIME	Post-hoc	Local	Model-agnostic	Forecast of central precocious puberty
	Ghafouri-Fard et al. [99]	LIME	Post-hoc	Local	Model-agnostic	Autism spectrum disorder diagnosis
	Kovalev et al. [100]	LIME	Post-hoc	Local	Model-agnostic	Survival models construction
	Meldo et al. [101]	LIME	Post-hoc	Local	Model-agnostic	Lung lesion segmentation
	Panigutti et al. [102]	LIME like with rule-based XAI	Post-hoc	Local	Model-agnostic	Prediction of patient readmission, diagnosis and medications
	Lauritsen et al. [103]	Layer-wise relevance propagation	Post-hoc	Local	Model-agnostic	Prediction of acute critical illness from EHR

The design of a computing system for the processing, storage and exchange of EHRs and other critical health data remains a problem [104]. Privacy-preserving approaches, e.g., via federated learning, can allow safe sharing of data or models across cloud providers [105]. However, the creation of interoperable systems that follow the requirement for the representation of clinical knowledge is important for the broad adoption of such technology [106]. Deep and seamless incorporation of data across healthcare applications and locations remains questionable and can be inefficient. However, new software interfaces for clinical data are starting to show substantial adoption through several EHR providers, such as Substitutable Medical Applications and Reusable Technologies on the Fast Health Interoperability Resources platform [107,108]. Most of the previously developed AI in healthcare applications were conducted on retrospective data for the proof of concept [109]. Prospective research and clinical trials to assess the efficiency of the developed AI systems in clinical environments are necessary to verify the real-world usefulness of these medical AI systems [110]. Prospective studies will help recognise the fragility of the AI models in real-world heterogeneous and noisy clinical settings and identify approaches to incorporate medical AI for existing clinical workflows.

AI in medicine would eventually result in safety, legal and ethical challenges [111] with respect to medical negligence attributed to complicated decision-making support structures, and have to face the regulation hurdles [112]. If malpractice lawsuits involving medical AI applications occur, the judicial system will continue to provide specific instructions as to which agency is responsible. Health providers with malpractice insurance have to be clear on coverage as health care decisions are taken in part by the AI scheme [8]. With the deployment of automatic AI for particular clinical activities, the criteria for diagnostic, surgical, supporting and paramedical tasks will need to be revised and the functions of healthcare practitioners will begin to change as different AI modules are implemented into the quality of treatment, and the bias needs to be minimised while the patient satisfaction must be maximised [113,114].

2.2. XAI in healthcare

Despite deep learning based AI technologies will usher in a new era of digital healthcare, challenges exist. XAI can play a crucial role, as an auxiliary development (Fig. 6), for potentially solving the small sample learning by filter out clinically meaningless features. Moreover, many high-performance deep learning models produce findings that are impossible for unaided humans to understand. While these models can produce better-than-human efficiency, it is not easy to express intuitive interpretations that can justify model findings, define model uncertainties, or derive additional clinical insights from these computational 'black-boxes.' With potentially millions of parameters in the deep learning model, it can be tricky to understand what the model sees in the clinical data, e.g., radiological images [115]. For example, research investigation has explicitly stated that being a black box is a "strong limitation" for AI in dermatology since it is not capable of doing a personalised evaluation by a qualified dermatologist that can be used to clarify clinical facts [116]. This black-box design poses an obstacle for the validation of the developed AI algorithms. It is necessary to demonstrate that a high-performance deep learning model actually identifies the appropriate area of the image and does not over-emphasise unimportant findings. Recent approaches have been developed to describe AI models including the visualisation methods. Some widely used levers include occlusion maps [117], salience maps [118], class activation maps [119], and attention maps [120]. Localisation and segmentation algorithms can be more readily interpreted since the output is an image. Model understanding, however, remains much more difficult for deep neural network models trained on non-imaging data other than images that is a current open question for ongoing research efforts [5].

Deep learning-based AI methods have gained popularity in the medical field, with a wide range of work in automatic triage, diagnosis, prognosis, treatment planning and patient management [6]. We can find many open questions in the medical field that have galvanised clinical trials leveraging deep learning and AI approaches (e.g., from grand-challenge.org). Nevertheless, in the medical field, the issue of interpretability is far from theoretical development. More precisely, it is noted that interpretabilities in the clinical sectors include considerations not recognised in other areas, including risk and responsibilities [7,121]. Life may be at risk as medical responses are made, and leaving those crucial decisions to AI algorithms that without explainabilities and accountabilities will be irresponsible [122]. Apart from legal concerns, this is a serious vulnerability that could become disastrous if used with malicious intent.

As a result, several recent studies [120,123–127] have been devoted to the exploration of explainability in medical AI. More specifically, specific analyses have been investigated, e.g., chest radiography [128], emotion analysis in medicine [129], COVID-19 detection and classification [43], and the research encourages understanding of the importance of interpretability in the medical field [130]. Besides, the exposition argues [131] that a certain degree of opaqueness is appropriate, that is, it would be more important for us to deliver empirically checked reliable findings than to dwell too hard on how to unravel the black-box. It is advised that readers consider these studies first, at least for an overview of interpretability in medical AI.

An obvious XAI approach has been taken by many researchers is to provide their predictive models with interpretability. These methods depend primarily on maintaining the interpretability of less complicated AI models while improving their performance by techniques of refinement and optimisation. For example, as Fig. 5 shows, decision tree based methods are normally interpretable, research studies have been done using automated pruning of decision trees for various classifications of illnesses [132] and accurate decision trees focused on boosting patient stratification [133]. However, such model optimisation is not always straightforward and it is not a trivial task.

Previous survey studies on XAI in healthcare can be found elsewhere, e.g., Tjoa and Guan [134] in medical XAI and Payrovnaziri et al. [135] in XAI for EHR. For specific applications, e.g., digital pathology, the readers can refer to Pocevivciute et al. [136] and Tosun et al. [137]. The research studies in XAI and medical XAI have been increased exponentially especially after 2018 alongside increasingly development of multimodal clinical information fusion (Fig. 7). In this mini-review, we only surveyed the most recent studies that were not covered by previous more comprehensive review studies. In this mini-review, we classified XAI in medicine and healthcare into five categories, which synthesised the approach by Payrovnaziri et al. [135], including (1) XAI via dimension reduction, (2) XAI via feature importance, (3) XAI via attention mechanism, (4) XAI via knowledge distillation, and (5) XAI via surrogate representations (Table 1).

2.2.1. XAI via dimension reduction

Dimension reduction methods, e.g., using principal component analysis (PCA) [138], independent component analysis (ICA) [139], and Laplacian Eigenmaps [140] and other more advanced techniques, are commonly and conventionally used approaches to decipher AI models by representing the most important features. For example, by integrating multi-label k-nearest neighbour and genetic algorithm techniques, Zhang et al. [66] developed a model for drug side effect estimation based on the optimal dimensions of the input features. Yang et al. [67] proposed a nonlinear dimension reduction method to improve unsupervised classification of the ^{1}H MRS brain tumour data and extract the most prominent features using Laplacian Eigenmaps. Zhao and Bolouri [68] stratified stage-one lung cancer patients by defining the most insightful examples via a supervised learning scheme. In order to recognise a group of "exemplars" to construct a "dense data matrix", they introduced a hybrid method for dimension reduction by combining

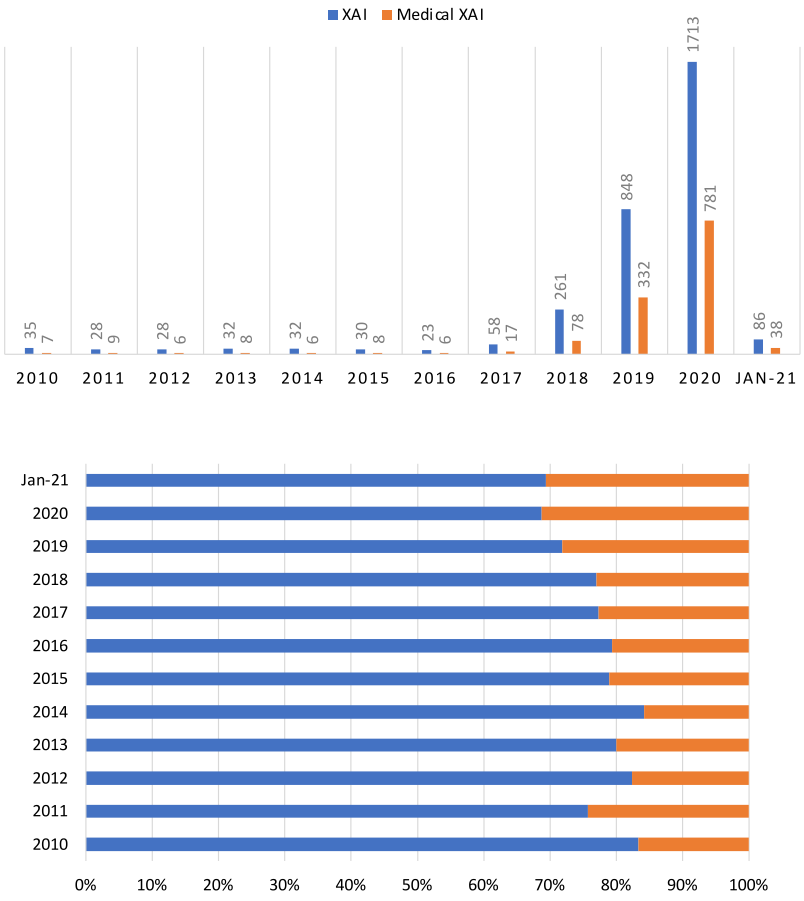


Fig. 7. Publication per year for XAI and medical XAI (top) and percentage for two categories of research (bottom). Data retrieved from Scopus® (Jan 8th, 2021) by using these commands when querying this database—XAI: (ALL("Explainable AI") OR ALL("Interpretable AI") OR ALL("Explainable Artificial Intelligence") OR ALL("Interpretable Artificial Intelligence") OR ALL("XAI")) AND PUBYEAR = 20XX; Medical XAI: (ALL("Explainable AI") OR ALL("Interpretable AI") OR ALL("Explainable Artificial Intelligence") OR ALL("Interpretable Artificial Intelligence") OR ALL("XAI")) AND (ALL("medical") OR ALL("medicine")) AND PUBYEAR = 20XX, in which XX represents the actual year.

pattern recognition with regression analytics. Then they used examples in the final model that are the most predictive for the outcome. Based on domain knowledge, Kim et al. [69] developed a deep learning method to extract and rank the most important features based on their weights in the model, and visualised the outcome for predicting cell-type-specific enhancers. To explore the gene pathways and their associations in patients with the brain tumour, Hao et al. [70] proposed a pathway-associated sparse deep learning method. Bernardini et al. [71] used the least absolute shrinkage and selection operator (LASSO) to prompt sparsity for SVMs for the early diagnosis of type 2 diabetes.

Simplifying the information down to a small subset using dimension reduction methods can make the underlying behaviour of the model understandable. Besides, with potentially more stable regularised models, they are less prone to overfitting, which may also be beneficial in general. Nevertheless, the possibility of losing crucial features, which may still be relevant for clinical predictions on a case-by-case basis, can be common and these important features may be neglected unintentionally by the dimensional reduced models.

2.2.2. XAI via feature importance

Researchers have leveraged the feature importance to explain the characteristics and significance of the extracted features and the correlations among features and between features and the outcomes for providing interpretability for AI models [2,141,142]. Ge et al. [73] used feature weights to rank the top ten extracted features to predict mortality of the intensive care unit. Suh et al. [75] developed a risk calculator model for prostate cancer (PCa) and clinically significant PCa with XAI modules that used Shapley value to determine the feature importance [143]. Sensitivity analysis of the extracted features can represent the feature importance, and essentially the more important features are those for which the output is more sensitive [144]. Eck et al. [72] defined the most significant features of a microbiota-based diagnosis task by roughly marginalising the features and testing the effect on the model performance.

Shrikumar et al. [145] implemented the Deep Learning Important FeaTures (DeepLIFT)—a backpropagation based approach to realise interpretability. Backpropagation approaches measure the output gradient for input through the backpropagation algorithm to report the significance of the feature. Zuallaert et al. [74] developed the DeepLIFT based method to create interpretable deep models for splice site prediction by measuring the contribution score for each nucleotide. A recent comparative study of different models of XAI, including DeepLIFT [145], Guided backpropagation (GBP) [146], Layer wise relevance propagation (LRP) [147], SHapley Additive exPlanations (SHAP) [148] and others, was conducted for ophthalmic diagnosis [76].

XAI, by the extraction of feature importance, can not only explain essential feature characteristics, but may also reflect their relative importance to clinical interpretation; however, numerical weights are either not easy to understand or maybe misinterpreted.

2.2.3. XAI via attention mechanism

The core concept behind the attention mechanism [149] is that the model "pays attention" only to the parts of the input where the most important information is available. It was originally proposed for tackling the relation extraction task in machine translation and other natural language processing problems. Because certain words are

more relevant than others in the relation extraction task, the attention mechanism can assess the importance of the words for the purpose of classification, generating a meaning representation vector. There are various types of attention mechanisms, including global attention, which uses all words to build the context, local attention, which depends only on a subset of words, or self-attention, in which several attention mechanisms are implemented simultaneously, attempting to discover every relation between pairs of words [150]. The attention mechanism has also been shown to contribute to the enhancement of interpretability as well as to technical advances in the field of visualisation [151].

