nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-023-00670-0

The importance of resource awareness in artificial intelligence for healthcare

Received: 1 September 2022

Accepted: 3 May 2023

Published online: 12 June 2023

Check for updates

Zhenge Jia[1], Jianxu Chen[2], Xiaowei Xu[3] ✉, John Kheir[4], Jingtong Hu[5], Han Xiao[6], Sui Peng[7], Xiaobo Sharon Hu[1], Danny Chen[1] & Yiyu Shi[1] ✉

Artificial intelligence and machine learning (AI/ML) models have been adopted in a wide range of healthcare applications, from medical image computing and analysis to continuous health monitoring and management. Recent data have demonstrated a clear trend that AI/ML model sizes, as well as their computational complexity, memory consumption and the scale of the required training data and costs, are experiencing an exponential increase. The developments in current computing hardware platforms, storage infrastructure, networking and domain expertise cannot keep up with this exponential growth in resources demanded by the AI/ML models. Here, we first analyse this recent trend and highlight that there are resource sustainability issues in AI/ML for healthcare. We then present various algorithm/system innovations that will help address these issues. We finally outline future directions to proactively and prospectively tackle these resource sustainability issues.

With the ever-growing volume of available data in the biomedical domain, artificial intelligence and machine learning (AI/ML) models are showing great potential across a wide range of healthcare applications, from medical image computing and analysis[1–4] to implantable health monitoring[5]. The performance of AI/ML models has been demonstrated to be comparable to or even superior to human expert performance in various healthcare applications[6,7]. The adoption of AI/ML models in healthcare has substantially reduced labour costs and freed doctors from tedious manual work[8]. With increasing computing power and the ever-growing available healthcare data, AI/ML models, while achieving better inference performance, are currently experiencing an exponential increase in terms of their size and demand for resources.

Unfortunately, with the exponential growth in the size of models, as well as the associated increase in computational complexity and rapid growth in the volume of health data, the development of accurate AI/ML models, which often consume extensive resources for training as well as testing, is facing critical sustainability issues in relation to energy, storage, computing power, networking and domain expertise. For healthcare applications with access to powerful computing infrastructure, the energy consumed in the operation of large AI/ML models may also be subject to unsustainability issues as a result of the slowdown in advances in hardware platforms. As model sizes grow, the amount of health data required for training will also increase dramatically. For healthcare applications for which cloud servers are easily accessible through network connections, a storage sustainability issue arises in relation to upgrading and maintaining current storage infrastructures, with the high associated costs (when budgets are often limited). For applications that are restricted to edge computation of embedded hardware, the energy, computing power, networking and storage sustainability issues are even more severe as there are already constraints in power and areas such as security, privacy and latency. In addition to hardware-related sustainability issues, unsustainable

[1]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA. [2]Leibniz-Institut für Analytische Wissenschaften – ISAS - e.V., Dortmund, Germany. [3]Guangdong Provincial Key Laboratory of South China Structural Heart Disease, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), South Medical University, Guangzhou, China. [4]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [5]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA. [6]Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. [7]Department of Gastroenterology and Hepatology, Clinical Trials Unit, Institute of Precision Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. ✉e-mail: xiao.wei.xu@foxmail.com; yshi4@nd.edu

https://doi.org/10.1038/s42256-023-00670-0

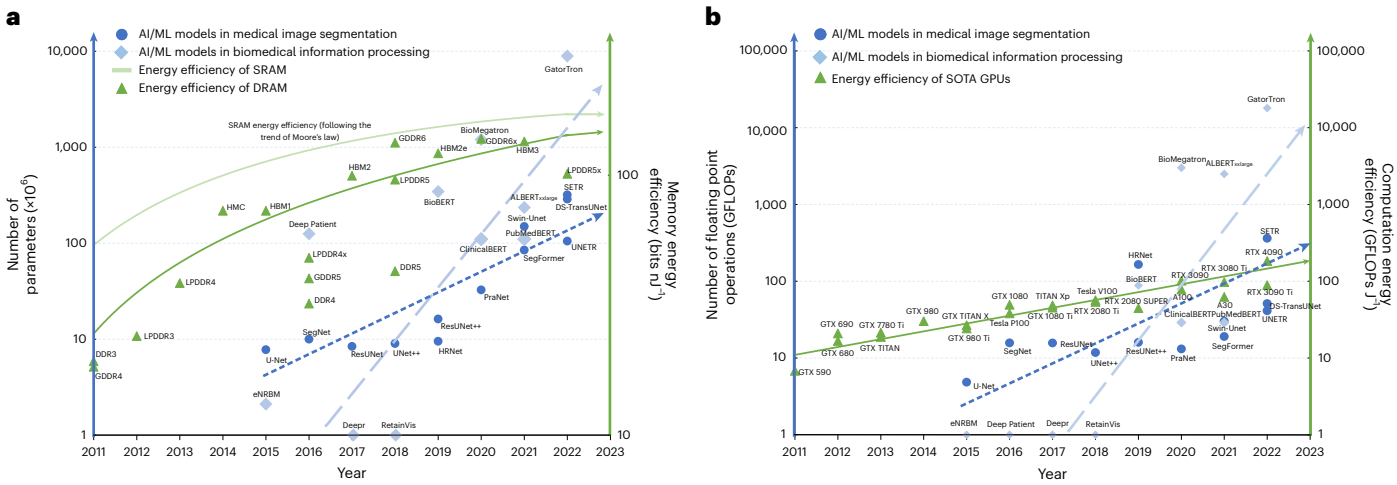

Fig. 1 | Unsustainable energy resource issues caused by the gap between model complexity and efficiency. a, Dashed lines show the trend of the number of parameters versus years for representative AI/ML models in medical image segmentation (from 7 million parameters with U-Net[73] to 200 million parameters with Swin Transformer[74]) and biomedical information processing (from 2 million parameters with autoencoder eNRBM[75] to 8.9 billion for BERT-based GatorTron[76]), respectively. Solid lines show the memory energy efficiency of DRAM and SRAM. Memory efficiency cannot accommodate the exponentially increasing number of memory accesses under an unsustainable energy budget. **b**, Dashed lines show the trend of the number of floating point operations versus years for representative AI/ML models in medical image segmentation (from 4.84 GFLOPs for U-Net[73] to 362.1 GFLOPs for SETR[77]) and biomedical information processing (from 88 GFLOPs for BioBERT[78] to 18,000 GLOPs for GatorTron[76]), respectively. The solid line shows the trend of computation energy efficiency of SOTA GPUs from 2011 to 2022. Computation energy efficiency cannot accommodate the exponentially increasing number of required floating point operations. Only representative AL/ML models are annotated each year. Both *y* axes are in log scale. Data are taken from refs. 11–13,21,73,75–106.

domain expertise in health data labelling has also restricted the development of AI/ML in healthcare.

Although there have been attempts to address certain types of resource constraint in AI/ML for healthcare, the proposed methods have mostly been devised to 'passively' deal with specific resource constraint issues. Few known methods have been systematically designed to proactively tackle resource sustainability issues for general current or future developments. We believe that addressing bottlenecks in algorithm and system design with sustainability awareness and promoting collaborations between academia and industry are key to resolve emerging resource sustainability issues. In this Perspective, we first demonstrate that resource sustainability issues are commonplace in AI/ML methods for healthcare applications, then discuss various algorithms and system approaches that can help alleviate these sustainability issues. Finally, we outline future directions for proactively and prospectively tackling these issues.

Resource sustainability issues

Sustainability is critical for AI/ML applications in healthcare. In previous developments of AI/ML healthcare systems, resource sustainability issues were often neglected, and it was implicitly assumed that there would always be adequate resources for future AI-based health data analysis. However, for AI/ML-based health applications where powerful servers and supercomputers are readily accessible, the current technology evolution trends may continue, but at the cost of unsustainable energy consumption. According to estimates, carbon emissions must be reduced by half over the next ten years to prevent an increase in the frequency of natural disasters[9]. For those applications that have to be conducted on the edge due to security, privacy and/or real-time constraints, the push for much more advanced deep learning technologies will very soon hit the wall in terms of unsustainable energy budget, computing power, network bandwidth, storage capacity and so on. Furthermore, the shortage of healthcare domain expertise and expert time in diagnosing, labelling and cross-validating diagnoses will become more severe as the volume of health data increase. We will focus on these critical resource sustainability issues in AI/ML for healthcare.

Energy consumption outpaces efficiency gains

We have examined data from recent years regarding the capacity of state-of-the-art (SOTA) deep neural networks for healthcare applications, as well as energy efficiency performances of hardware platforms. The results show that an energy sustainability issue exists between the increasing model complexity required for better accuracy and the advances in the hardware architectures needed to accommodate deep models.

