nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-021-00399-8

Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging

Rohan Shad[1], John P. Cunningham[2], Euan A. Ashley[3,4], Curtis P. Langlotz[4,5] and William Hiesinger[1,4] ✉

The National Institutes of Health in 2018 identified key focus areas for the future of artificial intelligence in medical imaging, creating a foundational roadmap for research in image acquisition, algorithms, data standardization and translatable clinical decision support systems. Among the key issues raised in the report, data availability, the need for novel computing architectures and explainable artificial intelligence algorithms are still relevant, despite the tremendous progress made over the past few years alone. Furthermore, translational goals of data sharing, validation of performance for regulatory approval, generalizability and mitigation of unintended bias must be accounted for early in the development process. In this Perspective, we explore challenges unique to high-dimensional clinical imaging data, in addition to highlighting some of the technical and ethical considerations involved in developing machine learning systems that better represent the high-dimensional nature of many imaging modalities. Furthermore, we argue that methods that attempt to address explainability, uncertainty and bias should be treated as core components of any clinical machine learning system.

Advances in computing power, deep learning architectures and expert labelled datasets have spurred the development of medical imaging artificial intelligence (AI) systems that rival clinical experts[1–8]. Yet it is remarkably challenging to deploy AI systems that assist with even simple clinical tasks[6,8]. Machine learning algorithms that were designed to reduce the time it took for clinically actionable inferences, when deployed in clinics, resulted in patients inadvertently experiencing event greater delays[9]. When taken out of siloed and controlled laboratory environments, end users of AI systems must contend with input quality control and network latency, and must devise ways to integrate these systems within established clinical practice. Some of these early forays into translatable clinical machine learning have shown that designing systems to work seamlessly within established clinical workflows requires substantial integrative efforts at the inception of algorithm development, given the drastically limited opportunities for iteration later at the time of prospective deployment[10]. Extensive open-source machine learning software libraries and advances in computer performance have made it easier for researchers to develop increasingly complex AI systems tailored towards specific clinical problems[11,12]. In addition to moving beyond detecting features diagnostic for disease, the next generation of AI systems must account for systemic biases in training data, intuitively alert end users to the uncertainty inherent in predictions and allow for opportunities to explore and explain the mechanisms by which predictions are made. This Perspective builds on these key priority areas for the acceleration of foundational AI research in medicine. We present an overview of dataset curation nuances and architectural considerations specific to machine learning for high-dimensional medical imaging, along with a discussion of explainability, uncertainty and bias in these systems. In the process, we provide a template for researchers interested in navigating some of the issues and challenges that come with building clinically translatable AI systems[13].

High-dimensional medical imaging data

We anticipate that the availability of high-quality 'AI-ready' annotated medical datasets will continue to lag behind demand for the foreseeable future. Retrospectively assigning clinical ground truth labels requires extensive investment of time from clinical experts, and there are substantial barriers to aggregating multi-institutional data for public release[13]. In addition to 'diagnostic AI' characterized by models trained on hard radiological ground truth labels, there will be demand for 'disease prediction AI' trained on potentially noisier clinical composite outcome targets[8,14–16]. Prospective data collection with standardized protocols for image acquisition and adjudication of clinical ground truth are essential steps towards building massive multicentre imaging datasets with paired clinical outcomes.

Large multicentre imaging datasets engender a multitude of privacy and liability concerns associated with potentially sensitive data embedded in the files. The Digital Imaging and Communication in Medicine (DICOM) standard was designed to capture, store and provide workflow management for medical images, and is nearly universally adopted[17]. Imaging files (stored either as .dcm files or within a nested folder structure) contain both pixel data and associated metadata. A multitude of open-source and proprietary tools can assist with de-identification of DICOM files[13,18]. Back-end hospital informatics frameworks such as the Google Healthcare API also support DICOM de-identification via 'safe lists'—a method to scrub out metadata fields that may contain sensitive information.