Kaji et al. [80] demonstrated particular occasions when the input features have mostly influenced the predictions of clinical events in ICU patients using the attention mechanism. Shickel et al. [81] presented an interpretable acuity score framework using deep learning and attention-based sequential organ failure assessment that can assess the severity of patients during an ICU stay. Hu et al. [82] provided "mechanistic explanations" for the accurate prediction of HIV genome integration sites. Zhang et al. [78] also built a method to learn how to represent EHR data that could document the relationship between clinical outcomes within each patient. Choi et al. [79] implemented the Reverse Time Attention Model (RETAIN), which incorporated two sets of attention weights, one for visit level to capture the effect of each visit and the other at the variable-level. RETAIN was a reverse attention mechanism intended to maintain interpretability, to replicate the actions of clinicians, and to integrate sequential knowledge. Kwon et al. [77] proposed a visually interpretable cardiac failure and cataract risk prediction model based on RETAIN (RetainVis). The general intention of these research studies is to improve the interpretability of deep learning models by highlighting particular position(s) within a sequence (e.g., time, visits, DNA) in which those input features can affect the prediction outcome.

Class activation mapping (CAM) [152] method and its variations have been investigated for XAI since 2016, and have been subsequently used for digital healthcare, especially the medical image analysis areas. Lee et al. [86] developed an XAI algorithm for the detection of acute intracranial haemorrhage from small datasets that is one of the most famous studies using CAM. Kim et al. [87] summarised AI based breast ultrasonography analysis with CAM based XAI. Zhao et al. [84] reported a Respond-CAM method that offered a heatmap-based saliency on 3D images obtained from cryo-tomography of cellular electrons. The region where macromolecular complexes were present was marked by the high intensity in the heatmap. Izadyyazdanabadi et al. [83] developed a multilayer CAM (MLCAM), which was used for brain tumour localisation. Coupling with CNN, Couture et al. [85] proposed a multi-instance aggregation approach to classify breast tumour tissue microarray for various clinical tasks, e.g., histologic subtype classification, and the derived super-pixel maps could highlight the area where the tumour cells were and each mark corresponded to a tumour class. Rajpurkar et al. [88] used Grad-CAM for the diagnosis of appendicitis from a small dataset of CT exams using video pretraining. Porumb et al. [89] combined CNN and RNN for electrocardiogram (ECG) analysis and applied Grad-CAM for the identification of the most relevant heartbeat segments for the hypoglycaemia detection. In Hu et al. [43], a COVID-19 classification system was implemented with multiscale CAM to highlight the infected areas. By the means of visual interpretability, these saliency maps are recommended. The clinician analysts who examine the AI output can realise that the target is correctly identified by the AI model, rather than mistaking the combination of the object with the surrounding as the object itself.

Attention based XAI methods do not advise the clinical end user specifically of the response, but highlight the areas of greater concern to facilitate easier decision-making. Clinical users can, therefore, be more tolerant of imperfect precision. However, it might not be beneficial to actually offer this knowledge to a clinical end user because of the major concerns, including information overload and warning fatigue. It can potentially be much more frustrating to have areas of attention without clarification about what to do with the findings if the end user is unaware of what the rationale of a highlighted segment is, and therefore the end user can be prone to ignore non-highlighted areas that could also be critical.

2.2.4. XAI via knowledge distillation and rule extraction

Knowledge distillation is one form of the model-specific XAI, which is about eliciting knowledge from a complicated model to a simplified model—enables to train a student model, which is usually explainable, with a teacher model, which is hard to interpret. For example, this can be accomplished by model compression [153] or tree regularisation [154] or through a coupling approach of model compression and dimension reduction [141]. Research studies have investigated this kind of technique for several years, e.g., Hinton et al. [155], but has recently been uplifted along with the development of AI interpretability [156–158]. Rule extraction is another widely used XAI method that is closely associated with knowledge distillation and can have a straightforward application for digital healthcare, for example, decision sets or rule sets have been studied for interpretability [159] and Model Understanding through Subspace Explanations (MUSE) method [160] has been developed to describe the projections of the global model by considering the various subgroups of instances defined by user interesting characteristics that also produces explanation in the form of decision sets.

Che et al. [92] introduced an interpretable mimic-learning approach, which is a straightforward knowledge-distillation method that uses gradient-boosting trees to learn interpretable structures and make the baseline model understandable. The approach used the information distilled to construct an interpretable prediction model for the outcome of the ICU, e.g., death, ventilator usage, etc. A rule-based framework that could include an explainable statement of death risk estimation due to pneumonia was introduced by Caruana et al. [90]. Letham et al. [91] also proposed an XAI model named Bayesian rule lists, which offered certain stroke prediction claims. Ming et al. [93] developed a visualisation approach to derive rules by approximating a complicated model via model induction at different tasks such as diagnosis of breast cancer and the classification of diabetes. Xiao et al. [94] built a deep learning model to break the dynamic associations between readmission to hospital and possible risk factors for patients by translating EHR incidents into embedded clinical principles to characterise the general situation of the patients. Classification rules were derived as a way of providing clinicians interpretable representations of the predictive models. Davoodi and Moradi [95] developed a rule extraction based XAI technique to predict mortality in ICUs and Das et al. [161] used a similar XAI method for the diagnosis of Alzheimer's disease. In the LSTM-based breast mass classification, Lee et al. [96] incorporated the textual reasoning for interpretability. For the characterisation of stroke and risk prediction, Prentzas et al. [97] implemented the argumentation theory for their XAI algorithm training process by extracting decision rules.

XAI approaches, which rely on knowledge distillation and rule extraction, are theoretically more stable models. The summarised representations of complicated clinical data can provide clinical end-users with the interpretable results intuitively. However, if the interpretation of these XAI results could not be intuitively understood by clinical end-users, then the representations are likely to make it much harder for the end-users to comprehend.

2.2.5. XAI via surrogate representation

An effective application of XAI in the medical field is the recognition of individual health-related factors that lead to disease prediction using the local interpretable model-agnostic explanation (LIME) method [5] that offers explanations for any classifier by approximating the reference model with a surrogate interpretable and "locally

faithful" representation. LIME disrupts an instance, produces neighbourhood data, and learns linear models in the neighbourhood to produce explanations [162].

Pan et al. [98] used LIME to analyse the contribution of new instances to forecast central precocious puberty in children. Ghafouri-Fard et al. [99] have applied a similar approach to diagnose autism spectrum disorder. Kovalev et al. [100] proposed a method named SurvLIME to explain AI base survival models. Meldo et al. [101] used a local post-hoc explanation model, i.e., LIME, to select important features from a special feature representation of the segmented lung suspicious objects. Panigutti et al. [102] developed the "Doctor XAI" system that could predict the readmission, diagnosis and medications order for the patient. Similar to LIME, the implemented system trained a local surrogate model to mimic the black-box behaviour with a rule-based explanation, which can then be mined using a multi-label decision tree. Lauritsen et al. [103] tested an XAI method using Layer-wise Relevance Propagation [163] for the prediction of acute critical illness from EHR.

Surrogate representation is a widely used scheme for XAI; however, the white-box approximation must accurately describe the black-box model to gain trustworthy explanation. If the surrogate models are too complicated or too abstract, the clinician comprehension might be affected.

3. Proposed method

3.1. Problem formulation

In this study, we have demonstrated two typical but important applications of using XAI, which have been developed for classification and segmentation—two mostly widely discussed problems in medical image analysis and AI-powered digital healthcare. Our developed XAI techniques have been manifested using CT images classification for COVID-19 patients and segmentation for hydrocephalus patients using CT and MRI datasets.

3.2. XAI for classification

In this subsection, we provide a practical XAI solution for explainable COVID-19 classification that is capable of alleviating the domain shift problem caused by multicentre data collected for distinguishing COVID-19 patients from other lung diseases using CT images. The main challenge for multicentre data is that hospitals are likely to use different scanning protocols and parameters for CT scanners when collecting data from patients leading to distinct data distribution. Moreover, it can be observed that images obtained from various hospitals are visually different although they are imaging the same organ. If a machine learning model is trained on data from one hospital and tested on the data from another hospital (i.e., another centre), the performance of the model often degrades drastically. Besides, another challenge is that only patient-level annotations are available commonly but image-level labels are not since it would take a large amount of time for radiologists to annotate them [164]. Therefore, we propose a weakly supervised learning based classification model to cope with these two problems. Besides, an explainable diagnosis module in the proposed model can also offer the auxiliary diagnostic information visually for radiologists. The overview of our proposed model is illustrated in Fig. 8.

3.2.1. Explainable diagnosis module (EDM)

As the predicting process of deep learning models is in a black-box, it is desirable to develop an explainable technique in medical image diagnosis, which provides an explainable auxiliary tool for radiologists. For common practice, CAM can generate the localisation maps for the prediction through the weighted sum of feature maps from the backbone networks such as ResNet [165]. Suppose $F^k \in \mathbb{R}^{H' \times W'}$ is the kth feature map with the shape of $H' \times W'$, and $W^{fc} \in \mathbb{R}^{K \times C}$, where K is the number of feature maps. Therefore, the class score for class c can be computed as

$$s_c = \sum_{k=1}^{K} W_{k,c}^{fc} \left(\frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} F_{i,j}^{k} \right). \tag{1}$$

Therefore, the activation map A_c^{fc} for class c can be defined by

$$(A_c^{fc})_{i,j} = \sum_{k=1}^{K} W_{k,c}^{fc} F_{i,j}^{k}. \tag{2}$$

However, generating CAMs is not an end-to-end process, in which the network should be firstly trained on the dataset and utilises the weights of the last fully connected layer to compute the CAMs, bringing extra computation. To tackle this drawback, in our explainable diagnosis module (EDM), we replace the fully connected layer with a 1×1 convolutional layer of which the weight W^{conv} shares the same mathematical form as W^{fc}. So we can reformulate Eq. (1) as

$$s_c = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \left(\sum_{k=1}^{K} W_{k,c}^{conv} F_{i,j}^{k} \right) = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} (A_c^{conv})_{i,j}, \tag{3}$$

where A_c^{conv} is the activation map for class c that can be learnt adaptively during the training procedure. The activation map produced by the EDM can not only accurately indicate the importance of the region from CT images and locate the infected parts of the patients, but can also offer the explainable results which are able to account for the prediction.

3.2.2. Slice integration module (SIM)

Intuitively, each COVID-19 patient case has a different severity. Some patients are severely infected with large lesions, while most of the positive cases can be mild of which only a small portion of the CT volume is infected. Therefore, if we directly apply the patient level annotations as the labels for the image slices, the data would be extremely noisy leading to poor performance as the consequence. To overcome this problem, instead of relying on single images, we propose a slice integration module (SIM) and use the joint distribution of the image slices to model the probability of the patient being infected or not. In our SIM, we assume that the lesions are consecutive and the distribution of the lesion positions is consistent. Therefore, we adopt a section based strategy to handle this problem and fit this into a Multiple Instance Learning (MIL) framework [166]. In the MIL, each sample is regarded as a bag, which is composed of a set of instances. A positive bag contains at least one positive instance, while the negative bag solely consists of negative instances. In our scenario, only patient annotations (bag labels) are provided, and the sections can be regarded as instances in the bags.

Given a patient $\mathcal{P} = [\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n]$ with n CT slices, we divide them into disjoint sections $\mathcal{P} = \{S_i\}_{i=1}^{|S|}$, where $|S|$ is the total amount of sections for patient $\mathcal{P}$, that is

$$|S| = \max\left(1, \left\lfloor \frac{n}{l_s} \right\rfloor\right). \tag{4}$$

Here l_s is the section length, which is a designed parameter. Then we integrate the probability of each section as the probability of the patient, that is

$$P(c \mid \mathcal{P}) = P(c \mid \{S_i\}_{i=1}^{|S|}) = \frac{1}{1 + \prod_{i=1}^{|S|} (\frac{1}{P(c \mid S_i)} - 1)}, \tag{5}$$

where $P(c \mid S_i)$ is the probability of the ith section S_i that belongs to class c. By taking the k-max probability of the images for each class to compute the section probability, we can mitigate the problem that some

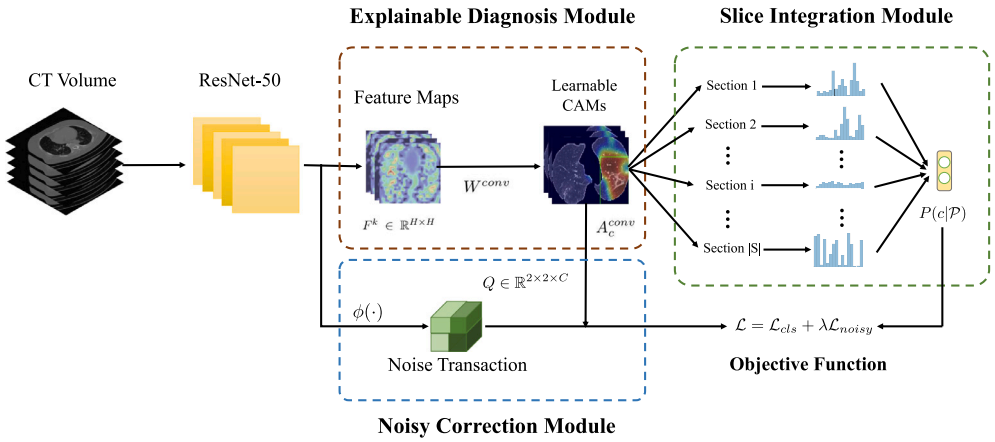


Fig. 8. The overview of our proposed model. $P(c \mid S_i)$ denotes the probability of the Section S_i, and $P(c \mid \mathcal{P})$ represents the probability of the patient who is COVID-19 infected or not. $Q \in \mathbb{R}^{2\times 2\times C}$ indicates the noise transaction from the probability of the true label $P(y_c \mid \mathcal{I})$ to the noise label $P(z_c \mid \mathcal{I})$. Besides, $\phi(\cdot)$ is a non-linear feature transformation function, which projects the feature into embedding space.

slices may contain few infections, which can hinder the prediction for the section. The k-max selection method can be formulated as

$$P(c \mid S_i) = \sigma\left(\frac{1}{k}\max_{s^{(j)}\in M}\sum_{j=1}^{k} s_c^{(j)}\right), \quad s.t. \quad M \subset S_i, |M| = k. \tag{6}$$

where $s_c^{(j)}$ is the top jth class score of the slice in the ith section for the class c, and $\sigma(\cdot)$ represents the Sigmoid function. Then we apply the patient annotations $\mathbf{y}$ to compute the classification loss, which can be formulated as

$$\mathcal{L}_{cls} = -\sum_{c=0}^{1}\left[y_c \log P(c \mid \mathcal{P}) + (1 - y_c)\log(1 - P(c \mid \mathcal{P}))\right]. \tag{7}$$