Figure 1a shows that the memory energy efficiency of hardware platforms is not keeping up with the increasing sizes of networks, resulting in a growing gap between the two. Representative networks were chosen from two prevalent healthcare tasks (medical image segmentation and biomedical information processing). These two tasks are representative examples of fundamental information extraction processes in the healthcare field, from visual (for example, medical images) and textual (for example, electronic healthcare records) modalities[10]. The dashed lines in Fig. 1a show that the number of parameters of deep neural networks has increased exponentially over the past few years. The solid lines in Fig. 1a present the energy efficiency of static random-access memory (SRAM) and dynamic random-access memory (DRAM). Data moving dominates the energy consumption of hardware platforms, and the total amount of energy consumed by memory is directly proportional to the number of deep model parameters[11]. From 2011 to 2017, the energy efficiency of DRAM, including double data rate, low-power double data rate, graphics double data rate and three-dimensional (3D) DRAM, increased exponentially, according to Moore's law-based complementary metal–oxide–semiconductor scaling[11,12]. Since 2017, DRAM has not experienced a dramatic improvement in terms of energy efficiency. The energy efficiency trend of SRAM is typically bounded by Moore's law, because SRAM is realized with complementary metal–oxide–semiconductor transistors[11]. The trend lines in Fig. 1a show that improvements in memory energy efficiency cannot keep up with the increasing sizes of deep models in healthcare. Accordingly, with the rising memory energy demand, sustainability issues will arise by exceeding the limited energy budget for network inference and training.

https://doi.org/10.1038/s42256-023-00670-0

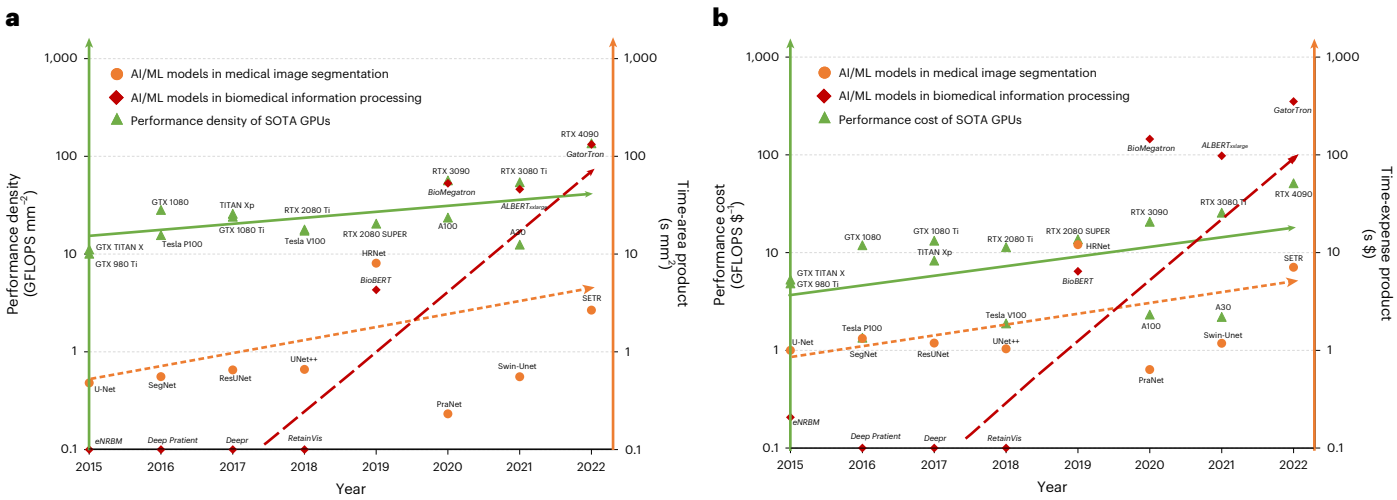

Fig. 2 | Unsustainable computing power resource issues caused by the gap between model complexity and computing capacity. a, The solid line shows the trend of the increasing performance density of leading GPUs from 2015 to 2022. Dashed lines show the trend of the time and chip area product of the SOTA AI/ML model inference conducted on the leading GPU in medical image segmentation (orange) and biomedical information processing (red), respectively. An exponential increase in computing time or chip area over time is needed for both tasks over the past five years. **b**, The solid line shows the trend of the increasing performance cost of the leading GPUs from 2015 to 2022. Dashed lines show the trend of the time and expense product of the SOTA AI/ML model inference conducted on the two healthcare tasks. An exponential increase in computing time or expense is needed for both tasks over the past five years. Both plots show that computing power cannot keep up with the increasing computational demand (in terms of chip area, expense and time) of AI/ML models in healthcare. Only the leading GPUs on desktop or servers for each year are selected. The *y* axis is in log scale. Data are taken from refs. 11–13,21,73,75–107.

An energy sustainability issue also arises in relation to the exponentially increasing computational demands of AI/ML models in comparison to the computation energy efficiency, which shows relatively slow improvement (Fig. 1b). The dashed lines in Fig. 1b show that the number of giga floating point operations (GFLOPs) of networks in medical image segmentation and biomedical information processing experienced an exponential increase from 2015 to 2022. A corresponding trend is shown in the advancement of hardware efficiency in computation (solid line in Fig. 1b). The efficiency of the float32 precision format of SOTA graphics processing units (GPUs) in desktops and servers from each year is reported. Units of GFLOPs per joule are utilized as a measure of hardware computation energy efficiency. Figure 1b shows that the trend of computation energy efficiency improvement cannot keep up with the increasing trend of the computational complexity of deep models in either the medical image segmentation or biomedical information processing tasks. Although the energy consumption of a single inference may not grow at the same rate as the number of model parameters, thanks to systematic optimization[13], the total energy consumption of frequent and multiple inferences in healthcare tasks remains unsustainable with the limited energy budgets available.

Furthermore, training large AI models requires massive amounts of energy, substantially higher than for simply performing inference[14–16]. For example, full training of a BERT model consumes ~103,593 kWh of electricity[9]. The use and deployment of AI models can also contribute to carbon emissions through the use of energy-intensive hardware, such as GPUs and tensor processing units (TPUs). To obtain a model architecture for better accuracy performance, running a neural architecture search over all BERT model parameters would consume even more electricity (656,347 kWh, which is as much as six times that for full training of a BERT model[17,18]) and cause more carbon emissions (626,155 pounds CO_2 emissions, as much as the lifetime emissions of five cars[9,18,19]).

As shown in Fig. 1, it is clear that the energy efficiency of leading hardware (that is, memory and computing capacity) cannot keep up with the deep model complexity required for better accuracy and broader usage. Simply increasing the number of hardware platforms to accommodate more complex deep models is not a sustainable solution due to the limited energy budget.

Computing power demand outpaces performance density gains

The use of AI/ML in healthcare requires large amounts of computing power, and this can have substantial sustainability implications. The rapid development of AI/ML models, with their growing computational demand, has led to a constant need for more powerful hardware, such as GPUs, to accommodate this increasing demand for computing power. As shown in Fig. 1b, the number of floating point operations performed by one inference increases from 4.84 GFLOPs for U-Net to 18,000 GFLOPs for GatorTron. The required computing power further increases in scale with the increasing number of model inferences conducted for various tasks. Training AI/ML models is an even more computing-intensive task. For example, it requires 1.9×10^5 peta FLOPs (PFLOPs) to train a BERT model[20] for biomedical information processing and 3.1×10^8 PFLOPs to train the GPT-3 model utilized for interactive computer-aided diagnosis[21,22]. A lack of computing power would be an obstacle to training an AI/ML model or conducting model inference, especially given the current pursuit of large models, further hindering technological progress.