[1]Department of Cardiothoracic Surgery, Stanford University, Palo Alto, CA, USA. [2]Department of Statistics, Columbia University, New York, NY, USA. [3]Department of Cardiovascular Medicine, Genetics, and Biomedical Data Science, Stanford University, Stanford, CA, USA. [4]Center for Artificial Intelligence in Medicine and Imaging, Stanford University, Stanford, CA, USA. [5]Department of Radiology and Biomedical Informatics, Stanford University, Stanford, CA, USA. ✉e-mail: willhies@stanford.edu

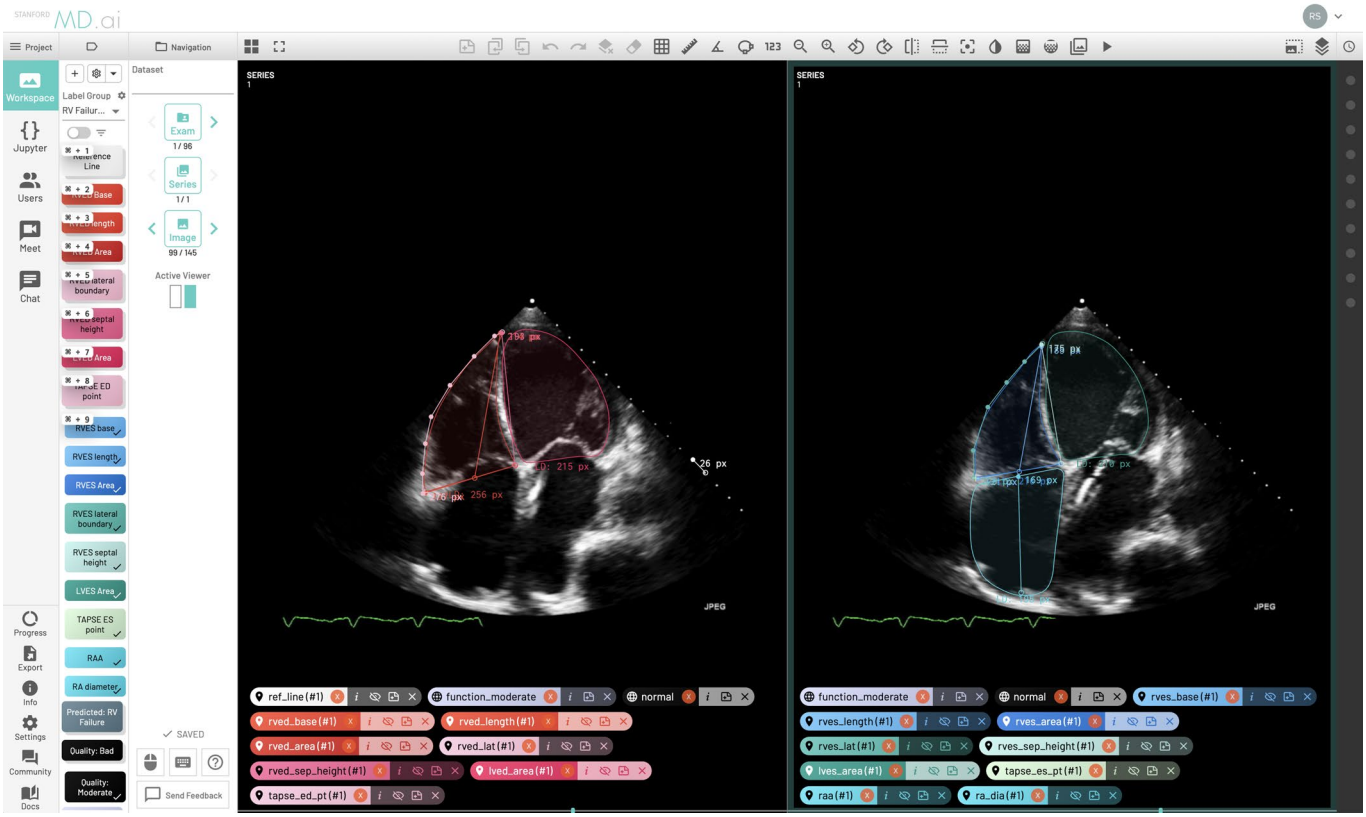

Fig. 1 | Cloud-based collaborative annotation workflows. Cloud-based tools such as MD.ai can be used to generate expert-annotated datasets and evaluate them against clinical experts via a secure connection. An implementation of MD.ai in which clinical experts make a variety of 2D measurements to quantify cardiac function is shown. Credit: MD.ai Inc, NY.