3.2.3. Noisy correction module (NCM)

In real-world applications, radiologists would only diagnose the disease from one image. Therefore, it is also significant for improving the prediction accuracy on single images. However, the image-level labels are extremely noisy since only patient-level annotations are available. To further alleviate the negative impact of patient-level annotations, we propose a noisy correction module (NCM). Inspired by [167], we model the noise transaction distribution $P(z_c = i \mid y_c = j, \mathcal{I})$, which transforms the true posterior distribution $P(y_c \mid \mathcal{I})$ to the noisy label distribution $P(z_c \mid \mathcal{I})$ by

$$P(z_c = i \mid \mathcal{I}) = \sum_j P(z_c = i \mid y_c = j, \mathcal{I}) P(y_c = j \mid \mathcal{I}). \tag{8}$$

In practice, we estimate the noise transaction distribution $Q_{ij}^c = P(z_c = i \mid y_c = j, \mathcal{I})$ for the class c via

$$Q_{ij}^c = P(z_c = i \mid y_c = j, \mathcal{I}) = \frac{\exp(w_{ij}^c \phi(\mathcal{I}) + b_{ij}^c)}{\sum_i \exp(w_{ij}^c \phi(\mathcal{I}) + b_{ij}^c)}, \tag{9}$$

where $i, j \in \{0, 1\}$; $\phi(\cdot)$ is a nonlinear mapping function implemented by convolution layers; w_{ij}^c and b_{ij}^c are the trainable parameters. The noise transaction score $T_{ij}^c = w_{ij}^c \phi(\mathcal{I}) + b_{ij}^c$ represents the confidence score of the transaction from the true label i to the noise label j for the class c. Therefore, Eq. (8) can be reformulated as

$$P(z_c = i \mid \mathcal{I}) = \sum_j Q_{ij}^c P(y_c = j \mid \mathcal{I}). \tag{10}$$

By estimating the noisy label distribution $P(z_c \mid \mathcal{I})$ for patient $\mathcal{P}$, the noisy classification loss can be computed by

$$\mathcal{L}_{noisy} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=0}^{1}[y_c^n \log P(z_c = 1 \mid \mathcal{I}_n) + (1 - y_c^n)\log P(z_c = 0 \mid \mathcal{I}_n)]. \tag{11}$$

By combining Eqs. (7) and (11), we can obtain the total loss function for our XAI solution of an explainable COVID-19 classification, that is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{noisy}, \tag{12}$$

where λ is a hyper-parameter to balance the classification loss $\mathcal{L}_{cls}$ and the noisy classification loss $\mathcal{L}_{noisy}$.

3.3. XAI for segmentation

In this subsection, we introduce an XAI model that is applicable for the explainable brain ventricle segmentation using multimodal MRI data acquired from the hydrocephalus patients. Previous methods [168,169] have conducted experiments using images with a slice thickness of less than 3 mm. This is because the smaller of the image thickness, the more images could be obtained, which helps improve the representation power of the model. However, in a real-world scenario, it is not practical for clinicians to use these models because labelling these image slices is extremely labour-intensive and time-consuming. Therefore, it is more common for the annotations of images with larger slice thicknesses, which are easily available while those images with smaller slice thickness are not. Besides, models trained only on thick-slice images have poor generalisation on thin-slice images. To alleviate these problems, we proposed a thickness agnostic image segmentation model, which can be applicable for both thick-slice and thin-slice images, but only requires the annotations of thick-slice images during the training procedure.

Suppose we have a set of thick-slice images $\mathcal{D}_S = \{(x_s, y_s) | x_s \in \mathbb{R}^{H\times W\times 3}, y_s \in \mathbb{R}^{H\times W}\}$ and a set of thin-slice images $\mathcal{D}_\mathcal{T} = \{x_t | x_t \in \mathbb{R}^{H\times W\times 3}\}$. The main idea of our model is to utilise the unlabelled thin-slice images $\mathcal{D}_\mathcal{T}$ to minimise the model performance gap between thick-slice and thin-slice images while a post-hoc XAI can also be developed.

3.3.1. Segmentation network

With the wide applications of deep learning methods, the encoder–decoder based architectures are usually adopted in automated high accuracy medical image segmentation. The workflow of our proposed segmentation network is illustrated in Fig. 9. Inspired by the U-Net [170] model, we replace the original encoder with ResNet-50 [165] pre-trained on ImageNet dataset [171] since it can provide better feature representation for the input images. In addition, the decoder of the U-Net has at least a couple of drawbacks: 1) the increase of low-resolution feature maps can bring a large amount of computational complexity, and 2) interpolation methods [172] such as bilinear interpolation and

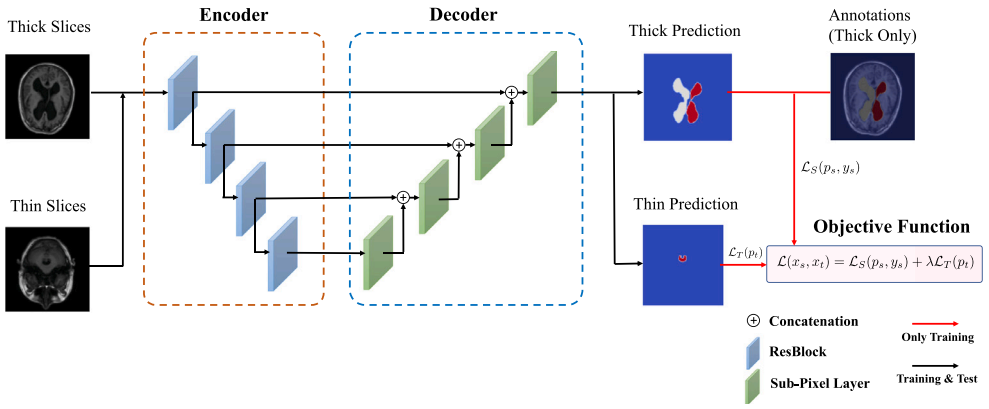


Fig. 9. Overview of our proposed XAI model for explainable segmentation. Here ResBlock represents the residual block proposed in the ResNet [165].

bicubic interpolation do not bring extra information to improve the segmentation. Instead, the decoder of our model adopts sub-pixel convolution for constructing segmentation results. The sub-pixel convolution can be represented as

$$F^L = SP(W_L * F^{L-1} + b_L), \tag{13}$$

where $SP(\cdot)$ operator transforms and arranges a tensor shaped in $H \times W \times C \times r^2$ into a tensor with the shape of $rH \times rW \times C$, and r is the scaling factor. F^{L-1} and F^L are the input feature maps and output feature maps. W_L and b_L are the parameters of the sub-pixel convolution operators for the layer L.

3.3.2. Multimodal training

As aforementioned, the thick-slice images with annotations are available. Therefore, in order to minimise the performance gap between thick-slice images and thin-slice images. We apply a multimodal training procedure to jointly optimise for both types of images. Overall, the objective function of our proposed multimodal training can be computed as

$$\mathcal{L}(x_t, x_s) = \mathcal{L}_S(p_s, y_s) + \beta \mathcal{L}_\mathcal{T}(p_t), \tag{14}$$

where β is a hyper-parameter for weighting the impact of $\mathcal{L}_S$ and $\mathcal{L}_\mathcal{T}$. p_s and p_t are the prediction of the segmentation probability maps shaped in $H \times W \times C$ for thick-slice images and thin-slices images, respectively. In particular, $\mathcal{L}_S$ is the cross-entropy loss defined as follows

$$\mathcal{L}_S(p_s, y_s) = -\frac{1}{HWC} \sum_{n=1}^{HW} \sum_{c=1}^{C} y_s^{n,c} \log p_s^{n,c}. \tag{15}$$

For the unlabelled thin-slice images, we assume that $\mathcal{L}_\mathcal{T}$ can push the features away from the decision boundary of the feature distributions of the thick-slice images, thus achieving distribution alignment. Besides, according to [173], minimising the distance between the prediction distribution p and the uniform distribution $\mathcal{U} = \frac{1}{C}$ can diminish the uncertainty of the prediction. To measure the distance of these two distributions, the objective function $\mathcal{L}_\mathcal{T}$ can be modelled by the f-divergence, that is

$$\mathcal{L}_\mathcal{T}(p_t) = -\frac{1}{HWC} \sum_{n=1}^{HW} \sum_{c=1}^{C} D_f(p_t^{n,c} \| \mathcal{U}) = -\frac{1}{HWC} \sum_{n=1}^{HW} \sum_{c=1}^{C} f(C p_t^{n,c}). \tag{16}$$

Most existing methods [173,174] tend to choose $f(x) = x \log x$, which is alternatively named as KL-divergence. However, one of the main obstacle is that when adopting $f(x) = x \log x$, the gradient of $\mathcal{L}_\mathcal{T}$ would be extremely imbalanced. To be more specific, it can assign a large gradient to the easily classified samples, while assigning a small gradient to hardly classified samples. Therefore, in order to mitigate the unbalancing problem during the optimisation, we incorporate Pearson χ^2-divergence (i.e., $f(x) = x^2 - 1$) rather than using the KL-divergence for $\mathcal{L}_\mathcal{T}$, that is

$$\mathcal{L}_\mathcal{T}(p_t) = -\frac{C}{HW} \sum_{n=1}^{HW} \sum_{c=1}^{C} (p_t^{n,c})^2. \tag{17}$$

After applying the Pearson χ^2-divergence, the gradient imbalanced issue can be mitigated since the slope of the gradient is constant, which can be verified by taking the second order derivative of $\mathcal{L}_\mathcal{T}$.

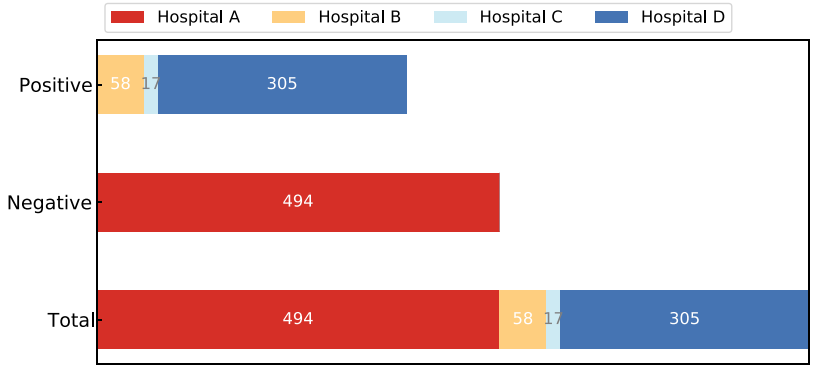
During the training procedure, $\mathcal{L}(x_t, x_s)$ is optimised alternatively for both thick-slice and thin-slice images.

3.3.3. Latent space explanation

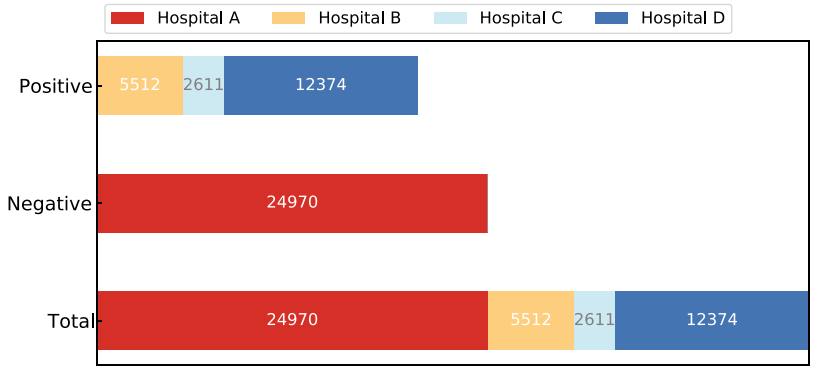
Once the model is trained using multimodal datasets, the performance of the network can be quantitatively evaluated by volumetric or regional overlapping metrics, e.g., Dice scores. However, the relation between network performance and input samples remains unclear. In order to provide information about the characteristics of data and their effect on model performance, through which users can set their expectations accordingly, we investigate the feature space and their correlation with the model performance. For feature space visualisation, we extract the outputs of the encoder module of our model, and then decompose them into a two-dimensional space via Principal Component Analysis (PCA). For estimating the whole space, we use a multi-layer perceptron to fit the decomposed samples and their corresponding Dice scores, which can provide an understanding of Dice scores for particular regions of interests in the latent space where there is no data available. Therefore, through analysing the characteristics of the samples in the latent space, we can retrieve the information about the relationships between samples and their prediction power.

3.4. Implementation details

For both our classification and segmentation tasks, we used ResNet-50 [165] as the backbone network pre-trained on ImageNet [171]. For classification, we resized these images into a spatial resolution of 224 × 224. During the training procedure, we set $\lambda = 1 \times 10^{-4}$, the dropout rate as 0.7, and the L_2 weight decay coefficient as 1×10^{-5}. Besides, l_s was set to 16, and k was set to 8 for the sake of computing the patient-level probability. For segmentation, we set $\beta = 1 \times 10^{-2}$ for balancing the impact of supervised loss and unsupervised loss. During the training, Adam [175] optimiser was utilised with a learning rate 1×10^{-3}. The training procedure is terminated after 4,000 iterations with batch size 8. All of the experiments were conducted on a workstation with 4 NVIDIA RTX GPUs using PyTorch framework with version 1.5.



(a) Patient-level Statistics



(b) Image-level Statistics

Fig. 10. Class distribution of the collected CT data. The numbers in the sub-figures (a) and (b) represent the counts for the patient-level statistics and image-level statistics, respectively. The data collected from several clinical centres can result in great challenges in learning discriminative features from those class-imbalanced centres.