The rapidly growing computing power demand for AI/ML model development and deployment has led to an unsustainable situation where simply increasing the number of hardware platforms cannot address the issue of a limited budget, as shown in Fig. 2. Figure 2a depicts how the performance density (the computing capacity per unit area, in giga floating point operations per second (GFLOPS) per square millimetre) of leading GPUs has gradually improved over the past few years. From 2015 to 2022, there has been a great improvement in the performance density, which improves by around tenfold, with the process size scaling down. However, the growth rate still cannot catch up with the growth rate of AI/ML model computational complexity. This has resulted in an exponentially increasing computation density demand (that is, the time and chip area product to complete one model inference on the leading GPU with the highest performance density), as indicated by dashed lines in Fig. 2a. The computation density demand is approximately obtained by dividing the total number of GFLOPs of a model by the performance density of the leading GPU (GFLOPS mm^{-2}) in the same year the model was developed.

https://doi.org/10.1038/s42256-023-00670-0

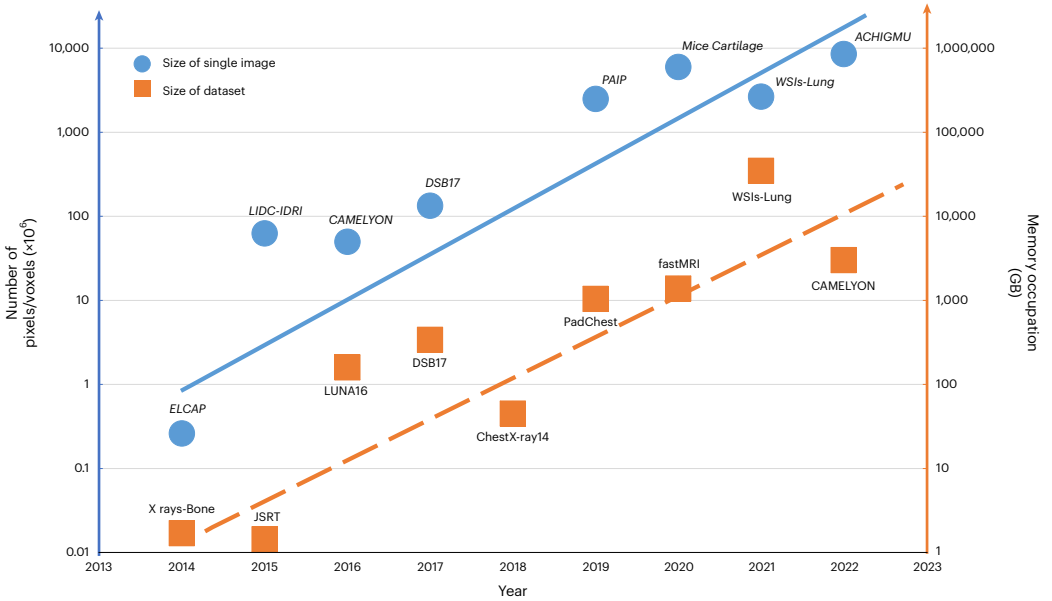

Fig. 3 | Unsustainable storage issue caused by the increasing data volume and resolution of medical data utilized in AI/ML model development. The solid line shows the trend of the number of pixels/voxels in a single medical image used in model inference. The dashed line shows the trend of the volume of a medical image dataset utilized for model training. The explosively increasing sizes of a single medical image and image dataset challenge storage infrastructure. Only representative medical image datasets are annotated each year. The *y* axis is in log scale. Data are taken from refs. 1,24–26,108–124.

The same applies for performance cost (the computing capacity per unit expense in GLOPS per US dollar), as shown in Fig. 2b. The price of each year's leading GPU is set to match the buying power in January 2023 by considering inflation, using the tool provided by the US Bureau of Labor Statistics[23]. As the performance cost of leading GPUs increases, the computation expense demand (the time and expense product to complete one model inference on the leading GPU with the lowest performance cost) grows exponentially in both healthcare tasks, as indicated by dashed lines. The computation expense demand is approximately obtained by dividing the total number of GFLOPs of a model by the performance expense of the leading GPU (GFLOPS $^{-1}$) in the same year the model was developed.

Accordingly, there is a sustainability issue in that the computing power of leading hardware platforms cannot keep up with the rapidly increasing computation complexity of AI/ML models in healthcare. Simply increasing the number of GPUs, performance density or expense budget is not a sustainable solution. Furthermore, computing hardware platforms are generally bounded by Moore's law[12], and the computing capacity of SOTA GPUs may not increase with an exponential growth rate when Moore's law ends as we approach the limit of process size[11].

Data storage outpaces infrastructure development

Along with the increasing model complexity, biomedical data involved in AI/ML model development have also experienced sharp growth in terms of volume and resolution. We now examine recent medical image datasets utilized in deep model training. The results show that the storage sustainability issue arises in the current storage infrastructure of research centres and medical institutes.

Figure 3 demonstrates a dramatic increase in storage requirements for medical image data. In particular, the solid line in Fig. 3 shows the number of pixels/voxels in a single medical image from representative datasets used in deep model inference. Recent advances in biomedical image acquisition technologies have led to an upscale in image resolution. For example, the size of three-dimensional (3D) micro-computed tomography (micro-CT) images of mouse skull cartilages and bones has increased from 200 × 5,122 to 1,500 × 20,002 voxels, requiring around 12 GB to store a single 3D image[1]. As for the resolution of 2D images, in the ACHIGMU (Affiliated Cancer Hospital and Institute of Guangzhou Medical University) dataset[24], a single histopathological image is scanned at ×20 magnification with an average size of 68,096 × 125,440 pixels, which requires ~16.3 GB for a single image. In addition, data volumes utilized in AI/ML healthcare applications have also become overwhelming. The dashed line in Fig. 3 demonstrates the volumes of representative datasets utilized in deep model training, showing that the memory capacity required to store datasets (utilized in deep model training) has increased from 1.7 GB (for the dataset X rays-Bone[25]) to 34,779 GB (for the dataset WSIs-Lung[26]). With increasing model sizes, the required data volume for properly training a deep model is also increasing. The multimodal imaging of proteomics[27], cell segmentation[28], super high-resolution 3D imaging[1–3] and the enormous amount of CT scan images[29] are further pressing the storage resources needed for cloud computing infrastructures. For example, the cancer prognostication described in ref. 30 proposes a deep learning-based multimodal fusion algorithm that uses both whole slice images and molecular profile features (that is, mutation status, copy-number variation, RNA sequencing expression), which amounts to over 7,000 GB in terms of data volume.

Furthermore, existing storage infrastructures may not be able to keep up with the increasing storage capacity required by medical image datasets. Health institutes and universities often provide two tiers of storage: locally maintained on-campus network-attached storage and storage infrastructure provided and maintained by cloud storage vendors[31]. The cost of on-campus network-attached storage can be up to US$3,000 per terabyte per year, and the cost of storage on high-performance computation cluster facilities can be up to US$3,600 per terabyte per year[31]. What is worse, universities and institutes do not generally offer more than a few terabytes of storage. With the increasing data volume and the total number of datasets for model training, the existing storage infrastructure is rather impractical and unsustainable in terms of memory capacity and budget. The limited storage capacity will greatly obstruct AI/ML development in healthcare. When it comes to cloud storage, integrating commercial cloud storage infrastructure into the AI/ML development process is not

a simple task due to concerns regarding the transparency, privacy and security of sensitive health data[7]. Although commercial cloud vendors can provide sufficient storage capacity, there is a risk of exposure and misuse of sensitive data without additional legislation and regulation in place for the third-party cloud storage platforms used for AI in healthcare[7,32]. As such, as of today, the utilization of commercial cloud storage for healthcare data is rare.

Data transmission outpaces network infrastructure development

The integration of AI/ML into healthcare systems generates large amounts of data and AI/ML models that need to be transmitted between devices and servers for a range of purposes, such as remote diagnosis[33] and collaborative learning[34]. The transmission efficiency of network communication infrastructure is particularly important in healthcare application scenarios where real-time data transmission is critical for patient care. Bandwidth limitations can result in delays or loss of data, which can have serious consequences for point-of-care patient outcomes.

This process thus comes with a substantial sustainability challenge related to network infrastructure. The network infrastructure in healthcare facilities is not designed to handle the massive amounts of data that deep learning requires[35]. A lack of high-speed data transmission required by real-time decision-making is putting further pressure on the current networking infrastructure, resulting in issues such as excessive latency.

Accordingly, the sustainability of network communication infrastructure is a critical issue in AI for healthcare. It requires immediate attention and efforts towards developing sustainable solutions to establish efficient communication while ensuring the effective processing and analysis of healthcare data using AI/ML.

Data preparation effort outpaces experts' load

In healthcare applications in particular, data preparation (that is, annotation and label verification) is a critical process to guarantee the performance of an AI/ML model and establish the trust of users and doctors. Often under-appreciated by AI/ML developers, data annotation and verification for AI/ML model training is a labour-intensive and time-consuming task, can take up to several hours for a single image, and results in expert load sustainability issues, as shown in Fig. 4.