On the user-facing side, The MIRC Clinical Trials Processor anonymizer is a popular alternative, although it requires working with certain legacy software[18]. Well-documented Python packages (such as pydicom) may also be used to process DICOM files before use or transfer to collaborating institutions[19]. Imaging data can then be extracted and stored in a variety of machine-readable formats[20]. These datasets can quickly become large and unwieldy, and while a discussion on the specifics of data storage formats is beyond the scope of this Perspective, a key consideration for medical imaging AI is the preservation of image resolution.

An oft-cited drawback of automated de-identification methods or scripts is the potential for 'burned in' protected health information to remain on the imaging files. Despite the DICOM standard, manufacturer-specific differences make it difficult to generate simple rules via tools such as the MIRC Clinical Trials Processor to mask out regions where protected health information may be located. We suggest using a simple machine learning system for masking 'burned in' protected health information. In the case of echocardiograms, for example, there is a pre-defined scanning sector where the heart is visualized. Other potential options are machine learning-based optical character recognition tools to identify and mask out regions with printed text. The DICOM tags themselves can be useful for the extraction of both scan-level information and modality-specific tags. In the cases of echocardiography and cardiac magnetic resonance imaging (MRI), for example, important scan-level information such as acquisition frame rates and date, or MRI sequence (T1/T2), can readily be extracted from the DICOM metadata (Fig. 1).

For research endeavours that involve head-to-head benchmarking of AI systems against clinicians, or for curating large datasets with the help of clinical annotators, we recommend that a copy of the scans be stored in the DICOM format. This allows for deployment over scalable and easy-to-use cloud-based annotation tools. Several solutions exist for assigning scans for assessment by clinical experts. The requirements may range from simple scan-level labels to detailed domain-specific anatomical segmentation masks. At our institution, we deployed MD.ai (New York, New York)—a cloud-based annotation system that natively works with DICOM files stored on institutionally approved cloud storage providers (Google Cloud Storage or Amazon AWS). Alternatives offer similar functionality, such as ePadLite (Stanford, California), which is available free of cost[21]. An additional advantage of the cloud-based annotation approach is that the scans are kept at native resolution and quality. Real-time collaboration simulates 'team-based' clinical decision-making. Annotations and labels can easily be exported for downstream analyses. Most importantly, many of these tools are accessible remotely from any modern web browser and are extremely easy to use, drastically improving user experience and reducing the technical burden on clinical collaborators.

Finally, newer machine learning training paradigms such as federated learning may help circumvent many of the barriers associated with data sharing. Kaissis et al. reviewed the principles, security risks and implementation challenges of federated learning[22]. The key feature of this method is that local copies of algorithms are trained at each institution, and the only information that is shared is the features learned by the neural network during training. At predetermined intervals, the information learned (trained weights) from each institutional algorithm is then pooled together and redistributed—effectively learning from a large multicentre dataset without the need to transmit or share any of the medical imaging data[23,24]. This has been instrumental in rapidly training algorithms to detect features of COVID-19 from computed tomography scans of the chest[25]. Although there have been successful demonstrations

of federated learning in medical imaging, there remain substantial technical challenges in implementing these methods for routine clinical use[25]. Specifically in the context of high-dimensional imaging machine learning systems, the network latency introduced by the need to transmit and update trained weights from multiple participating centres becomes a fundamental rate-limiting step in training larger neural networks. Researchers must also ensure that the transmission of the trained weights is secure and encrypted between participating institutions, which further increases network latency[26]. Furthermore, curating datasets for quality and consistency while designing a study can be extremely challenging without access to the source data. Many conceptually similar federated learning frameworks still assume a degree of access to the source data[27].