4. Experimental settings and results

4.1. Showcase I: Classification for COVID-19

4.1.1. Datasets

We collected CT data from four different local hospitals in China and removed the personal information to ensure data privacy. The information of our collected data is summarised in Fig. 10. In total, there were 380 CT volumes of the patients who tested COVID-19 positive (reverse transcription polymerase chain reaction test confirmed) and 424 COVID-19 negative CT volumes. For a fair comparison, we trained the model on the cross-centre datasets collected from hospital A, B, C, and D. For an unbiased independent testing, CC-CCII data [176], a publicly available dataset, which contained 2,034 CT volumes with 130,511 images, was adopted to verify the effectiveness of the trained models.

4.1.2. Data standardisation, pre-processing and augmentation

Following the protocol described in [176], we used the U-Net segmentation network [170] to segment the CT images. Then, we randomly cropped a rectangular region whose aspect ratio was randomly sampled in $[3/4, 4/3]$, the area was randomly sampled in $[90\%, 100\%]$, and the region was then resized into 224×224. Meanwhile, we randomly flipped the input volumes horizontally with 0.5 probability. The input data would be a set of CT volumes, which were composed of consecutive CT image slices.

4.1.3. Quantitative results

We compared our proposed classification model with several state-of-the-art COVID-19 CT classification models [165,177–179]. Table 2 summarises the experimental results of COVID-19 classification on the CC-CCII data. For image-level annotations, ResNet-50 [165] and COVID-Net [177] simply treated patient-level labels as the image labels. Different from methods proposed by [165,177,178], VBNet [179] utilised the 3D residual convolutional neural network to train with patient-level annotations on the whole CT volumes rather than single slices. Besides, COVNet [178] extracted prediction scores from each slice in the CT volumes with ResNet and aggregated the prediction scores via a max-pooling operator to get the patient-level probability.

In Table 2, we can find that our method achieved the best performance among these SOTA methods. In particular, our method obtained a better performance by 7.2% on AUC compared to VB-Net [179] on the patient-level indicating that our method can be applicable for the real-world scenario. This also verified the benefit of modelling section information in the CT volumes via our proposed SIM, which we believe is also vital to the improvement of the classification performance. Besides, our method significantly outperformed other methods by at least 40% with respect to the specificity while maintaining high sensitivity, which is also a crucial indication for diagnosing COVID-19. In addition, models trained on patient-level annotations could achieve better performance compared to those trained on image-level labels. This is because the noise in the image labels could have a negative impact during the training, which might degrade the representation ability of the model. According to [180], models trained on images may rely on learning the textures of images that were highly discriminative among multiple centres. Therefore, these trained models might be overfitted and biased to the texture features of the images collected from different centres, which could explain the phenomenon that these methods (i.e., [165,177]) were poorly performed on the unseen centres.

In another aspect, for CT volumes, the sequential ordering of CT image slices is also informative. COVID-Net [178] took the most discriminative slice as the representation of the whole CT volume, which

Table 2
Comparison results of our method vs. state-of-the-art methods performed on the CC-CCII dataset.

Annotation	Method	Patient Acc. (%)	Precision (%)	Sensitivity (%)	Specificity (%)	AUC (%)
Patient-level	ResNet-50 [165]	53.44	64.45	63.03	35.71	53.24
	COVID-Net [177]	57.13	62.53	84.70	6.16	49.58
	COVNet [178]	69.96	70.20	**93.33**	26.75	81.61
	VB-Net [179]	76.11	75.84	92.73	45.38	88.34
	Ours	**89.97**	**92.99**	91.44	**87.25**	**95.53**
Image-level	ResNet-50 [165]	52.56	61.60	71.27	18.06	50.19
	COVID-Net [177]	60.03	64.81	**83.91**	15.98	58.39
	COVNet [178]	75.55	79.90	83.24	61.37	79.48
	Ours	**80.41**	**88.56**	80.15	**80.89**	**86.06**

ignored the encoding of adjacent slices. This would enforce the model only detect the most discriminative slice, leading to the bias towards positive cases, which could impede the detecting of negative cases that resulted in a low specificity. On the contrary, VBNet proposed by Ouyang et al. [179] preserved the sequential information by training on the whole CT volumes. In contrast, we partitioned the CT volume into several sections in order to preserve the sequential information to some extent. Besides, VB-Net was trained with stronger supervision that it utilised additional masks for its supervised training. For our method, we only used patient-level annotations that were much more efficient. More importantly, our method achieved better performance on both AUC and accuracy compared to VBNet [179] and COVNet [178].

In addition, we also provided the Precision–Recall (PR) and Receiver Operating Characteristic (ROC) curves to compare different methods on patient-level annotations and image-level annotations (Figs. 11 and 12). From the figure, we can observe that models trained on image-level annotations (e.g., ResNet [165] and COVID-Net [177]) were poorly performed since their AUCs were close to 50% which indicated a random guess. In contrast, models trained on patient-level was more reliable since their AUCs were greater than 50%. In particular, we found that overall our proposed method remained the best-performed algorithm with an AUC of 95.53% at the patient-level and 86.06% at the image-level. These results verified our assumption that for mild COVID-19 cases, most of the image slices are disease-free. It is of note that the AUC of the ROC or the AUC of the precision recall curve could be too general because they contain unrealistic decision thresholds, while metrics like accuracy, sensitivity or the F1 score are measured at a single threshold that reflects an individual single probability or predicted risk, rather than a range of individuals or risk. Further study using deep ROC analysis will be investigated [181].

4.1.4. Qualitative results

In order to make the prediction to be more explainable, we used the trained model to visualise the CAMs and bounding boxes generated by our EDM as described above. Fig. 13 shows the visualisation results of the derived CAMs (i.e., A^{conv}). In this figure, we can clearly observe that our method tended to pay more attention to the discriminative part of the images so as to make the predictions. For example, in the first column, the lower left part of the lung was seriously infected and had a large area of lesions. Therefore, our method would make the predictions that the image was classified as COVID-19 positive, demonstrating the capability of our XAI model to make explainable predictions.

In addition, based on the results of the derived CAMs, we also extracted the lesion bounding boxes from the CAMs. It can be found that our method was capable of yielding accurate bounding boxes from the salient part of the CAMs, as illustrated in Fig. 13, which further confirmed that our XAI method was applicable to be an auxiliary diagnosis tool for the clinicians.

To further illustrate the learnt features from our proposed method, we extracted the feature from the backbone network of our architecture, and used T-SNE [182] visualisation technique to transform the features extracted from the backbone network of our proposed model, and visualised the distribution of the classified images as shown in Fig. 14. In this figure, we can find the distinctive visual characteristics of the CT images from different hospitals (i.e., Hospital A, B, C, and D). Besides, it can be observed that the COVID-19 positive images were mostly clustered together, and negative images were mainly distributed in another cluster. More interestingly, in the cluster of the negative images, we can find several positive images in this cluster since these images were scanned from patients who were tested COVID-19 positive. In our intuition, we assume that for some mild cases, lesions were not presented in all of the CT slices. Therefore, there were indeed disease-free CT slices that could be falsely labelled as COVID-19 positive, which verified our assumption.

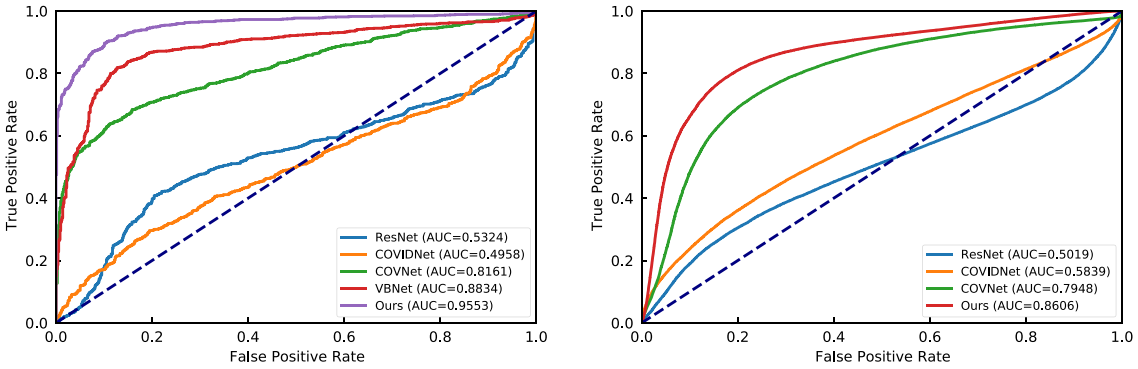
Additionally, in order to explain each individual prediction, we adopted the LIME method [5] to investigate the contribution of each pixel for the prediction. Instead of using the individual pixel, we divided an image into super-pixels, which were composed of inter-connected pixels with similar imaging patterns. Fig. 15 shows the explanations via LIME for COVID-19 positive images. In each pair of images, we visualised the super-pixels that contributed to the COVID-19 positive prediction results. We can observe that the lesion parts would explain for the positive prediction, which is reasonable to our deep learning model.

However, the LIME method could only quantitatively estimate the importance according to how close the combination of super-pixels was to the original instance. It discarded the global view of the individual feature contributed by the super-pixels. To overcome this drawback, we further leveraged Kernel SHapley Additive exPlanations (Kernel SHAP) method [183] to estimate the contribution of each super-pixel quantitatively by the SHAP value. Samples explained by the Kernel SHAP are demonstrated in Fig. 16. We can observe that the super-pixels contained lesion areas positively contributed to the positive prediction, while those super-pixels related to the backgrounds or disease-free areas would reflect the contribution to negative prediction.

4.2. Showcase II: Segmentation for hydrocephalus

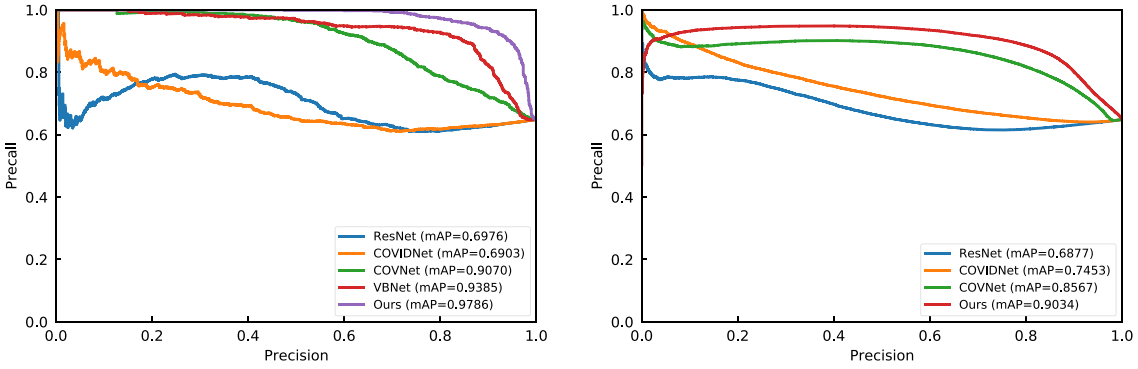
4.2.1. Datasets

The studied cohort included 20 normal elderly people, 20 patients with cerebral atrophy, 64 patients with normal pressure hydrocephalus, and 51 patients with acquired hydrocephalus (caused by subarachnoid haemorrhage, brain trauma or brain tumour). CT scans of the head were performed using two CT instruments, one of which was the SOMATOM Definition Flash from Siemens, Germany, and the other was the SOMATOM Emotion 16 from Siemens, Germany. Secondly, MRI examinations were conducted using a 1.5T MR scanner(Avanto, Siemens, Erlangen, Germany) and a 3.0T MRI scanner(Prisma, Siemens, Erlangen, Germany). The slice thickness of the CT images includes: 0.5 mm, 1.0 mm, 1.5 mm, 2.0 mm, 4.8 mm, 5.0 mm. The slice thickness of the MRI images includes: 1.0 mm, 7.8 mm, 8.0 mm. For experiments, we randomly split the thick-slice and thin-slice images into training, validation and testing sets. The details of the dataset are summarised in Table 3.



(a) Patient-level Annotation (b) Image-level Annotation

Fig. 11. The Receiver Operating Characteristic (ROC) curves of different compared methods.



(a) Patient-level Annotation (b) Image-level Annotation

Fig. 12. The Precision–Recall (PR) curves of different compared methods.

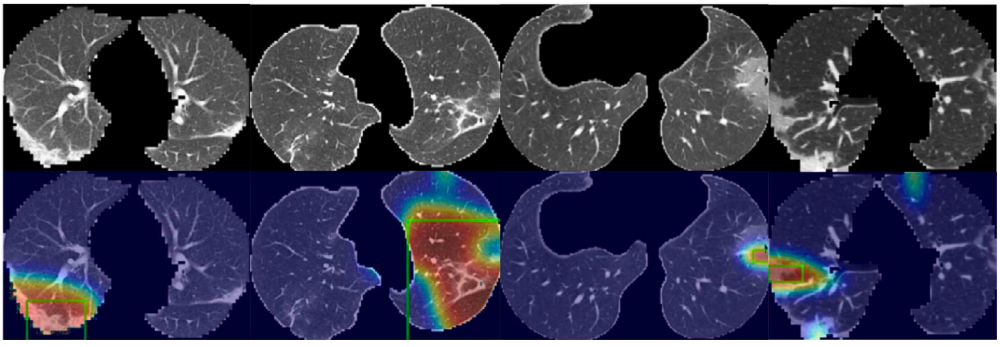


Fig. 13. Examples of the CAMs A^{conv} generated by our proposed EDM for classifying COVID-19 positive patients. The first row contains the original CT-scan image slices, and the second row illustrates the heatmaps of CAMs A^{conv} with bounding boxes confined to the infected areas.

Table 3
The number of thick-slice and thin-slice images used in our study.