As AI/ML models with an increasing number of parameters are adopted in healthcare applications, more images are required for model training to achieve better accuracy. As shown in Fig. 4a, the number of X-ray, CT and magnetic resonance imaging (MRI) images utilized in AI/ML model training is exponentially increasing. Alongside the explosively increasing number and size of medical images utilized in AI/ML model training, the time spent by domain experts on data preparation of a medical image has remained constant for years as a result of mature diagnostic procedures[36,37]. For a sophisticated case, a domain expert will take a long time to complete a specific diagnostic image analysis for interpretation and annotation verification. As shown in Fig. 4b, detailed manual contouring of COVID-19 infection regions on one chest CT scan can take 187 ± 38.5 min (ref. 38), and labelling neuroblastic tumours manually requires a mean time of 56 min per case in MRI images[39]. Indeed, current qualified domain experts cannot keep up with the data preparation demand. The orange dashed line with diamond-like points in Fig. 4a demonstrates a fairly small increase in the number of registered radiologists in the USA, from 36,000 to 41,000, in the past ten years[40]. This slow growth is due to the long period of training time required to become a qualified domain expert.

The total load of all domain experts devoted to labelling and verification cannot keep up with the explosively increasing number of medical images needed for training. This could lead to the waste of the abundant medical data made available by advances in healthcare technologies, which could have been effectively used to boost very large-scale deep learning models. Limited domain expertise can also potentially lead to biased models if only a few people contribute to training dataset preparation.

Resource sustainability issues considered from lifecycle perspective

It is worth noting that the costs and benefits associated with developing and deploying AI systems are not restricted to a single stage such as the training or testing phase. Rather, they span from the design and development stages all the way through to adaptation and implementation, and may continue to evolve with each subsequent use or contribution. To assess the resource impact of an AI system, it is important to consider its entire lifecycle, taking into account the different stages and their associated costs. For example, during the early stages of model development, resources such as energy, computing power, domain expertise and time may be heavily invested in failed experiments and testing with various libraries. As the model evolves into a prototype, testing, software, hardware and computational resources become increasingly focused on reliability, stability and generalizability. Once the system is deployed to users, additional resource costs are incurred to enable efficient human–machine collaboration, potential domain adaptation and/or model upgrade. These may further exaggerate the sustainability issues and require a holistic approach to consider the resource sustainability of an AI system over its entire lifecycle.

Resource sustainability issues considered in edge computing

The application of edge computing is gaining attention in AI for healthcare as it allows for data processing and analysis to be done closer to the source of data generation (for example, in implantable[5] or wearable devices[41]) or at the point of care, rather than sending it to a centralized cloud server. This approach offers several benefits, including reduced latency, improved data privacy and security, and more efficient use of network bandwidth. However, resource sustainability issues would become more acute due to the stringent resource capability of these edge platforms. For example, the existing storage infrastructure of edge computing systems generally cannot accommodate the enormous volume of healthcare data for on-device learning. On the other hand, frequent communication to acquire data would increase the burden on network bandwidth and edge device power consumption, and pose potential security risks. The computing power provided by these edge computing devices is also inherently limited due to physical size constraints and limited energy budgets. Therefore, as the prevalent AI models in healthcare sharply increase in size and require more data to feed, training the model or even performing inference locally will become prohibitively expensive in the near future.

In summary, sustainability issues are critical for AI in healthcare in relation to various types of resource. AI, on the other hand, has the potential to have a substantial positive impact on carbon emission reduction by enabling the electronic delivery of targeted clinical expertise needed in remote areas or underdeveloped regions so that healthcare providers can reduce travel. For example, AI-CHD[42] has recently demonstrated its power in providing automatic diagnoses of congenital heart disease, which previously would have required radiologists with very specialized training.

Approaches to solving resource scarcity

The current solutions for addressing resource constraints in AI/ML for healthcare span the perspectives of algorithms and system optimization.

Algorithm perspective

Domain adaptation to reduce resource consumption. In recent years, domain adaptation techniques have been applied to pretrained

https://doi.org/10.1038/s42256-023-00670-0

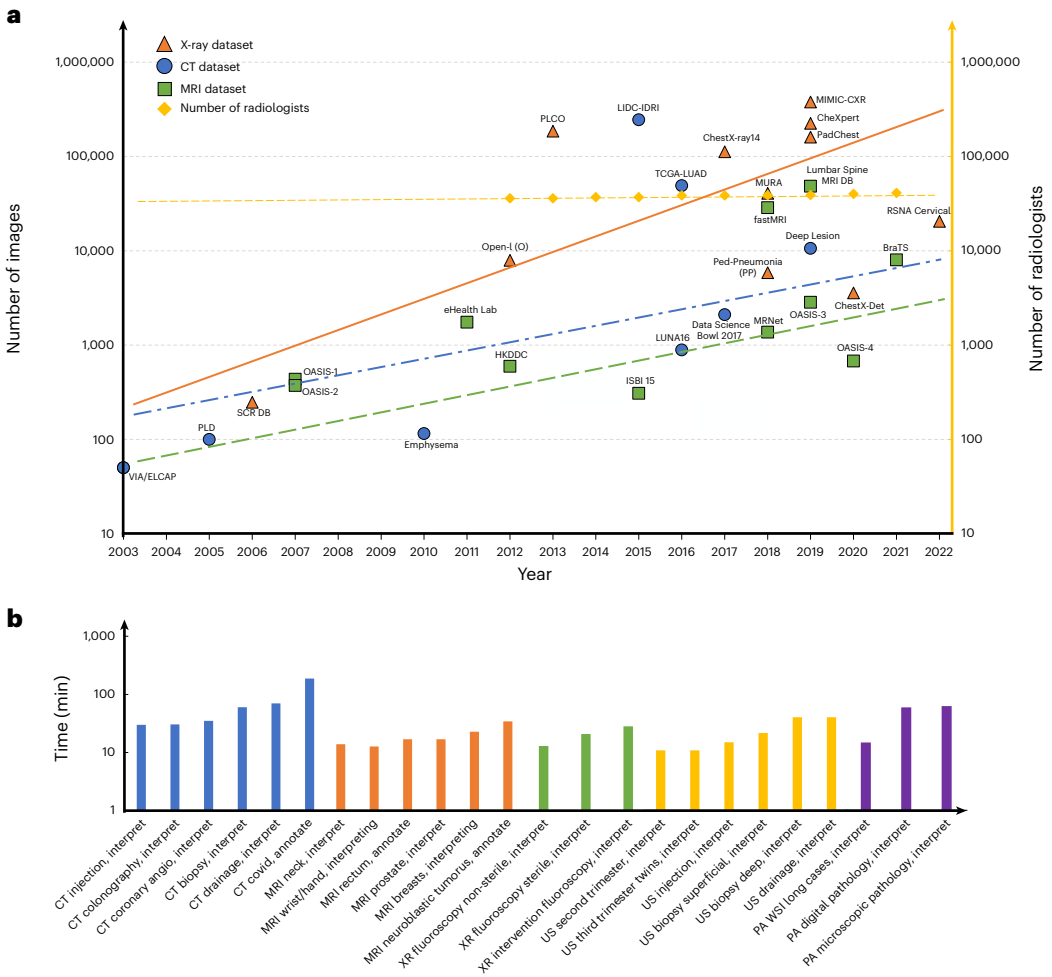

Fig. 4 | Unsustainable expert load on health data preparation for AI/ML model training. a, The unsustainable expert load issue caused by the gap between the increasing number of acquired radiological images and the radiologists required for annotation and interpretation. The solid orange, dash-dotted blue and dashed green lines respectively show the numbers of medical images in datasets of X-ray, CT and MRI modalities utilized in AI/ML model training over the past ten years. The yellow dashed line shows the number of registered radiologists in the USA. To catch up with the required annotation of medical images, training more qualified radiologists in a short period is not feasible. Only representative medical image datasets are annotated each year. The *y* axis is in log scale. Data are taken from refs. 40,108,110,113,116,119–122,125–155. **b**, Time taken by healthcare domain experts in the diagnosis and annotation of medical images under different modalities. The histograms show the time (in minutes) spent by the domain expert for CT, MRI, X-ray (XR), ultrasound (US) and pathology (PA) images, respectively. Most diagnostic and annotation tasks require more than 10 min for each image. Data are taken from refs. 36–39,156–159.

models based on abundant health data to adapt the networks to downstream tasks. In this way, a pretrained model trained for one domain can be reused in another domain, with minimal energy, computing power and domain expertise costs for the model development process. As a result, an extensively pretrained network will require less training than if trained from scratch for a new healthcare application. Hence, the amount of training and the number of labelled training samples will be reduced, and will cost less in terms of different resources. In the medical image domain, the pretrain-finetune paradigm has been applied to electron microscopy data[4] and mass spectroscopy[43] by fine-tuning the pretrained model with a limited amount of task-specific data to adapt to the new domain. A novel method that learns to ignore the scanner-related features present in MRI images when performing domain adaptation for a multi-site MRI dataset has been presented[44]. In the physiological signal domain, the authors of ref. 45 proposed a model personalization method based on meta-learning and fine-tuning for personalized arrhythmia detection and human activities recognition. An on-device model personalization based on the generative adversarial network for ventricular arrhythmia detection is proposed in ref. 46.