Computational architectures

Neural network architectures used in modern clinical machine learning are largely derived from those optimized for large photo or video recognition tasks[28]. These architectures are remarkably robust even in the otherwise challenging task of fine-grained classification, where classes have subtle intraclass variance (breeds of dogs), rather than obviously different objects with high interclass variance (airplanes versus dogs). With adequate pre-training on large datasets (for example, ImageNet) these 'off the shelf' architectures outperform their tailor-made fine-grained classifier counterparts[29]. Many of these architectures are available for use in popular machine learning frameworks such as TensorFlow and Pytorch[30–34]. Most importantly, these frameworks often provide ImageNet pre-trained weights for a variety of different neural network architectures, allowing researchers to rapidly repurpose them for specialized medical imaging tasks[35].

Unfortunately, the vast majority of clinical imaging modalities are not simply static 'images'. An echocardiogram, for example, is a two-dimensional (2D) ultrasonographic video of the heart. These 'videos' can be taken from multiple different view planes, allowing for a more complete assessment of the heart. CT and MRI scans can be thought of as a stack of 2D images that must be analysed in sequence, or practitioners run the risk of missing valuable relationships between organs along one axis or another. These 'imaging' modalities are thus are more similar to videos, where unstacking them as images may lead to the loss of spatial or temporal context: processing a video by analysing each frame as a separate independent image, for example, leads to the loss of temporal information between each video frame[4,36,37]. In a variety of tasks utilizing echocardiography and CT and MRI scans, video-based neural network algorithms have shown considerable improvements over their 2D counterparts, yet integrating multiple different view planes brings an additional layer of dimensionality that is challenging to incorporate into current frameworks[2,4,38]. Unlike the extensive libraries of pre-trained image-based networks, support for video algorithms remains limited. Researchers interested in deploying newer architectures will probably need to perform pre-training steps on large publicly available video datasets (such as Kinetics and UCF101 (University of Central Florida 101 – Action Recognition Data Set)) themselves[39]. Furthermore, video networks can be orders of magnitude more computationally expensive to train. While pre-training using large natural scenery datasets is an accepted strategy in developing clinical imaging machine learning systems, performance gains are not guaranteed[40]. Reports of performance improvements are common with pre-training, especially when working with smaller datasets, but the benefits taper off with larger training datasets[2].

The lack of medical imaging-specific architectures was raised as a key challenge in the 2018 National Institutes of Health roadmap[13]. We extend this further by proposing that how we train these architectures has a large role to play in how well these systems will translate to the real world. We believe that the next generation of high-dimensional medical imaging AI will require training on richer, contextually more meaningful targets, rather than simple categorical labels. Most medical imaging AI systems today focus on diagnosing a handful of diseases from a normal background. The typical approach is to assign a numeric label (disease: 1; normal: 0) when training these algorithms. This is quite different from how clinical trainees learn to diagnose different diseases from imaging scans. In an effort to provide more 'medical knowledge' as opposed to simply pre-training on natural images or videos, Taleb et al.[37] proposed a series of novel self-supervised pre-training techniques using large unlabelled medical imaging datasets with the aim of assisting the development of 3D medical imaging–based AI systems. Neural networks learn to 'describe' the imaging scans provided as inputs by first performing a set of 'proxy tasks'[37]. For example, by tasking networks to 'reassemble' scrambled input scans as one would a jigsaw puzzle, they can be trained to 'understand' which anatomical structures line up with one another in various pathological and physiological states. Pairing data from imaging scans with their radiology reports is another interesting strategy that saw considerable success with chest X-ray–based AI systems[41]. In the spirit of providing more nuanced clinical context and embedding more 'knowledge' into neural networks, the text in the reports is processed via state-of-the-art natural language machine learning algorithms that subsequently train the vision network to better understand what makes various diseases appear 'different'. Most importantly, however, they show that using such approaches can reduce the amount of labelled data by up to two orders of magnitude for specific downstream classification tasks[41]. Unlabelled imaging studies—either alone or in combination with paired text reports—can therefore serve as the groundwork for effective pre-training. This would be followed by fine-tuning on a smaller sample of high-quality ground truth data towards a specific supervised learning task.