Modality	Training set		Validation set		Test set	
	Thick-slice	Thin-slice	Thick-slice	Thin-slice	Thick-slice	Thin-slice
MRI	810	1,303	203	326	189	982
CT	2,088	2,076	523	519	309	492

4.2.2. Data standardisation, pre-processing and augmentation

For the pre-processing of these data, we normalised images using the z-score normalisation scheme, which was done by subtracting its mean then divided by its standard deviation. For anomaly pixels, we clipped them within the range of 1-quantile and 99-quantile. For data augmentation, we resized the images using a bicubic interpolation method and resized masks with the nearest interpolation. Then we flipped the images horizontally with 0.5 probability, and scaled the hue, saturation, and brightness with coefficients uniformly drawn from [0.8, 1.2].

4.2.3. Quantitative results

Table 4 shows the segmentation performance of various compared models on different modalities. All of the models were trained on the thick-slice images with annotations and the unlabelled thin-slice images. We can observe that our proposed method outperformed all of the compared state-of-the-art methods by a large margin on the mixed datasets (i.e., the mixture of thick-slice and thin-slice images) with at least 4.4% of the Dice scores. It is of note that all three models achieved

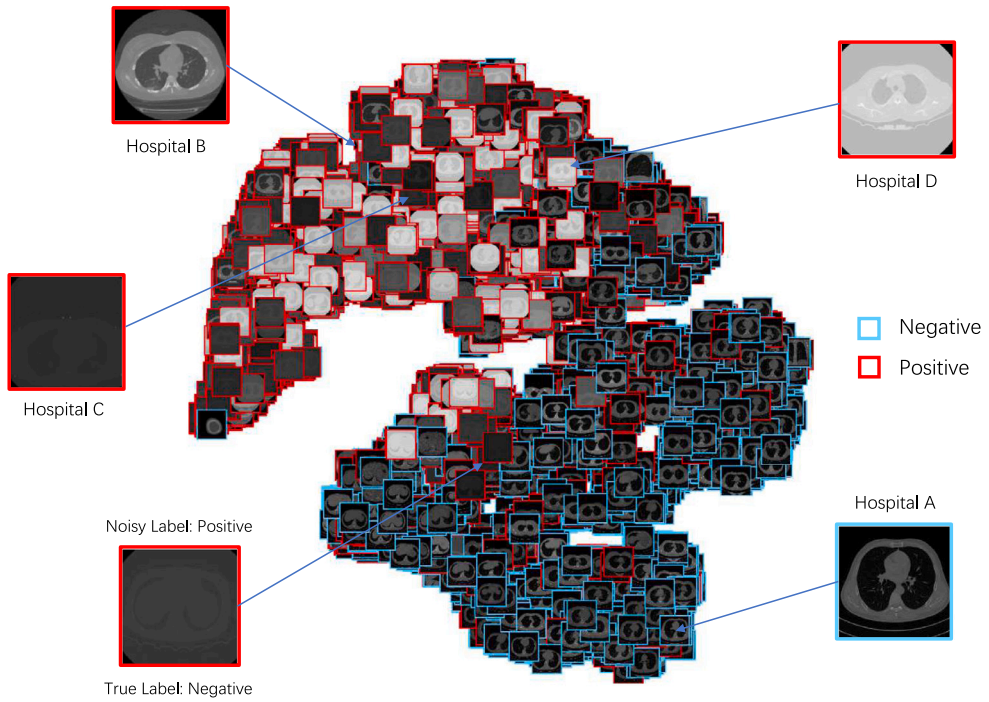


Fig. 14. T-SNE visualisation [182] of the learnt features from CT images. Original images are sampled from four different hospitals and represented in the figure. Besides, a falsely annotated image is drawn from the negative cluster.

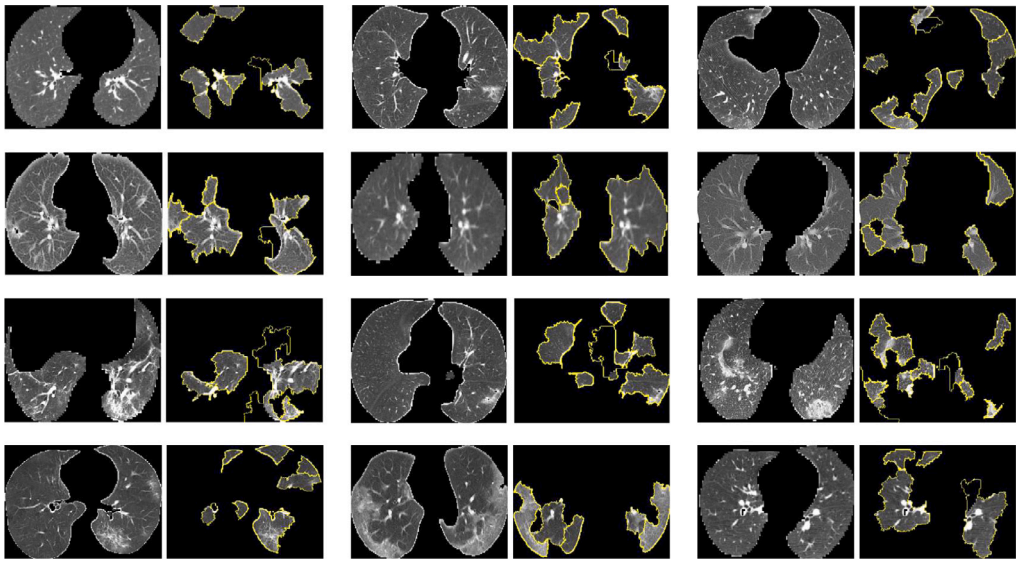


Fig. 15. Visualisation of the super-pixels that are positively contributed to the predictions via the LIME method [5].

similar segmentation performance on thick-slice images. However, our proposed method gained a significant improvement on the thin-slice images for both MRI and CT scans. The primary reason is that our model could diminish the uncertainty of these thin-slice images while achieving the distribution alignment between thick-slice and thin-slice images, which could enhance the representation and generalisation capabilities of our model. Besides, we also investigated the effectiveness of $\mathcal{L}_S$ and $\mathcal{L}_T$ in Table 5. In the table, we can find that when only training on thick-slice images, the model performed perfectly on thick-slice images while performing poorly on thin-slice images, since the distribution of these two kinds of slices could vary. Moreover, the performance of the models trained only on unlabelled thin-slice images degraded sharply because of the lack of annotations to guide the segmentation. In the Exp.3 as shown in Table 5, our model could gain significant improvement on the thin-slice images while preserving good performance on the thick-slice images, which demonstrated that our trained model was applicable for both types of images for both CT and MRI modalities.

Besides, in order to interpret the black-box segmentation model, we extracted the lowest bottom features and projected them into a 2D latent space using the PCA technique. We then computed the Dice score for each sample and visualised it in Fig. 17. In this figure, we can observe that slices sampled from the orange circle all contained a small region of ventricle where the model could not perform well. However, images from the green and yellow circle had multiple ventricles, which took a large proportion of the images. Therefore, these images could be well-predicted by our model.

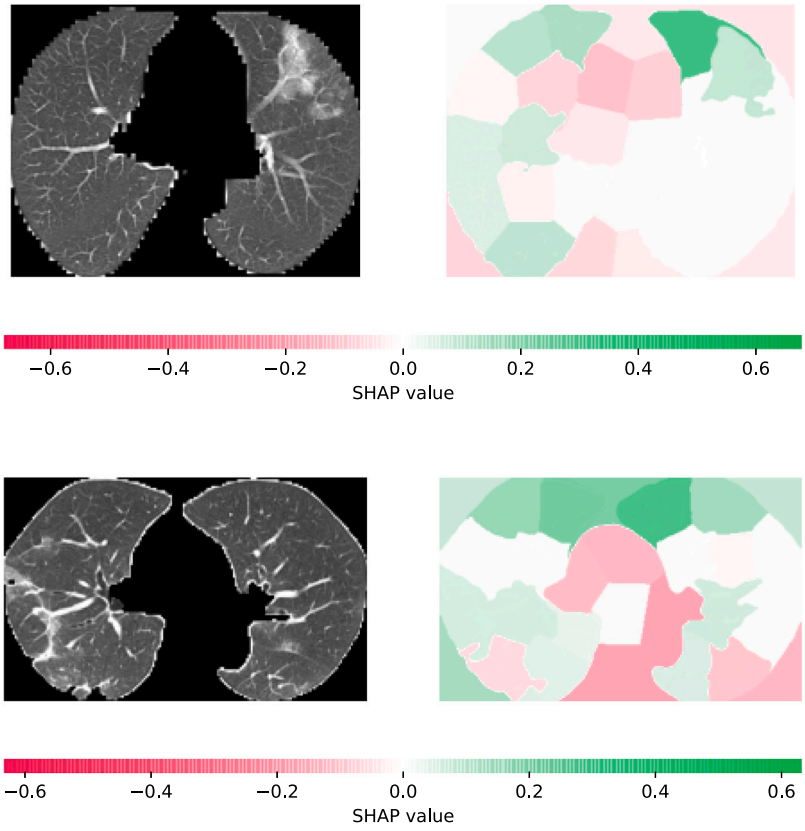


Fig. 16. The SHAP values for different super-pixels of the sampled images. We computed the SHAP values through the Kernel SHAP method [183]. The super-pixel with positive SHAP value indicates the positive impact to the positive prediction, while the negative value means that the super-pixel contributes to the negative prediction.

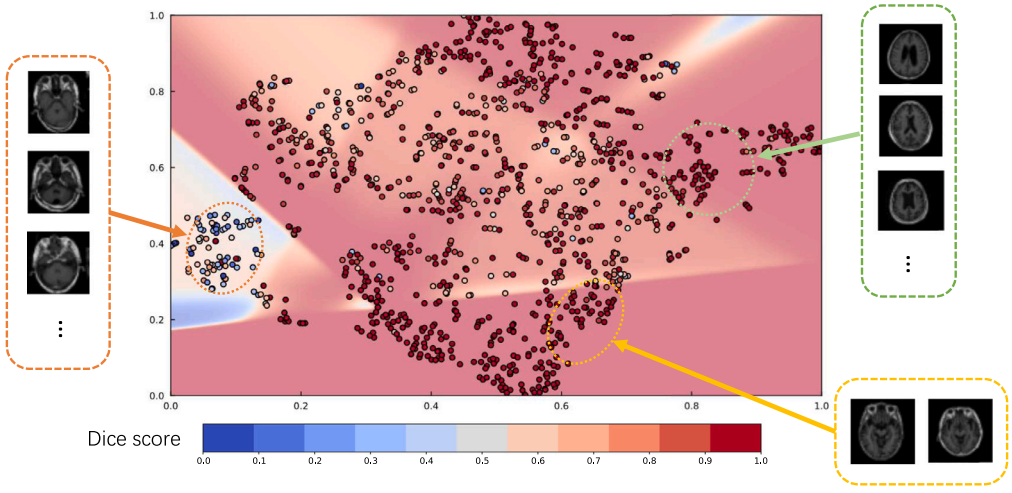


Fig. 17. The visualisation of the Dice scores of the projected images. The plane was computed by smoothing the Dice scores. It is of note that images sharing similar characteristics were clustered together. On the left-hand side and right-hand side, samples from different regions of the plane are presented.

Table 4
Comparison results (Dice scores) of our method vs. other state-of-the-art methods. Mixed represents the test set containing both thick-slice and thin-slice images.

Method	MRI			CT		
	Thick	Thin	Mixed	Thick	Thin	Mixed
U-Net [170]	0.9226	0.7665	0.8353	0.9351	0.7987	0.8513
U-Net++ [184]	0.9159	0.8495	0.8602	**0.9421**	0.7797	0.8424
Ours	**0.9323**	**0.9056**	**0.9099**	0.9365	**0.8697**	**0.8954**

Table 5
Dice scores comparison for verifying the effectiveness of each loss term. Mixed represents the test set containing both thick-slice and thin-slice images.

Exp.	$\mathcal{L}_S$	$\mathcal{L}_T$	MRI			CT		
			Thick	Thin	Mixed	Thick	Thin	Mixed
1	✓		**0.9390**	0.8199	0.8391	**0.9438**	0.8345	0.8767
2		✓	0.0034	0.0108	0.0110	0.0109	0.0006	0.0069
3	✓	✓	0.9323	**0.9056**	**0.9099**	0.9365	**0.8697**	**0.8954**

4.2.4. Qualitative results

To qualitatively examine the performance of our model and other state-of-the-art models, we presented some visualisation results of the CT and MRI images with thin-slices in Fig. 18, and computed the Dice scores for each segmentation result. For MRI images, our model and U-Net++ [184] were able to segment four ventricles in the brain. In particular, our model could predict the third ventricle in the brain more completely compared to the prediction generated by the U-Net++ [184] due to the informative feature representation by the pre-trained encoder. However, for CT images, the performance varied among different models. The primary reason is that original CT volumes contained the skull which could cause the brain to be visually unclear, after removing the skull, the contrast of the images could be largely distinct. More concretely, for those images with low contrast, (e.g., the row 1 and row 5 in Fig. 18), all of the three compared methods were capable of predicting the left lateral and right lateral ventricles. However, for those images with high contrast (e.g., the row 2 and row 4 in Fig. 18), our proposed method could predict most of the ventricle part in the brain while U-Net and U-Net++ failed.

In addition, we used the segmentation results generated by compared models to reconstruct the 3D images of each ventricle. The example is illustrated in Fig. 19. We can observe that U-Net [170] could hardly predict the ventricles on thin-slice images, while U-Net++ [184] was able to segment the left lateral and right lateral ventricles by taking advantage of dense connections of the intermediate feature maps. In contrast, our proposed method could not only predict the two ventricles mentioned above, but could also segment the third ventricle and the fourth ventricle well. One limitation of our model is that it could not predict the connection region between the third ventricle and fourth ventricle because the area is too small to be distinguished.

4.3. Discussions

In missions increasingly vital to human healthcare, AI is being deployed. Automated decisions should be explainable in order to create trust in AI and prevent an algorithm-based totalitarian society. This is not just a human right, for example, enshrined in the European GDPR, but an ultimate goal for algorithm developers who want to know if the necessary clinical characteristics are captured by the decision support systems. XAI should be possible to provide explanations in a systematic manner in order to make the explainability scalable. To construct a surrogate while-box model for the black-box model used to make a prediction, a typical solution is to use simpler, more intuitive decision algorithms. There is a chance, though, that the surrogate model is too complicated or too abstract for it to be truly understandable for humans.