Model compression and architecture search to save resources. There are algorithms designed specifically to find tiny networks, thus using the least amount of parameters so as to reduce energy consumption and storage in healthcare applications. For example, for medical image segmentation tasks, a compressed CeNN framework[47] has been devised to perform incremental quantization and early exit, which substantially reduces computational demands while maintaining acceptable performance with a field-programmable gate array. A fairness-aware pruning strategy, FairPrune[48], has been proposed to prune the network weights while maintaining the fairness and accuracy of medical image classification, and a network quantization method was proposed to reduce the model size to fit the constrained resources of edge devices for medical image segmentation[49]. There are also works in resource-aware neural architecture search to find the best-fit tiny network architecture for real-time electrocardiogram reconstruction[50].

Self-supervised learning paradigm to reduce expertise involved. There are algorithms designed for self- and semi-supervised learning that reduce the dependency on expert annotation for model training. For example, a self-supervised learning strategy based on context

restoration has been proposed to better exploit unlabelled images of CT scans and 2D ultrasound results[51]. A self-supervised learning method for MRI images was devised in which the network is trained using a contrastive loss on whether the scan is from the same person (that is, longitudinal) or not, together with a classification loss on predicting the level of vertebral bodies[52]. A novel multi-instance contrastive learning[53] method that used multiple images of the underlying pathology was proposed to improve classification accuracy on dermatology and chest X-ray images. ConVIRT was proposed in ref. 54 to learn medical visual representations through naturally occurring paired descriptive text in an unsupervised way, which required only 10% of the labelled training data compared with the supervised learning approach, while achieving better performance. As for the physiological signal domain, the authors in ref. 55 proposed a contrastive learning approach, CLOCS, that enabled the model to learn representations across space, time and patients and achieved comparable atrial fibrillation detection accuracy but with 25% the labelling required by supervised training approaches. An intra–inter subject self-supervised learning model was presented in ref. 56 to effectively learn from intra–inter subject differences and achieved 10% improvement over supervised training, but required only 1% of the labelled data.

System perspective

Federated learning to balance workload and resource consumption. From the system perspective, federated learning has become prevalent in balancing the resources required in training by offloading the computational workload from a single central server to multiple servers or other edge devices with computing system development. Federated learning is a paradigm that addresses the problem of data governance by training a deep neural network collaboratively without uploading data to a centralized server[34]. The training process of federated learning can occur locally at each participating site's end, and only model characteristics (such as parameters and gradients) are transferred. In this way, the energy consumption and computational workload of deep learning-based analysis on overwhelming healthcare data could be properly managed by distributing the tasks among all participants, which alleviates the sustainability issues on the central server. Cross-institute federated learning is an emerging technique in the healthcare industry that enables multiple hospitals or organizations to collaboratively train a global model without aggregating the raw data into a central server. This technique could preserve data privacy and further balance the workload of data processing by avoiding the conventional centralized training paradigm. This learning paradigm in computing system development has been adopted in the medical domain on multi-site institutions' servers[34], and the global model being collaboratively trained by ten institutions reached 99% of the model quality achieved with centralized data. The work in ref. 57 shows that model training tasks can be offloaded across multiple institutions with real-world private clinical data while generating the model, demonstrating improved generalization. Clustered federated learning[58], which offloads training tasks from servers to local edge devices, was proposed for an automatic COVID-19 diagnosis and resulted in performances comparable to those of the central baseline on X-ray and ultrasound datasets. The federated learning framework FedHealth[41] was introduced to include wearable devices in collaborative model training to build personalized models for Parkinson's disease auxiliary diagnosis.

Decentralized storage system to mitigate the sustainability issue while preserving data privacy. To tackle the storage constraint, decentralized storage systems for medical data have been developed for healthcare applications. With advances in medical image acquisition techniques, image sizes have dramatically increased. Meanwhile, the amount of medical image data is growing quickly as the utilization of massive medical images for clinical decision support has grown. As a result, there is a big demand to provide highly scalable data management and sharing in AI/ML for healthcare. Decentralized storage systems have been applied in the field to substantially reduce the storage overhead in the centralized cloud system. For example, a scalable and flexible decentralized medical image management system based on DCMRL/XMLStore and DCMDocStore has been devised[59]. A healthcare data-sharing scheme has also been proposed to enable efficient and secure medical data sharing via blockchain for a decentralized storage system[60].

Outlook

Current attempts to tackle resource constraints in AI/ML for healthcare propose algorithm- and system-perspective optimizations. However, there is still a lack of sustainability awareness when developing systems and algorithms in AI/ML for healthcare. In this section we present an outlook for potential solutions for resource sustainability issues.

Cost-aware cross-layer co-design for AI/ML healthcare systems

It is well known that hardware and the performance of AI models, including their accuracy[61], confidence[62] and even security[63], are entangled. To enable efficient design space exploration, it is critical to develop a cross-layer co-exploration framework that spans hardware, algorithms and models to identify the best configurations of resource-sustainable solutions for healthcare applications[64–67]. In addition, a resource cost model is needed that accurately estimates or predicts resource consumption in terms of different combinations of hardware/software resources, neural network components and algorithm design options for each specific AI/ML-based healthcare application[68]. Moreover, the prediction model should also be able to predict the cost to maintain sufficient resource availability in terms of energy, computing power, storage, networks, algorithms, domain expertise and data volumes for model upgrades or possible domain adaptation from a lifecycle perspective. AI/ML models can potentially help in making accurate predictions to reduce waste and carbon emissions. With the cost prediction model and co-design framework, designers can customize the optimization goals in terms of a newly updated infrastructure, specific resource budget and other optimization criteria for the entire lifecycle.

Consensus-based distributed learning

A novel consensus-based distributed learning framework should be developed to fully utilize the storage resources of existing and future computing infrastructures. Existing distributed learning paradigms, such as cross-institute FL in the healthcare field, still heavily rely on tedious verification and authorization before the actual learning process. They also largely rely on the data centre infrastructure for data storage and sharing. Consensus-based distributed learning can be a future direction in the field by incorporating a large number (for example, millions) of Internet-of-Things devices, edge servers and cloud centres altogether into the learning infrastructure, which fully utilizes the storage capacity and computational power of these devices based on a smart consensus strategy to protect data privacy while enabling fast data sharing and learning. Effective consensus mechanisms must be developed to be scalable, flexible and lightweight so as to fit heterogeneous system structures and requirements into distributed learning.

Stable infrastructures with AI-enhanced resource allocation

Out of regulatory considerations on health data security and privacy, instead of pushing towards better use of general-purpose commercial cloud storage, an alternative, perhaps more viable approach, is to establish dedicated healthcare AI infrastructures that maintain full compliance with government regulations, which may evolve over time[7]. These infrastructures can be fully funded by governments or private sectors, and will secure the preservation of data and support the further development of algorithms. Additionally, AI-based techniques can be

applied to optimize resource allocation in terms of storage, computing power and energy efficiency in these infrastructures.

Interpretable self-supervised learning

The conflict between the increasing demand for domain expertise and the increasing volume of healthcare data will become more severe. Self-supervised learning can be a key approach to addressing the sustainability issue in domain expertise. Currently, there is still a big barrier to obtaining the trust of providers, doctors and patients due to the blackbox nature of deep learning-based diagnosis. A future direction could be self-supervised learning with interpretability/explainability. Interpretable self-supervised learning algorithms can substantially alleviate the expertise sustainability issue by extracting clinically useful features, explaining analysis/diagnosis results with human-interpretable evidence, and training models without massive expert-labelled data, while further gaining the trust of domain experts, providers and patients. Future steps of interpretable self-supervised learning in healthcare applications could include (1) full exploration and usage of core medical features (for example, human-understandable features) by self-supervised learning algorithms to improve performance and explainability; (2) human-interpretable inference and reasoning; (3) evolution of deep models with closed interaction with and input from human experts (for example, human-in-the-loop for error correction and function enhancement).