Although these steps help adapt existing neural network architectures for medical imaging, designing new architectures to specific tasks requires rare expertise. A model architecture is analogous to the brain, and the trained weights (the mathematical functions optimized during training) are analogous to the mind. Advances in evolutionary search algorithms make use of machine learning methods to discover new architectures tailored to a specific task, resulting in hyper-efficient and higher-performance architectures than those constructed by humans[42,43]. These offer a unique opportunity in the development of imaging-modality-specific architecture. Training deep learning algorithms rely on graphical processing units (GPUs) to perform the massively parallel matrix multiplication operations. The availability of cloud computing 'pay as you go' GPU resources and consumer grade GPUs with high memory capacities have all helped reduce the barrier to entry for researchers interested in developing machine learning systems for medical imaging. Despite these advances, training complex modern network architectures on large video datasets requires multiple GPUs running for weeks[33]. Clinical research groups should note that while training a single model might be feasible on a relatively inexpensive computer, finding the right combination of settings for the best performance almost always requires the use of specialized hardware and computing clusters to return results within a reasonable timeframe. Powerful abstraction layers (Pytorch Lightning, for example) also allow research groups to establish internal standards for structuring their code in a modular format. Adopting such modular approaches—where neural network architectures and datasets can be swapped out easily—helps to rapidly repurpose systems designed for clinical imaging modalities in the past to newer use cases. This approach also helps extend the capabilities of these systems by integrating subcomponents in novel ways.

Time-to-event analyses and uncertainty quantification

As medical AI systems shift from 'diagnostic' to more 'prognostic' applications, time-to-event predictions (rather than simple

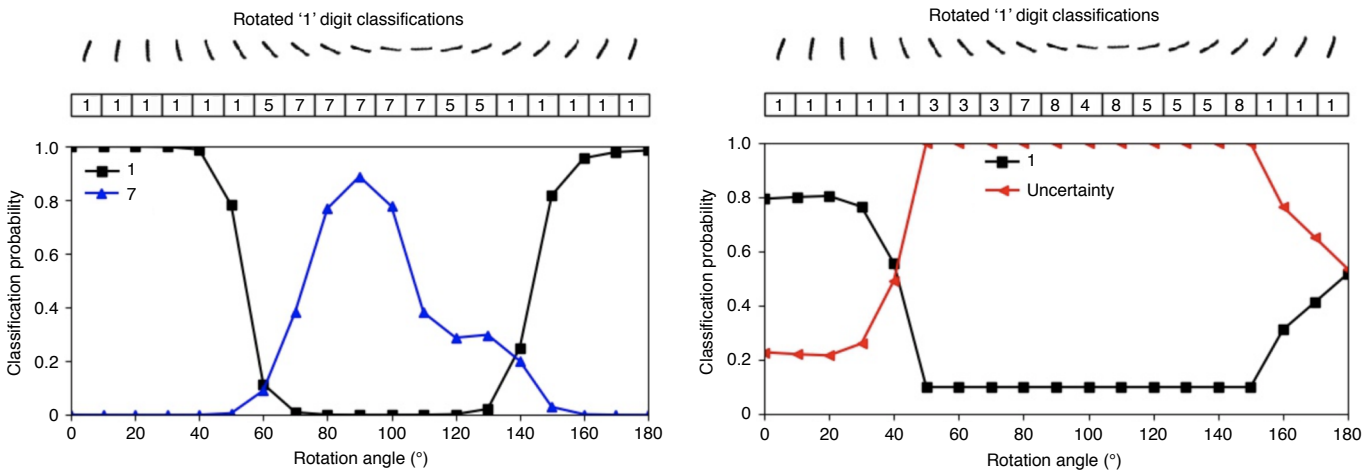

Fig. 2 | Quantifying uncertainty in machine learning outputs. Machine learning models trained with standard methods can be extremely confident even when incorrect, as described by Sensoy et al.[47]. Left: as a digit is rotated 180°, the system confidently assigns a label from '1' to '7'. Right: with methods that account for classification uncertainty, however, the system assigns an uncertainty score that can help alert clinicians to potentially erroneous predictions. Figure courtesy of the authors of ref. [47].