In this study, we have firstly provided a mini-review for XAI methods and their specific applications in medicine and digital healthcare that is followed by two example showcases that we have developed. From our two showcases, we have explored the classification model and segmentation model in terms of sensitivity (i.e., LIME [5] and Kernel SHAP [183]) and decomposition (i.e., T-SNE [182] and CAMs). For LIME and Kernel SHAP methods, the individual sample can be analysed and interpreted with each super-pixel, which is useful for individual diagnosis. These methods can provide a straightforward view of how local explanations affect the final predictions. In this study, although we have not proposed novel XAI methods, the two showcases have demonstrated new pipelines of using XAI techniques for COVID-19 classification and segmentation for hydrocephalus. It is of note that both classification and segmentation are the two most widely investigated problems in medical data analysis with the arising of machine learning and deep learning. For example, machine and deep learning algorithms have been widely used for chest X-ray and CT images analysis for COVID-19 patients; however, there are still some common pitfalls in detection, diagnosis and prognosis of using these algorithms, e.g., data duplication, image quality issues and labelling accuracy and reproducibility, and most importantly, the lack of model interpretability [185–187].

On the other hand, T-SNE provides us with an insight into the strength and weakness of our proposed models. For example, in Fig. 17, the distribution of the decomposed image features has an association with the prediction performance, which indicates the weakness of the black-box segmentation models. Meanwhile, the distribution of decomposed image features also reveals the clustered characteristics of the raw inputs (Fig. 14), which can help us to find the reason why a model would make such predictions.

In consequence, these methods can also be classified into two categories named as perceptive interpretability and mathematical interpretability. When visual evidence is not useful or erroneous, the mathematical evidence can be used as the complement for interpretability. Therefore, various methods should be applied simultaneously for the sake of providing reliable interpretability.

Nevertheless, a significant drawback of the current studies on XAI is that the interpretations are focused on the intuition of experts rather than from the demands of the end-users [188]. Current local explanations are typically provided in a feature-importance vector format, which is a full causal attribution and a low-level interpretation. This format would be satisfactory if the description viewers were the developers and analysts, since they could use the mathematical study of the distribution of features to debug the models. However, this type of XAI is less accommodating if the description receivers are lay-users of the AI. XAI can explain the complete judgement logic of the model, which includes a large amount of repetitive knowledge which can confuse the lay-users. The presentation of the XAI algorithms should be further improved to increase customer satisfaction.

The poor abstraction level of explanations is another drawback. For example, despite XAI derived heatmaps can indicate that individual pixels are important, there is normally no correlation computed between these significance regions to more abstract principles such as the anatomical or pathological regions shown in the images. More importantly, the explanations ought to be understood by humans to make sense of them and to grasp the understandable actions of the model. It is indeed desirable to provide meta-explanations that can integrate evidence from these low-level heatmaps to describe the behaviour of the model at a more abstract, more humanly understandable level. However, this level of understanding can be hard and erroneous. Previously proposed methods have recently been suggested to aggregate low-level explanations and measure the semantics of neural representations. Thus, a constructive topic for future study is the development of more advanced meta-explanations that leverages multimodal information fusion.

Because the audiences of XAI results are essentially human users, an important future research direction is the use of XAI in human–machine interaction; therefore, research studies in XAI need to explore human factors. A prerequisite for good human–machine interaction is to construct explanations for the right user focus, for instance, develop XAI to ask the correct questions in the proper manner, which is crucial in the clinical environment. Optimisation of the reasoning procedure for optimal human use, however, is still a problem that demands more research. Eventually, a broad open gap in XAI is the use of interpretabilities beyond using visualisation techniques. Future studies will demonstrate how to incorporate XAI into a broader optimisation mechanism in order to, e.g., boost the efficiency of the model and reduce the model complexity.

5. Conclusion

The recent confluence of large-scale annotated clinical databases, the innovation of deep learning approaches, open-source software packages, and inexpensive and rapidly increasing computing capacity and cloud storage has fuelled the recent exponential growth in AI. This foretells to change the landscape of medical practice in the near future.

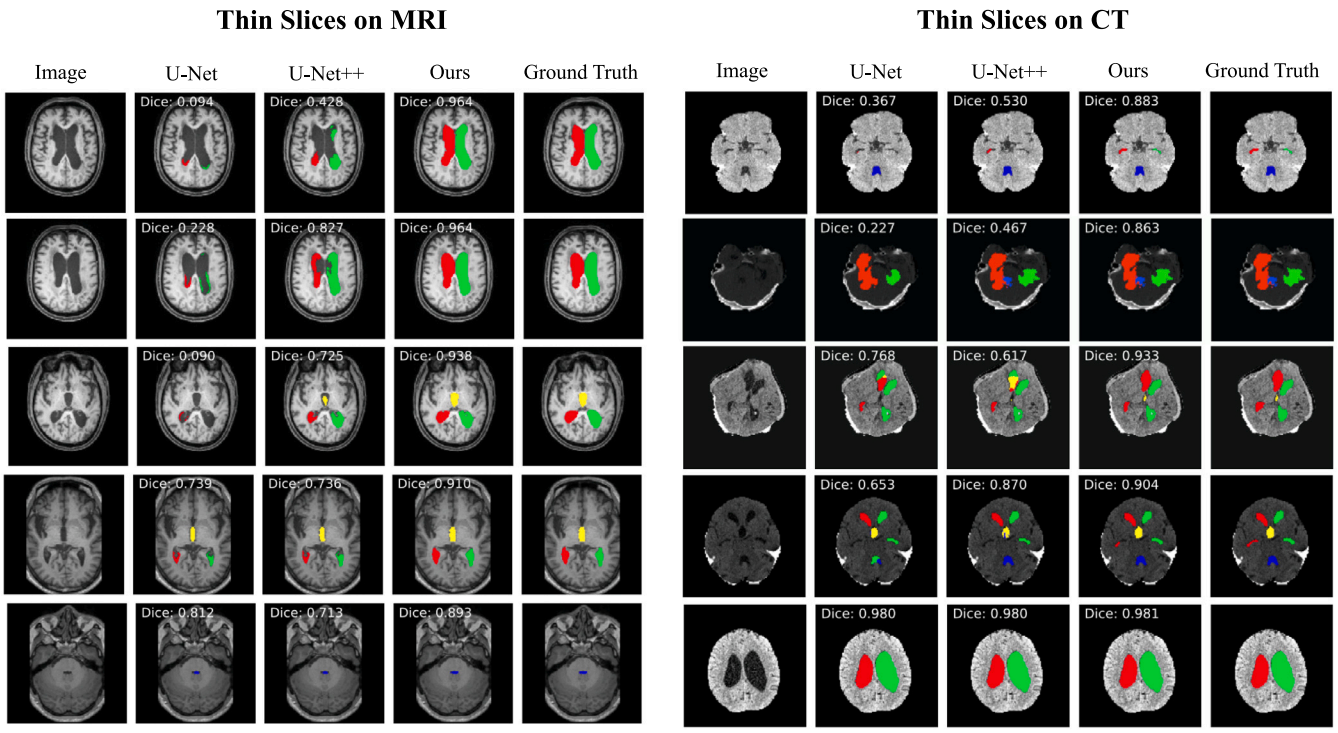


Fig. 18. The visualisation of the 3D brain ventricles segmentation results using different compared models. The right lateral ventricle is coloured in red; the left lateral ventricle is coloured in green; the yellow coloured region represents the third ventricle; and the blue region represents the fourth ventricle. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

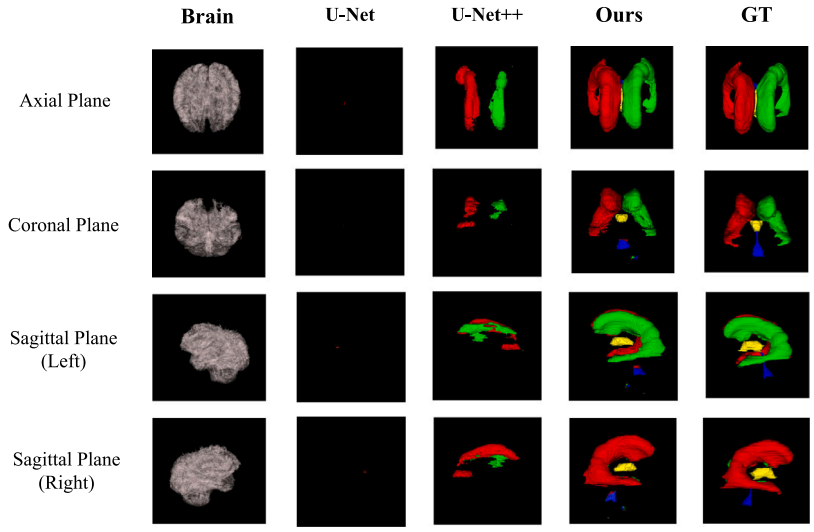


Fig. 19. Three-dimensional visualisation of the predictions on thin-slice MRI images for each ventricle segmented by different comparison models. The 3D segmentation results were visualised from the axial plane, the coronal plane, and the sagittal plane. Colouring scheme is consistent with Fig. 18.

AI systems have specialised success in certain clinical activities that are more able to assess patient prognosis compared to doctors, and can help in surgical procedures. If deep learning models continue to advance, there is a growing chance that AI could revolutionise medical practice and redefine the role of clinicians in the process. Our mini-review has demonstrated the research trends towards the trustable AI or trustworthy AI, which promotes the XAI globally, and XAI methods in medicine and digital healthcare are highly in demand. Additionally, our two showcases have shown promising XAI results for the two most widely investigated classification and segmentation problems in medical image analysis. We can envisage further development of XAI in medicine and digital healthcare by integrating information fusion from cross-modalities imaging and non-imaging clinical data can be a stepping stone toward a more general acceptance of AI in clinical practice. Ultimately, the trustable AI will promote confidence and openness of its deployment in the clinical arena and also make it easier to comply with the legislation of the GDPR and regulations of the NHS^X in the UK, CE-mark in the EU, FDA in the USA, and NMPA in China.

CRediT authorship contribution statement

Guang Yang: Conceived and designed the study, Data analysis, Data interpretation, Writing of the report, Literature search. **Qinghao Ye:** Conceived and designed the study, Data analysis, Data interpretation, Writing of the report, Data curation, Contributed to the tables and

figures. **Jun Xia:** Conceived and designed the study, Data analysis, Data interpretation, Writing of the report, Data collection.

Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.inffus.2021.07.016.QY is employed by Hangzhou Ocean's Smart Boya Co., Ltd., China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

This work was supported in part by the Hangzhou Economic and Technological Development Area Strategical Grant [Imperial Institute of Advanced Technology], in part by the Project of Shenzhen International Cooperation Foundation (GJHZ20180926165402083), in part by the Clinical Research Project of Shenzhen Health and Family Planning Commission (SZLY2018018), in part by the European Research Council Innovative Medicines Initiative on Development of Therapeutics and Diagnostics Combatting Coronavirus Infections Award 'DRAGON: rapiD and secuRe AI imaging based diaGnosis, stratification, fOllow-up, and preparedness for coronavirus paNdemics' [H2020-JTI-IMI2 101005122], in part by the AI for Health Imaging Award 'CHAIMELEON: Accelerating the Lab to Market Transition of AI Tools for Cancer Management' [H2020-SC1-FA-DTS-2019-1 952172], in part by the British Heart Foundation [TG/18/5/34111, PG/16/78/32402], and in part by the UK Research and Innovation [MR/V023799/1]. All authors approved the submitted version.