Few-shot learning on large language model for automatic labelling and annotation

Advancements in AI/ML algorithms have made it possible to automate the process of data labelling. Using few-shot learning techniques, large language models can be fine-tuned to automatically label meta-features in medical images or biomedical information text. This is achieved by providing prompts and a series of labelled examples, which allow the AI system to learn to recognize patterns and features in the images or text and label new instances accordingly. Automated labelling can substantially speed up the process of annotating large datasets, while also increasing the accuracy and consistency of the labels. Although automation can greatly reduce the time and effort required for annotation, it is important to validate and cross-check automated methods with manual annotations to ensure the highest level of accuracy.

The past few years have witnessed great success in large model design, exemplified by models such as ChatGPT, with its 175 billion parameters[69], PaLM with its 540 billion parameters[70] and Visual ChatGPT[71]. It is anticipated that, in the very near future, large AI models will also revolutionize healthcare, enabling improved performance in a broader spectrum of tasks beyond what we have focused on in this Perspective, such as evidence-based medicine and personal health advisors. The explosive increase in model complexity will, however, further stress the already existing resource sustainability issues. With the deep learning stack developed for predictable scalability in GPT-4[72], the performance of a small model with a limited number of parameters on a healthcare-specific task can be utilized to precisely predict the performance of a standard-scale model, which will substantially alleviate the resource sustainability issue. It is also important to investigate methods that can better support not only model generalization but also specialization, through weakly supervised or unsupervised on-device learning and model personalization.

References

1. Zheng, H. et al. Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* 802–812 (Springer, 2020).
2. Perrine, S. M. M. et al. A dysmorphic mouse model reveals developmental interactions of chondrocranium and dermatocranium. *eLife* **11**, e76653 (2022).
3. Pitirri, M. K. et al. Meckel's cartilage in mandibular development and dysmorphogenesis. *Front. Genet* **13**, 871927 (2022).
4. Nightingale, L. et al. Automatic instance segmentation of mitochondria in electron microscopy data. Preprint *bioRxiv* https://doi.org/10.1101/2021.05.24.444785 (2021).
5. Jia, Z. et al. Learning to learn personalized neural network for ventricular arrhythmias detection on intracardiac EGMs. In *Proc. International Joint Conference on Artificial Intelligence* 2606–2613 (2021).
6. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
7. Lekadir, K., Quaglio, G., Garmendia, A. T. & Gallin, C. *Artificial Intelligence in Healthcare. Applications, Risks, and Ethical and Societal Impacts* (European Parliamentary Research Service, 2022).
8. Banerjee, M. et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. *BMC Med. Educ.* **21**, 429 (2021).
9. Dodge, J. et al. Measuring the carbon intensity of AI in cloud instances. In *Proc. ACM Conference on Fairness, Accountability and Transparency* 1877–1894 (ACM, 2022).
10. Bayoudh, K., Knani, R., Hamdaoui, F. & Mtibaa, A. A survey on deep multimodal learning for computer vision: advances, trends, applications and datasets. *Visual Comput.* **38**, 2939–2970 (2021).
11. Xu, X. et al. Scaling for edge inference of deep neural networks. *Nat. Electron.* **1**, 216–222 (2018).
12. Sutter, H. et al. The free lunch is over: a fundamental turn toward concurrency in software. *Dr. Dobb's J.* **30**, 202–210 (2005).
13. Desislavov, R., Martínez-Plumed, F. & Hernández-Orallo, J. Compute and energy consumption trends in deep learning inference. Preprint at https://arxiv.org/abs/2109.05472 (2021).
14. Hestness, J. et al. Deep learning scaling is predictable, empirically. Preprint at https://arxiv.org/abs/1712.00409 (2017).
15. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
16. Henighan, T. et al. Scaling laws for autoregressive generative modeling. Preprint at https://arxiv.org/abs/2010.14701 (2020).
17. Jassim, H. S., Lu, W. & Olofsson, T. Predicting energy consumption and CO_2 emissions of excavators in earthwork operations: an artificial neural network model. *Sustainability* **9**, 1257 (2017).
18. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 3645–3650 (Association for Computational Linguistics, 2019).
19. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 86 (2021).
20. Clark, K., Luong, M.-T., Le, Q. V. & Manning, C. D. Electra: pre-training text encoders as discriminators rather than generators. Preprint at https://arxiv.org/abs/2003.10555 (2020).
21. Gholami, A., Kim, S. & Yao, Z. Memory footprint and FLOPs for SOTA models in CV/NLP/Speech https://github.com/amirgholami/ai_and_memory_wall (2020).
22. Wang, S., Zhao, Z., Ouyang, X., Wang, Q. & Shen, D. ChatCAD: interactive computer-aided diagnosis on medical image using large language models. Preprint at https://arxiv.org/abs/2302.07257 (2023).
23. CPI inflation calculator. https://www.bls.gov/data/inflation_calculator.htm (2023).

24. Xu, Y. et al. Computer-aided detection and prognosis of colorectal cancer on whole slide images using dual resolution deep learning. *J. Cancer Res. Clin. Oncol.* **149**, 91–101 (2022).
25. Cernazanu-Glavan, C. & Holban, S. Segmentation of bone structure in X-ray images using convolutional neural network. *Adv. Electr. Comput. Eng* **13**, 87–94 (2013).
26. Chen, C.-L. et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat. Commun.* **12**, 1193 (2021).
27. Mund, A. et al. Deep visual proteomics defines single-cell identity and heterogeneity. *Nat. Biotechnol.* **40**, 1231–1240 (2022).
28. Ghahremani, P. et al. Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nat. Mach. Intell.* **4**, 401–412 (2022).
29. Jin, C.-B. et al. Deep CT to MR synthesis using paired and unpaired data. *Sensors* **19**, 2361 (2019).
30. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878 (2022).
31. Andreev, A., Morrell, T., Briney, K., Gesing, S. & Manor, U. Biologists need modern data infrastructure on campus. Preprint at https://arxiv.org/abs/2108.07631 (2021).
32. Gourraud, P.-A. & Simon, F. Differences between Europe and the United States on AI/digital policy: comment response to roundtable discussion on AI. *Gender Genome* **4**, 1–18 (2020).
33. Ghosh, A., Raha, A. & Mukherjee, A. Energy-efficient IoT-health monitoring system using approximate computing. *Internet Things* **9**, 100166 (2020).
34. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
35. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng* **6**, 1330–1345 (2022).
36. Pitman, A., Cowan, I. A., Floyd, R. A. & Munro, P. L. Measuring radiologist workload: progressing from RVUs to study ascribable times. *J. Med. Imag. Rad. Oncol.* **62**, 605–618 (2018).
37. Dora, J. M., Torres, F. S., Gerchman, M. & Fogliatto, F. S. Development of a local relative value unit to measure radiologists' computed tomography reporting workload. *J. Med. Imag. Rad. Oncol.* **60**, 714–719 (2016).
38. Ghayvat, H. et al. AI-enabled radiologist in the loop: novel AI-based framework to augment radiologist performance for COVID-19 chest CT medical image annotation and classification from pneumonia. *Neural Comput. Appl.* https://doi.org/10.1007/s00521-022-07055-1 (2022).
39. Veiga-Canuto, D. et al. Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images. *Cancers* **14**, 3648 (2022).
40. *Physician Specialty Data Report;* https://www.aamc.org/data-reports/workforce/report/physician-specialty-data-report (Association of American Medical Colleges, 2022).
41. Chen, Y., Qin, X., Wang, J., Yu, C. & Gao, W. FedHealth: a federated transfer learning framework for wearable healthcare. *IEEE Intell. Syst.* **35**, 83–93 (2020).
42. Xu, X. et al. AI-CHD: an AI-based framework for cost-effective surgical telementoring of congenital heart disease. *Commun. ACM* **64**, 66–74 (2021).
43. Bittremieux, W., May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. *Nat. Methods* **19**, 675–678 (2022).
44. Wolleb, J. et al. Learn to ignore: domain adaptation for multi-site MRI analysis. In *Proc. Medical Image Computing and Computer Assisted Intervention* 725–735 (Springer, 2022).
45. Jia, Z., Shi, Y. & Hu, J. Personalized neural network for patient-specific health monitoring in IoT: a meta-learning approach. In *Proc. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* Vol. 41, 5394–5407 (IEEE, 2022).
46. Jia, Z., Hong, F., Ping, L., Shi, Y. & Hu, J. Enabling on-device model personalization for ventricular arrhythmias detection by generative adversarial networks. In *Proc. ACM/IEEE Design Automation Conference* (*DAC*) 163–168 (IEEE, 2021).
47. Xu, X. et al. Efficient hardware implementation of cellular neural networks with incremental quantization and early exit. *ACM J. Emerg. Technol. Comput. Syst.* **14**, 1–20 (2018).
48. Wu, Y., Zeng, D., Xu, X., Shi, Y. & Hu, J. FairPrune: achieving fairness through pruning for dermatological disease diagnosis. In *Proc. Medical Image Computing and Computer Assisted Intervention: 25th International Conference Part I* 743–753 (Springer, 2022).
49. Zhang, R. & Chung, A. C. MedQ: lossless ultra-low-bit neural network quantization for medical image segmentation. *Med. Image Anal.* **73**, 102200 (2021).
50. Zhang, Y. et al. RT-RCG: neural network and accelerator search towards effective and real-time ECG reconstruction from intracardiac electrograms. *ACM J. Emerg. Technol. Comput. Syst.* **18**, 1–25 (2022).
51. Chen, L. et al. Self-supervised learning for medical image analysis using image context restoration. *Med. Image Anal.* **58**, 101539 (2019).
52. Jamaludin, A., Kadir, T. & Zisserman, A. Self-supervised learning for spinal MRIs. In *Proc. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 294–302 (Springer, 2017).
53. Azizi, S. et al. Big self-supervised models advance medical image classification. In *Proc. IEEE/CVF International Conference on Computer Vision* 3478–3488 (IEEE, 2021).
54. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Proc. Machine Learning for Healthcare Conference* 2–25 (PMLR, 2022).
55. Kiyasseh, D., Zhu, T. & Clifton, D. A. CLOCS: contrastive learning of cardiac signals across space, time and patients. In *Proc. International Conference on Machine Learning* 5606–5615 (PMLR, 2021).
56. Lan, X., Ng, D., Hong, S. & Feng, M. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 36, 4532–4540 (AAAI, 2022).
57. Sarma, K. V. et al. Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Inform. Assoc.* **28**, 1259–1264 (2021).
58. Qayyum, A., Ahmad, K., Ahsan, M. A., Al-Fuqaha, A. & Qadir, J. Collaborative federated learning for healthcare: multi-modal COVID-19 diagnosis at the edge. *IEEE Open J. Comput. Soc.* **3**, 172–184 (2022).
59. Teng, D., Kong, J. & Wang, F. Scalable and flexible management of medical image big data. *Distrib. Parallel Databases* **37**, 235–250 (2019).
60. Shen, B., Guo, J. & Yang, Y. MedChain: efficient healthcare data sharing via blockchain. *Appl. Sci.* **9**, 1207 (2019).
61. Lu, Q., Jiang, W., Xu, X., Shi, Y. & Hu, J. On neural architecture search for resource-constrained hardware platforms. In *Proc. International Conference on Computer-Aided Design;* https://doi.org/10.48550/arXiv.1911.00105 (Association for Computing Machinery, 2019).
62. Ding, Y. et al. Hardware design and the competency awareness of a neural network. *Nat. Electron.* **3**, 514–523 (2020).
63. Bian, S., Jiang, W., Lu, Q., Shi, Y. & Sato, T. NASS: optimizing secure inference via neural architecture search. In *Proc. ECAI 2020 24th European Conference on Artificial Intelligence* 1746–1753 (IOS Press, 2020).

64. Jiang, W. et al. Device-circuit-architecture co-exploration for computing-in-memory neural accelerators. *IEEE Trans. Comput.* **70**, 595–605 (2020).
65. Jiang, W. et al. Hardware/software co-exploration of neural architectures. *IEEE Trans. Comput. Aided Design Integrated Circuits Syst.* **39**, 4805–4815 (2020).
66. Jiang, W., Yang, L., Dasgupta, S., Hu, J. & Shi, Y. Standing on the shoulders of giants: hardware and neural architecture co-search with hot start. *IEEE Trans. Comput. Aided Design Integrated Circuits Syst.* **39**, 4154–4165 (2020).
67. Yang, L. et al. Co-exploration of neural architectures and heterogeneous ASIC accelerator designs targeting multiple tasks. In *Proc. Design Automation Conference* (*DAC*) 1–6 (IEEE, 2020).
68. Cao, Q., Lal, Y. K., Trivedi, H., Balasubramanian, A. & Balasubramanian, N. IrEne: interpretable energy prediction for transformers. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* 2145–2157 (ACL, 2021).
69. Baruffati, A. *Chat GPT Statistics 2023: Trends and the Future Perspectives* https://blog.gitnux.com/chat-gpt-statistics (2023).
70. Narang, S. & Chowdhery, A. *Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance* https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html (2022).
71. Wu, C. et al. Visual ChatGPT: talking, drawing and editing with visual foundation models. Preprint at https://arxiv.org/abs/2303.04671 (2023).
72. Stokes, J. *With GPT-4, OpenAI Is Deliberately Slow Walking To AGI* https://www.piratewires.com/p/openai-slowing-walking-gpt (2023).
73. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
74. Wei, C., Ren, S., Guo, K., Hu, H. & Liang, J. High-resolution Swin transformer for automatic medical image segmentation. *Sensors* **23**, 3420 (2023).
75. Tran, T., Nguyen, T. D., Phung, D. & Venkatesh, S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *J. Biomed. Inform.* **54**, 96–105 (2015).
76. Yang, X. et al. GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. Preprint at https://www.medrxiv.org/content/10.1101/2022.02.27.22271257v2 (2022).
77. Zheng, S. et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 6881–6890 (IEEE, 2021).
78. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
79. Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495 (2017).
80. Zhang, Z., Liu, Q. & Wang, Y. Road extraction by deep residual U-net. *IEEE Geosci. Remote Sensing Lett.* **15**, 749–753 (2018).
81. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imag.* **39**, 1856–1867 (2019).
82. Jha, D. et al. ResUNet++: an advanced architecture for medical image segmentation. In *Proc. 2019 IEEE International Symposium on Multimedia* (*ISM*) 225–2255 (IEEE, 2019).
83. Sun, K., Xiao, B., Liu, D. & Wang, J. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5693–5703 (IEEE, 2019).
84. Fan, D.-P. et al. PraNet: parallel reverse attention network for polyp segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* 263–273 (Springer, 2020).
85. Xie, E. et al. SegFormer: simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **34**, 12077–12090 (2021).
86. Cao, H. et al. Swin-UNet: UNet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022 Proceedings* Part III, 205–218 (Springer Nature, 2023).
87. Lin, A. et al. DS-TransUNet: dual Swin transformer U-Net for medical image segmentation. In *Proc. IEEE Transactions on Instrumentation and Measurement* Vol. 71, 1–15 (IEEE, 2022).
88. Hatamizadeh, A. et al. UNETR: transformers for 3D medical image segmentation. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision* 574–584 (IEEE, 2022).
89. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 (2016).
90. Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. Deepr: a convolutional net for medical records. *IEEE J. Biomed. Health Inform.* **21**, 22–30 (2016).
91. Kwon, B. C. et al. RetainVis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Trans. Visual. Comput. Graph.* **25**, 299–309 (2018).
92. Huang, K., Altosaar, J. & Ranganath, R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. Preprint at https://arxiv.org/abs/1904.05342 (2019).
93. Shin, H.-C. et al. BioMegatron: larger biomedical domain language model. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 4700–4706 (Online Association for Computational Linguistics, 2020).
94. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare* **3**, 1–23 (2021).
95. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 86 (2021).
96. Shamsolmoali, P., Zareapoor, M., Wang, R., Zhou, H. & Yang, J. A novel deep structure U-Net for sea-land segmentation in remote sensing images. *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.* **12**, 3219–3232 (2019).
97. Wang, Z. & Blaschko, M. MRF-UNets: searching UNet with Markov random fields. In *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases* 599–614 (ACM, 2022).
98. Gao, J. et al. AutoBERT-Zero: evolving Bert backbone from scratch. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 36, 10663–10671 (AAAI, 2022).
99. Mutlu, O. Memory scaling: a systems architecture perspective. In *Proc. IEEE International Memory Workshop* 21–25 (IEEE, 2013).
100. Rajagopalan, V. et al. *Using Next-Generation Memory Technologies: DRAM and Beyond HC28-T1* https://www.youtube.com/watch?v=61oZhHwBrh8 (2016).
101. Samsung HBM2E https://semiconductor.samsung.com/dram/hbm/hbm2e-flashbolt/ (2019).
102. Micron GDDR6X https://www.micron.com/products/ultra-bandwidth-solutions/gddr6x (2020).
103. Samsung HBM3 https://semiconductor.samsung.com/dram/hbm/hbm3/ (2021).