References

[1] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[2] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[3] W. Samek, K.-R. Müller, Towards explainable artificial intelligence, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 5–22.
[4] A. Rai, Explainable AI: From black box to glass box, J. Acad. Mark. Sci. 48 (1) (2020) 137–141.
[5] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[6] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke Vasc. Neurol. 2 (4) (2017) 230–243.
[7] T. Panch, H. Mattie, L.A. Celi, The "inconvenient truth" about AI in healthcare, NPJ Digit. Med. 2 (1) (2019) 1–3.
[8] K.-H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (10) (2018) 719–731.
[9] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.
[10] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[11] J. Ker, L. Wang, J. Rao, T. Lim, Deep learning applications in medical image analysis, IEEE Access 6 (2017) 9375–9389.
[12] S.E. Dilsizian, E.L. Siegel, Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment, Curr. Cardiol. Rep. 16 (1) (2014) 441.
[13] V.L. Patel, E.H. Shortliffe, M. Stefanelli, P. Szolovits, M.R. Berthold, R. Bellazzi, A. Abu-Hanna, The coming of age of artificial intelligence in medicine, Artif. Intell. Med. 46 (1) (2009) 5–17.
[14] S. Jha, E.J. Topol, Adapting to artificial intelligence: radiologists and pathologists as information specialists, JAMA 316 (22) (2016) 2353–2354.
[15] E. Strickland, IBM Watson, heal thyself: How IBM overpromised and underdelivered on ai health care, IEEE Spectr. 56 (4) (2019) 24–31.
[16] N.S. Weingart, R.M. Wilson, R.W. Gibberd, B. Harrison, Epidemiology of medical error, Bmj 320 (7237) (2000) 774–777.
[17] M.L. Graber, N. Franklin, R. Gordon, Diagnostic error in internal medicine, Arch. Intern. Med. 165 (13) (2005) 1493–1499.
[18] B. Winters, J. Custer, S.M. Galvagno, E. Colantuoni, S.G. Kapoor, H. Lee, V. Goode, K. Robinson, A. Nakhasi, P. Pronovost, et al., Diagnostic errors in the intensive care unit: a systematic review of autopsy studies, BMJ Qual. Saf. 21 (11) (2012) 894–902.
[19] C.S. Lee, P.G. Nagy, S.J. Weaver, D.E. Newman-Toker, Cognitive and system factors contributing to diagnostic errors in radiology, Am. J. Roentgenol. 201 (3) (2013) 611–617.
[20] D.B. Neill, Using artificial intelligence to improve hospital inpatient care, IEEE Intell. Syst. 28 (2) (2013) 92–95.
[21] R.A. Miller, Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary, J. Am. Med. Inform. Assoc. 1 (1) (1994) 8–27.
[22] M.A. Musen, B. Middleton, R.A. Greenes, Clinical decision-support systems, in: Biomedical Informatics, Springer, 2014, pp. 643–674.
[23] M. Kundu, M. Nasipuri, D.K. Basu, Knowledge-based ECG interpretation: a critical review, Pattern Recognit. 33 (3) (2000) 351–373.
[24] F. De Dombal, D. Leaper, J.R. Staniland, A. McCann, J.C. Horrocks, Computer-aided diagnosis of acute abdominal pain, Br. Med. J. 2 (5804) (1972) 9–13.
[25] E.H. Shortliffe, R. Davis, S.G. Axline, B.G. Buchanan, C.C. Green, S.N. Cohen, Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system, Comput. Biomed. Res. 8 (4) (1975) 303–320.
[26] G.O. Barnett, J.J. Cimino, J.A. Hupp, E.P. Hoffer, DXplain: an evolving diagnostic decision-support system, JAMA 258 (1) (1987) 67–74.
[27] R.A. Miller, M.A. McNeil, S.M. Challinor, F.E. Masarie Jr, J.D. Myers, The internist-1/quick medical reference project—Status report, West. J. Med. 145 (6) (1986) 816.
[28] E.S. Berner, G.D. Webster, A.A. Shugerman, J.R. Jackson, J. Algina, A.L. Baker, E.V. Ball, C.G. Cobbs, V.W. Dennis, E.P. Frenkel, et al., Performance of four computer-based diagnostic systems, N. Engl. J. Med. 330 (25) (1994) 1792–1796.
[29] P. Szolovits, S.G. Pauker, Categorical and probabilistic reasoning in medical diagnosis, Artificial Intelligence 11 (1–2) (1978) 115–144.
[30] P. Szolovits, S.G. Pauker, Categorical and probabilistic reasoning in medicine revisited, in: Artificial Intelligence in Perspective, MIT Press, Cambridge, MA, 1994.
[31] R.C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.
[32] K.-H. Yu, M. Snyder, Omics profiling in precision oncology, Mol. Cell. Proteom. 15 (8) (2016) 2525–2536.
[33] K. Roberts, M.R. Boland, L. Pruinelli, J. Dcruz, A. Berry, M. Georgsson, R. Hazen, R.F. Sarmiento, U. Backonja, K.-H. Yu, et al., Biomedical informatics advancing the national health agenda: the AMIA 2015 year-in-review in clinical and consumer informatics, J. Am. Med. Inform. Assoc. 24 (e1) (2017) e185–e190.
[34] W. Rogers, S. Thulasi Seetha, T.A. Refaee, R.I. Lieverse, R.W. Granzier, A. Ibrahim, S.A. Keek, S. Sanduleanu, S.P. Primakov, M.P. Beuque, et al., Radiomics: from qualitative to quantitative imaging, Br. J. Radiol. 93 (1108) (2020) 20190948.
[35] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, vol. 1, MIT press Cambridge, 2016.
[36] Y.E. Wang, G.-Y. Wei, D. Brooks, Benchmarking TPU, GPU, and CPU platforms for deep learning, 2019, arXiv Preprint arXiv:1907.10701.
[37] K. Tomczak, P. Czerwińska, M. Wiznerowicz, The cancer genome atlas (TCGA): an immeasurable source of knowledge, Contemp. Oncol. 19 (1A) (2015) A68.
[38] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al., UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, Plos Med. 12 (3) (2015) e1001779.
[39] V. Ljosa, K.L. Sokolnicki, A.E. Carpenter, Annotated high-throughput microscopy image sets for validation, Nature Methods 9 (7) (2012) 637.
[40] E. Williams, J. Moore, S.W. Li, G. Rustici, A. Tarkowska, A. Chessel, S. Leo, B. Antal, R.K. Ferguson, U. Sarkans, et al., Image data resource: a bioimage data integration and publication platform, Nature Methods 14 (8) (2017) 775–781.
[41] C.M. DesRoches, E.G. Campbell, S.R. Rao, K. Donelan, T.G. Ferris, A. Jha, R. Kaushal, D.E. Levy, S. Rosenbaum, A.E. Shields, et al., Electronic health records in ambulatory care—a national survey of physicians, N. Engl. J. Med. 359 (1) (2008) 50–60.
[42] C.-J. Hsiao, A.K. Jha, J. King, V. Patel, M.F. Furukawa, F. Mostashari, Office-based physicians are responding to incentives and assistance by adopting and using electronic health records, Health Aff. 32 (8) (2013) 1470–1477.

[43] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E.F. Fang, W. Menpes-Smith, J. Xia, et al., Weakly supervised deep learning for COVID-19 infection detection and classification from CT images, IEEE Access (2020).

[44] G. Litjens, C.I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, J. Van Der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. 6 (2016) 26286.

[45] N. Zhang, G. Yang, Z. Gao, C. Xu, Y. Zhang, R. Shi, J. Keegan, L. Xu, H. Zhang, Z. Fan, et al., Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI, Radiology 291 (3) (2019) 606–617.

[46] Y. Cao, Z. Wang, Z. Liu, Y. Li, X. Xiao, L. Sun, Y. Zhang, H. Hou, P. Zhang, G. Yang, Multiparameter synchronous measurement with IVUS images for intelligently diagnosing coronary cardiac disease, IEEE Trans. Instrum. Meas. (2020).

[47] A. Cheerla, O. Gevaert, Deep learning with multimodal representation for pancancer prognosis prediction, Bioinformatics 35 (14) (2019) i446–i454.

[48] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., Machine learning for COVID-19 detection and prognostication using chest radiographs and CT scans: a systematic methodological review, 2020, arXiv Preprint arXiv:2008.06388.

[49] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, A. Telenti, A primer on deep learning in genomics, Nature Genet. 51 (1) (2019) 12–18.

[50] S.M. Waldstein, P. Seeböck, R. Donner, A. Sadeghipour, H. Bogunović, A. Osborne, U. Schmidt-Erfurth, Unbiased identification of novel subclinical imaging biomarkers using unsupervised deep learning, Sci. Rep. 10 (1) (2020) 1–9.

[51] L. Li, F. Wu, G. Yang, L. Xu, T. Wong, R. Mohiaddin, D. Firmin, J. Keegan, X. Zhuang, Atrial scar quantification via multi-scale CNN in the graph-cuts framework, Med. Image Anal. 60 (2020) 101595.

[52] N.D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, F. Kawsar, An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices, in: Proceedings of the 2015 International Workshop on Internet of Things Towards Applications, 2015, pp. 7–12.

[53] A.I. Chen, M.L. Balter, T.J. Maguire, M.L. Yarmush, Deep learning robotic guidance for autonomous vascular access, Nat. Mach. Intell. 2 (2) (2020) 104–115.

[54] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature Med. 25 (1) (2019) 24–29.

[55] Z. Obermeyer, E.J. Emanuel, Predicting the future—big data, machine learning, and clinical medicine, N. Engl. J. Med. 375 (13) (2016) 1216.

[56] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G.S. Corrado, L. Peng, D.R. Webster, Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy, Ophthalmology 125 (8) (2018) 1264–1272.

[57] D. Rebholz-Schuhmann, A.J. Jimeno-Yepes, E. van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, K. Tomanek, E. Beisswanger, et al., The calbc silver standard corpus for biomedical named entities-a study in harmonizing the contributions from four independent named entity taggers, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010.

[58] J.C. Kirby, P. Speltz, L.V. Rasmussen, M. Basford, O. Gottesman, P.L. Peissig, J.A. Pacheco, G. Tromp, J. Pathak, D.S. Carrell, et al., PheKB: a Catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inform. Assoc. 23 (6) (2016) 1046–1052.

[59] Z.-Q. Zhao, P. Zheng, S.-t. Xu, X. Wu, Object detection with deep learning: A review, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019) 3212–3232.

[60] S.P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, S. Jain, Machine translation using deep learning: An overview, in: 2017 International Conference on Computer, Communications and Electronics (Comptelix), IEEE, 2017, pp. 162–167.

[61] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., Deep speech 2: End-to-end speech recognition in English and Mandarin, in: International Conference on Machine Learning, 2016, pp. 173–182.

[62] A.B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, Speech recognition using deep neural networks: A systematic review, IEEE Access 7 (2019) 19143–19165.

[63] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep learning for audio signal processing, IEEE J. Sel. Top. Sign. Proces. 13 (2) (2019) 206–219.

[64] T. Elsken, J.H. Metzen, F. Hutter, et al., Neural architecture search: A survey, J. Mach. Learn. Res. 20 (55) (2019) 1–21.

[65] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37.

[66] W. Zhang, F. Liu, L. Luo, J. Zhang, Predicting drug side effects by multi-label learning and ensemble learning, BMC Bioinformatics 16 (1) (2015) 365.

[67] G. Yang, F. Raschke, T.R. Barrick, F.A. Howe, Manifold learning in MR spectroscopy using nonlinear dimensionality reduction and unsupervised clustering, Magn. Reson. Med. 74 (3) (2015) 868–878.

[68] L.P. Zhao, H. Bolouri, Object-oriented regression for building predictive models with high dimensional omics data from translational studies, J. Biomed. Inform. 60 (2016) 431–445.

[69] S.G. Kim, N. Theera-Ampornpunt, C.-H. Fang, M. Harwani, A. Grama, S. Chaterji, Opening up the blackbox: an interpretable deep neural network-based classifier for cell-type specific enhancer predictions, BMC Syst. Biol. 10 (2) (2016) 243–258.

[70] J. Hao, Y. Kim, T.-K. Kim, M. Kang, PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data, BMC Bioinformatics 19 (1) (2018) 1–13.

[71] M. Bernardini, L. Romeo, P. Misericordia, E. Frontoni, Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine, IEEE J. Biomed. Health Inf. 24 (1) (2019) 235–246.

[72] A. Eck, L.M. Zintgraf, E. de Groot, T.G. de Meij, T.S. Cohen, P. Savelkoul, M. Welling, A. Budding, Interpretation of microbiota-based diagnostics by explaining individual classifier decisions, BMC Bioinformatics 18 (1) (2017) 441.

[73] W. Ge, J.-W. Huh, Y.R. Park, J.-H. Lee, Y.-H. Kim, A. Turchin, An interpretable ICU mortality prediction model based on logistic regression and recurrent neural networks with LSTM units, in: AMIA Annual Symposium Proceedings, 2018, American Medical Informatics Association, 2018, p. 460.

[74] J. Zuallaert, F. Godin, M. Kim, A. Soete, Y. Saeys, W. De Neve, Splicerover: interpretable convolutional neural networks for improved splice site prediction, Bioinformatics 34 (24) (2018) 4180–4188.

[75] J. Suh, S. Yoo, J. Park, S.Y. Cho, M.C. Cho, H. Son, H. Jeong, Development and validation of explainable AI-based decision-supporting tool for prostate biopsy, BJU Int. (2020).

[76] A. Singh, A.R. Mohammed, J. Zelek, V. Lakshminarayanan, Interpretation of deep learning using attributions: application to ophthalmic diagnosis, in: Applications of Machine Learning 2020, vol. 11511, International Society for Optics and Photonics, 2020, p. 115110A.

[77] B.C. Kwon, M.-J. Choi, J.T. Kim, E. Choi, Y.B. Kim, S. Kwon, J. Sun, J. Choo, Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records, IEEE Trans. Vis. Comput. Graphics 25 (1) (2018) 299–309.

[78] J. Zhang, K. Kowsari, J.H. Harrison, J.M. Lobo, L.E. Barnes, Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record, IEEE Access 6 (2018) 65333–65346.

[79] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Adv. Neural Inf. Process. Syst. 29 (2016) 3504–3512.

[80] D.A. Kaji, J.R. Zech, J.S. Kim, S.K. Cho, N.S. Dangayach, A.B. Costa, E.K. Oermann, An attention based deep learning model of clinical events in the intensive care unit, PLoS One 14 (2) (2019) e0211057.

[81] B. Shickel, T.J. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, P. Rashidi, DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning, Sci. Rep. 9 (1) (2019) 1–12.

[82] H. Hu, A. Xiao, S. Zhang, Y. Li, X. Shi, T. Jiang, L. Zhang, L. Zhang, J. Zeng, DeepHINT: understanding HIV-1 integration via deep learning with attention, Bioinformatics 35 (10) (2019) 1660–1667.

[83] M. Izadyyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L.B. Moreira, J. Eschbacher, P. Nakaji, M.C. Preul, Y. Yang, Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 300–308.

[84] G. Zhao, B. Zhou, K. Wang, R. Jiang, M. Xu, Respond-CAM: Analyzing deep models for 3D imaging data by visualizations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 485–492.

[85] H.D. Couture, J.S. Marron, C.M. Perou, M.A. Troester, M. Niethammer, Multiple instance learning for heterogeneous images: Training a CNN for histopathology, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 254–262.

[86] H. Lee, S. Yune, M. Mansouri, M. Kim, S.H. Tajmir, C.E. Guerrier, S.A. Ebert, S.R. Pomerantz, J.M. Romero, S. Kamalian, et al., An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets, Nat. Biomed. Eng. 3 (3) (2019) 173.

[87] J. Kim, H.J. Kim, C. Kim, W.H. Kim, Artificial intelligence in breast ultrasonography, Ultrasonography (Seoul, Korea) (2020).

[88] P. Rajpurkar, A. Park, J. Irvin, C. Chute, M. Bereket, D. Mastrodicasa, C.P. Langlotz, M.P. Lungren, A.Y. Ng, B.N. Patel, AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining, Sci. Rep. 10 (1) (2020) 1–7.