104. Talluri, R. *LPDDR5X: Memory Performance that Pushes the Limits of What's Possible* https://www.micron.com/about/blog/2022/february/lpddr5x-memory-performance-that-pushes-the-limits (2022).
105. Samsung LPDDR5X. https://semiconductor.samsung.com/dram/lpddr/lpddr5x/ (2022).
106. Alrowili, S. & Vijay-Shanker, K. BioM-transformers: building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proc. 20th Workshop on Biomedical Language Processing* 221–227 (Association for Computational Linguistics. 2021).
107. GPU specs database https://www.techpowerup.com/gpu-specs/ (2023).
108. Early lung cancer action program (ELCAP) dataset https://www.via.cornell.edu/lungdb.html (2014).
109. Shafiee, M. J. et al. Discovery radiomics via stochasticnet sequencers for cancer detection. Preprint at https://arxiv.org/abs/1511.03361 (2015).
110. Armato, S. G.III et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* **38**, 915–931 (2011).
111. Armato, S. G.III et al. Lung image database consortium: developing a resource for the medical imaging research community. *Radiology* **232**, 739–748 (2004).
112. Litjens, G. et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience* **7**, giy065 (2018).
113. Kuan, K. et al. Deep learning for lung cancer detection: tackling the Kaggle Data Science Bowl 2017 challenge. Preprint at https://arxiv.org/abs/1705.09435 (2017).
114. PAIP 2019: Liver cancer segmentation https://paip2019.grand-challenge.org/ (2019).
115. Ngo, T. A. & Carneiro, G. Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference. In *Proc. IEEE International Conference on Image Processing* 2140–2143 (IEEE, 2015).
116. LUng Nodule Analysis (LUNA) https://luna16.grand-challenge.org/Home/ (2016).
117. Dou, Q., Chen, H., Yu, L., Qin, J. & Heng, P.-A. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Trans. Biomed. Eng.* **64**, 1558–1567 (2016).
118. Venkatesan, N. J., Shin, D. R. & Nam, C. S. Nodule detection with convolutional neural network using Apache Spark and GPU frameworks. *Appl. Sci.* **11**, 2838 (2021).
119. Yan, C., Yao, J., Li, R., Xu, Z. & Huang, J. Weakly supervised deep learning for thoracic disease classification and localization on chest X-rays. In *Proc. ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 103–110 (ACM, 2018).
120. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. PadChest: a large chest X-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **66**, 101797 (2020).
121. Lee, J., Kim, H., Chung, H. & Ye, J. C. Deep learning fast MRI using channel attention in magnitude domain. In *Proc. International Symposium on Biomedical Imaging* 917–920 (IEEE, 2020).
122. Knoll, F. et al. fastMRI: a publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Artif. Intell.* **2**, e190007 (2020).
123. Linmans, J., Elfwing, S., van der Laak, J. & Litjens, G. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Med. Image Anal.* **83**, 102655 (2023).
124. Dandu, R. V. Storage media for computers in radiology. *Ind. J. Radiol. Imag.* **18**, 287–289 (2008).
125. Reeves, A. P. et al. A public image database to support research in computer aided diagnosis. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 3715–3718 (IEEE, 2009).
126. Computed tomography emphysema database https://lauge-soerensen.github.io/emphysema-database/ (2010).
127. TCGA-LUAD collection https://www.cancerimagingarchive.net/collections/tcga-luad/ (2016).
128. DeepLesion dataset https://nihcc.app.box.com/v/DeepLesion/ (2019).
129. SCR database: segmentation in chest radiographs https://www.isi.uu.nl/Research/Databases/SCR/ (2006).
130. Demner-Fushman, D., Antani, S., Simpson, M. & Thoma, G. R. Design and development of a multimodal biomedical information retrieval system. *J. Comput. Sci. Eng.* **6**, 168–177 (2012).
131. Zhu, C. S. et al. The prostate, lung, colorectal and ovarian cancer screening trial and its associated research resource. *J. Natl Cancer Institute* **105**, 1684–1693 (2013).
132. Guendel, S. et al. Learning to recognize abnormalities in chest X-rays with location-aware dense networks. In *Proc. Iberoamerican Congress on Pattern Recognition* 757–765 (Springer, 2018).
133. Rajpurkar, P. et al. MURA dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. In *Proc. Medical Imaging with Deep Learning* (2018).
134. Kermany, D., Zhang, K. & Goldbaum, M. Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images. *Mendeley Data* **3**, 10-17632 (2018).
135. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 33, 590–597 (AAAI, 2019).
136. Johnson, A. E. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
137. ChestX-Det-Dataset https://github.com/Deepwise-AILab/ChestX-Det-Dataset (2020).
138. RSNA cervical spine train image PNG CSFD + CSV https://www.kaggle.com/datasets/saberghaderi/rsna-cervical-spine-train-image-png-csfd?select=RSNA+Cervical+Spine+CSFD (2022).
139. Marcus, D. S. et al. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented and demented older adults. *J. Cogn. Neurosci.* **19**, 1498–1507 (2007).
140. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* **22**, 2677–2684 (2010).
141. LaMontagne, P. J. et al. IC-P-164: OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimers disease. *Alzheimers Dement.* **14**, 138 (2018).
142. Koenig, L. N. et al. Select atrophied regions in Alzheimer disease (SARA): an improved volumetric model for identifying Alzheimer disease dementia. *NeuroImage Clin.* **26**, 102248 (2020).
143. MRI lesion segmentation in multiple sclerosis database http://www.medinfo.cs.ucy.ac.cy/index.php/facilities/32-software/218-datasets (2011).
144. Loizou, C. P. et al. Multiscale amplitude-modulation frequency-modulation (AM–FM) texture analysis of multiple sclerosis in brain MRI images. *IEEE Trans. Inf. Technol. Biomed.* **15**, 119–129 (2010).
145. Samartzis, D., Karppinen, J., Chan, D., Luk, K. D. & Cheung, K. M. The association of lumbar intervertebral disc degeneration on magnetic resonance imaging with body mass index in overweight and obese adults: a population-based study. *Arthritis Rheum.* **64**, 1488–1496 (2012).

146. Kuang, X. et al. Spine-GFlow: a hybrid learning framework for robust multi-tissue segmentation in lumbar MRI without manual annotation. *Comput. Med. Imag. Graph.* **99**, 102091 (2022).
147. Longitudinal multiple sclerosis lesion segmentation challenge https://smart-stats-tools.org/lesion-challenge-2015 (2015).
148. Carass, A. et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage* **148**, 77–102 (2017).
149. MRNet dataset: a knee MRI dataset and competition https://stanfordmlgroup.github.io/competitions/mrnet/ (2018).
150. Kara, A. C. & Hardalaç, F. Detection and classification of knee injuries from MR images using the MRNet dataset with progressively operating deep learning methods. *Mach. Learn. Knowledge Extraction* **3**, 1009–1029 (2021).
151. Lumbar spine MRI dataset https://data.mendeley.com/datasets/k57fr854j2/2 (2019).
152. RSNA-ASNR-MICCAI brain tumor segmentation (BraTS) challenge http://braintumorsegmentation.org/ (2021).
153. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G. & Murphy, K. Deep learning for chest X-ray analysis: a survey. *Med. Image Anal.* **72**, 102125 (2021).
154. Gu, Y. et al. A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning. *Comput. Biol. Med.* **137**, 104806 (2021).
155. Shoeibi, A. et al. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: a review. *Comput. Biol. Med.* **136**, 104697 (2021).
156. Forsberg, D., Rosipko, B. & Sunshine, J. L. Radiologists' variation of time to read across different procedure types. *J. Digit. Imag.* **30**, 86–94 (2017).
157. Randell, R., Ruddle, R. A., Quirke, P., Thomas, R. G. & Treanor, D. Working at the microscope: analysis of the activities involved in diagnostic pathology. *Histopathology* **60**, 504–510 (2012).
158. Vodovnik, A. Diagnostic time in digital pathology: a comparative study on 400 cases. *J. Pathol. Inform.* **7**, 4 (2016).
159. Obaro, A. E., Plumb, A. A., North, M. P., Halligan, S. & Burling, D. N. Computed tomographic colonography: how many and how fast should radiologists report? *Eur. Radiol.* **29**, 5784–5790 (2019).

Acknowledgements

J.C. is funded by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF, Germany) under funding reference 161L0272 and supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia (Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen, MKW NRW).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Xiaowei Xu or Yiyu Shi.

Peer review information *Nature Machine Intelligence* thanks Andrey Andreev and Fernando Martínez-Plumed for their contribution to the peer review of this work.