[89] M. Porumb, S. Stranges, A. Pescapè, L. Pecchia, Precision medicine and artificial intelligence: A pilot study on deep learning for hypoglycemic events detection based on ECG, Sci. Rep. 10 (1) (2020) 1–16.

[90] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1721–1730.

[91] B. Letham, C. Rudin, T.H. McCormick, D. Madigan, et al., Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model, Ann. Appl. Stat. 9 (3) (2015) 1350–1371.

[92] Z. Che, S. Purushotham, R. Khemani, Y. Liu, Interpretable deep models for ICU outcome prediction, in: AMIA Annual Symposium Proceedings, vol. 2016, American Medical Informatics Association, 2016, p. 371.

[93] Y. Ming, H. Qu, E. Bertini, Rulematrix: Visualizing and understanding classifiers with rules, IEEE Trans. Vis. Comput. Graphics 25 (1) (2018) 342–352.

[94] C. Xiao, T. Ma, A.B. Dieng, D.M. Blei, F. Wang, Readmission prediction via deep contextual embedding of clinical concepts, PLoS One 13 (4) (2018) e0195024.

[95] R. Davoodi, M.H. Moradi, Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier, J. Biomed. Inform. 79 (2018) 48–59.

[96] H. Lee, S.T. Kim, Y.M. Ro, Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer, 2019, pp. 21–29.

[97] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, C. Pattichis, Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction, in: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2019, pp. 817–821.

[98] L. Pan, G. Liu, X. Mao, H. Li, J. Zhang, H. Liang, X. Li, Development of prediction models using machine learning algorithms for girls with suspected central precocious puberty: retrospective study, JMIR Med. Inform. 7 (1) (2019) e11728.

[99] S. Ghafouri-Fard, M. Taheri, M.D. Omrani, A. Daaee, H. Mohammad-Rahimi, H. Kazazi, Application of single-nucleotide polymorphisms in the diagnosis of autism spectrum disorders: a preliminary study with artificial neural networks, J. Mol. Neurosci. 68 (4) (2019) 515–521.

[100] M.S. Kovalev, L.V. Utkin, E.M. Kasimov, Survlime: A method for explaining machine learning survival models, Knowl.-Based Syst. 203 (2020) 106164, http://dx.doi.org/10.1016/j.knosys.2020.106164, http://www.sciencedirect.com/science/article/pii/S0950705120304044.

[101] A. Meldo, L. Utkin, M. Kovalev, E. Kasimov, The natural language explanation algorithms for the lung cancer computer-aided diagnosis system, Artif. Intell. Med. 108 (2020) 101952.

[102] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI: An ontology-based approach to black-box sequential data classification explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, in: FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 629–639, http://dx.doi.org/10.1145/3351095.3372855, https://doi.org/10.1145/3351095.3372855.

[103] S.M. Lauritsen, M. Kristensen, M.V. Olsen, M.S. Larsen, K.M. Lauritsen, M.J. Jø rgensen, J. Lange, B. Thiesson, Explainable artificial intelligence model to predict acute critical illness from electronic health records, Nature Commun. 11 (1) (2020) 1–11.

[104] L.M. Lee, L.O. Gostin, Ethical collection, storage, and use of public health data: a proposal for a national privacy protection, JAMA 302 (1) (2009) 82–84.

[105] S. Narayan, M. Gagné, R. Safavi-Naini, Privacy preserving EHR system using attribute-based infrastructure, in: Proceedings of the 2010 ACM Workshop on Cloud Computing Security Workshop, 2010, pp. 47–52.

[106] R.H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron, A. Shabo, HL7 clinical document architecture, release 2, J. Am. Med. Inform. Assoc. 13 (1) (2006) 30–39.

[107] K.D. Mandl, I.S. Kohane, Escaping the EHR trap—the future of health IT, N. Engl. J. Med. 366 (24) (2012) 2240–2242.

[108] J.C. Mandel, D.A. Kreda, K.D. Mandl, I.S. Kohane, R.B. Ramoni, SMART On FHIR: a standards-based, interoperable apps platform for electronic health records, J. Am. Med. Inform. Assoc. 23 (5) (2016) 899–908.

[109] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Med. 25 (1) (2019) 44–56.

[110] K.-H. Yu, I.S. Kohane, Framing the challenges of artificial intelligence in medicine, BMJ Qual. Saf. 28 (3) (2019) 238–241.

[111] D.D. Miller, The medical AI insurgency: what physicians must know about data to practice with intelligent machines, NPJ Digit. Med. 2 (1) (2019) 1–5.

[112] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, K. Tsaneva-Atanasova, Artificial intelligence, bias and clinical safety, BMJ Qual. Saf. 28 (3) (2019) 231–237.

[113] M. DeCamp, C. Lindvall, Latent bias and the implementation of artificial intelligence in medicine, J. Am. Med. Inform. Assoc. 27 (12) (2020) 2020–2023.

[114] P. Esmaeilzadeh, Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 1–19.

[115] J.R. England, P.M. Cheng, Artificial intelligence for medical image analysis: a guide for authors and reviewers, Am. J. Roentgenol. 212 (3) (2019) 513–519.

[116] A. Gomolin, E. Netchiporouk, R. Gniadecki, I.V. Litvinov, Artificial intelligence applications in dermatology: Where do we stand? Front. Med. 7 (2020).

[117] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.

[118] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Workshop At International Conference on Learning Representations, 2014.

[119] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[120] Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, Mdnet: A semantically and visually interpretable medical image diagnosis network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6428–6436.

[121] P. Croskerry, K. Cosby, M.L. Graber, H. Singh, Diagnosis: Interpreting the Shadows, CRC Press, 2017.

[122] T.P. Quinn, S. Jacobs, M. Senadeera, V. Le, S. Coghlan, The three ghosts of medical AI: Can the black-box present deliver? 2020, arXiv preprint arXiv:2012.06000.

[123] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain? 2017, arXiv preprint arXiv:1712.09923.

[124] S. Tonekaboni, S. Joshi, M.D. McCradden, A. Goldenberg, What clinicians want: Contextualizing explainable machine learning for clinical end use, in: F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), Proceedings of the 4th Machine Learning for Healthcare Conference, in: Proceedings of Machine Learning Research, vol. 106, PMLR, Ann Arbor, Michigan, 2019, pp. 359–380, http://proceedings.mlr.press/v106/tonekaboni19a.html.

[125] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) e1312.

[126] M.S. Hossain, G. Muhammad, N. Guizani, Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics, IEEE Netw. 34 (4) (2020) 126–132.

[127] E. Khodabandehloo, D. Riboni, A. Alimohammadi, HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline, Future Gener. Comput. Syst. 116 (2021) 168–189.

[128] K. Kallianos, J. Mongan, S. Antani, T. Henry, A. Taylor, J. Abuya, M. Kohli, How far have we come? Artificial intelligence for chest radiograph interpretation, Clin. Radiol. 74 (5) (2019) 338–345.

[129] C. Zucco, H. Liang, G. Di Fatta, M. Cannataro, Explainable sentiment analysis with applications in medicine, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018, pp. 1740–1747.

[130] C.P. Langlotz, B. Allen, B.J. Erickson, J. Kalpathy-Cramer, K. Bigelow, T.S. Cook, A.E. Flanders, M.P. Lungren, D.S. Mendelson, J.D. Rudie, et al., A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The academy workshop, Radiology 291 (3) (2019) 781–791.

[131] A.J. London, Artificial intelligence and black-box medical decisions: accuracy versus explainability, Hastings Cent. Rep. 49 (1) (2019) 15–21.

[132] G. Stiglic, S. Kocbek, I. Pernek, P. Kokol, Comprehensive decision tree models in bioinformatics, PLoS One 7 (3) (2012) e33812.

[133] G. Valdes, J.M. Luna, E. Eaton, C.B. Simone II, L.H. Ungar, T.D. Solberg, Mediboost: a patient stratification tool for interpretable decision making in the era of precision medicine, Sci. Rep. 6 (2016) 37854.

[134] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. (2020).

[135] S.N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J.H. Chen, X. Liu, Z. He, Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, J. Am. Med. Inform. Assoc. (2020).

[136] M. Pocevičiūtė, G. Eilertsen, C. Lundström, Survey of XAI in digital pathology, in: Artificial Intelligence and Machine Learning for Digital Pathology, Springer, 2020, pp. 56–88.

[137] A.B. Tosun, F. Pullara, M.J. Becich, D.L. Taylor, S.C. Chennubhotla, J.L. Fine, Histomapr™: An explainable AI (xAI) platform for computational pathology solutions, in: Artificial Intelligence and Machine Learning for Digital Pathology, Springer, 2020, pp. 204–227.

[138] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.

[139] P. Comon, Independent component analysis, a new concept? Signal Process. 36 (3) (1994) 287–314.

[140] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 14 (2001) 585–591.

[141] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832.

[142] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, Entropy 23 (1) (2021) 18.

[143] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 2522–5839.

[144] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2018) 1–15.

[145] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, 2017, pp. 3145–3153.

[146] J. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: the all convolutional net, in: ICLR (Workshop Track), 2015, http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a.

[147] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.

[148] H. Chen, S. Lundberg, S.-I. Lee, Explaining models by propagating Shapley values of local components, in: Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability, Springer International Publishing, Cham, 2021, pp. 261–270.

[149] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, http://arxiv.org/abs/1409.0473.

[150] L. Putelli, A.E. Gerevini, A. Lavelli, I. Serina, Applying self-interaction attention for extracting drug-drug interactions, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2019, pp. 445–460.

[151] D. Mascharka, P. Tran, R. Soklaski, A. Majumdar, Transparency by design: Closing the gap between performance and interpretability in visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4942–4950.

[152] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[153] A. Polino, R. Pascanu, D. Alistarh, Model compression via distillation and quantization, in: International Conference on Learning Representations, 2018, https://openreview.net/forum?id=S1XolQbRW.

[154] M. Wu, M. Hughes, S. Parbhoo, M. Zazzi, V. Roth, F. Doshi-Velez, Beyond sparsity: Tree regularization of deep models for interpretability, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[155] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: NIPS Deep Learning and Representation Learning Workshop, 2015, http://arxiv.org/abs/1503.02531.

[156] G. Hinton, N. Frosst, Distilling a neural network into a soft decision tree, 2017, https://arxiv.org/pdf/1711.09784.pdf.

[157] X. Yang, H. Zhang, J. Cai, Auto-encoding and distilling scene graphs for image captioning, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

[158] X.-H. Li, C.C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, et al., A survey of data-driven and knowledge-aware explainable AI, IEEE Trans. Knowl. Data Eng. (2020).

[159] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, F. Doshi-Velez, An evaluation of the human-interpretability of explanation, 2019, arXiv preprint arXiv:1902.00006.

[160] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Faithful and customizable explanations of black box models, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 131–138.

[161] D. Das, J. Ito, T. Kadowaki, K. Tsuda, An interpretable machine learning model for diagnosis of Alzheimer's disease, PeerJ 7 (2019) e6543.

[162] Y. Liang, S. Li, C. Yan, M. Li, C. Jiang, Explaining the black-box model: A survey of local interpretation methods for deep neural networks, Neurocomputing 419 (2021) 168–182.

[163] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 193–209.

[164] Y. Lin, S. Dong, Y. Yeh, Y. Wu, G. Lan, C. Liu, T.C. Chu, Emergency management and infection control in a radiology department during an outbreak of severe acute respiratory syndrome, Br. J. Radiol. 78 (931) (2005) 606–611.

[165] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[166] Z.-H. Zhou, Multi-instance learning: A survey, Tech. Rep 2, Department of Computer Science & Technology, Nanjing University, 2004.

[167] A.J. Bekker, J. Goldberger, Training deep neural-networks based on unreliable labels, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2682–2686.

[168] X. Qian, Y. Lin, Y. Zhao, X. Yue, B. Lu, J. Wang, Objective ventricle segmentation in brain CT with ischemic stroke based on anatomical knowledge, BioMed Res. Int. 2017 (2017).

[169] V. Cherukuri, P. Ssenyonga, B.C. Warf, A.V. Kulkarni, V. Monga, S.J. Schiff, Learning based segmentation of CT brain images: application to postoperative hydrocephalic scans, IEEE Trans. Biomed. Eng. 65 (8) (2017) 1871–1884.

[170] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[171] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[172] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2015) 295–307.

[173] Y. Grandvalet, Y. Bengio, Semi-supervised learning by entropy minimization, in: Advances in Neural Information Processing Systems, 2005, pp. 529–536.

[174] T.-H. Vu, H. Jain, M. Bucher, M. Cord, P. Pérez, Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2517–2526.

[175] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[176] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, et al., Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography, Cell (2020).

[177] L. Wang, A. Wong, COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images, 2020, arXiv preprint arXiv:2003.09871.

[178] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al., Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT, Radiology (2020) 200905.

[179] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, F. Yan, Z. Ding, Q. Yang, B. Song, et al., Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia, 2020, arXiv preprint arXiv:2005.02690.

[180] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F.A. Wichmann, W. Brendel, Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, 2018, arXiv preprint arXiv:1811.12231.

[181] A.M. Carrington, D.G. Manuel, P.W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, M. McInnes, O. Magwood, et al., Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding and interpretation, 2021, arXiv preprint arXiv:2103.11357.

[182] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.

[183] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[184] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-Net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.

[185] C. Long, H. Xu, Q. Shen, X. Zhang, B. Fan, C. Wang, B. Zeng, Z. Li, X. Li, H. Li, Diagnosis of the coronavirus disease (COVID-19): rRT-PCR or CT? Eur. J. Radiol. 126 (2020) 108961.

[186] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nat. Mach. Intell. 3 (3) (2021) 199–217.

[187] D. Driggs, I. Selby, M. Roberts, E. Gkrania-Klotsas, J.H. Rudd, G. Yang, J. Babar, E. Sala, C.-B. Schönlieb, A.-C. collaboration, Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise, Radiological Society of North America, 2021.

[188] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Commun. ACM 63 (1) (2019) 68–77.