binary predictions) will find more relevance in the clinical setting. Time-to-event analyses are characterized by the ability to predict event probabilities as a function of time, whereas binary classifiers can provide predictions for only one predetermined duration. Unlike binary classifiers, time-to-event analyses account for censoring of data to allow for individuals who either were lost to follow-up or did not experience the event of interest within the observation timeframe. Survival analyses are commonplace in clinical research, and are central to the development of evidence-based practice guidelines. Extending traditional survival models with image- and video-based machine learning may provide powerful insight into the prognostic value of features within histological sections or medical imaging scans. For example, integrating extensions of Cox proportional-hazards loss functions into traditional neural network architectures made cancer outcome prediction from histopathology slides alone possible[44,45]. We do not advocate using such vision networks to dictate how care should be administered, but instead advocate their use as a method to flag cases where features of advanced malignancy were missed by clinicians. Incorporating time-to-event analyses will be increasingly relevant in clinical situations where indolent and early stages of disease have detectable features that that may progress rapidly after a certain amount of time. Retinal features diagnostic of macular degeneration, for example, often take years to manifest[8]. Patients with incipient features of disease may be labelled as 'normal', muddying the waters for neural networks attempting to make predictions about the future risk of developing complications of macular degeneration. Incorporating concepts of survival and censoring may help train systems to better separate normal individuals from those with mild, moderate and rapidly advancing disease. Similarly, training vision networks for time-to-event analyses may find use in screening for lung cancer, helping with risk stratification based on expected potential for aggressive spread. Critical for such translational efforts is the availability of robust and well-validated deep learning extensions of the Cox regression. Over the past several years, a number of deep learning implementations of the Cox model have been described. Kvamme et al. proposed a series of proportional and non-proportional extensions of the Cox model, with additional implementations of survival methods described in the past, such as DeepSurv and DeepHit[46] (Fig. 2).

Time-to-event predictions can, however, prove to be problematic from an actionable standpoint. In the hypothetical example of lung cancer screening, a suspicious nodule on a computed tomography scan of the chest might yield a prediction for median survival with and without appropriate therapeutic interventions. It might be interesting for the clinician to know how certain the machine learning system is about its prediction for an individual patient. Humans tend to err on the side of caution when unsure about a task. This is mirrored by machine learning systems where the output is a 'class probability' or 'likelihood of being correct' on a scale of 0 to 1. Most medical imaging machine learning systems described in literature today, however, lack the implicit ability to say 'I don't know' when provided input data that are out of distribution for the model. A classifier trained to predict pneumonia from computed tomography scans (for example) is by design coerced to provide an output (of either pneumonia or no pneumonia) even if the input image is that of a cat. In their paper on uncertainty quantification in deep learning, Sensoy et al. addressed these issues with a series of loss functions that assign an 'uncertainty score' as a way to avoid erroneous, but confident, predictions[47]. The benefits of uncertainty quantification arises later in the translational phase of a project, when AI systems are deployed in environments working alongside human users. Confidence measures were a key element of AlphaFold2, the protein-folding machine learning system that achieved unparalleled levels of accuracy in the 14th Critical Assessment of Protein Structure Prediction (CASP14) challenge, giving the DeepMind research team a way to gauge how much trust they should place in the predictions being generated[48,49]. Numerous implementations of uncertainty quantification methods are available under permissive licenses and are compatible with commonly used machine learning frameworks[50]. The incorporation of uncertainty quantification may help increase both the interpretability and the reliability of high-stakes medical imaging machine learning systems, and reduce the likelihood of automation bias—a phenomenon whereby clinicians may over-rely on automation[51].

Explainable AI and risk of harm

Aside from quantifying how certain machine learning systems are of their predictions, understanding how these machine learning systems arrive at their conclusions is of considerable interest to both the engineers building these systems and the clinicians using them. Saliency maps and class activation maps remain the de-facto standard for explaining how machine learning algorithms make their

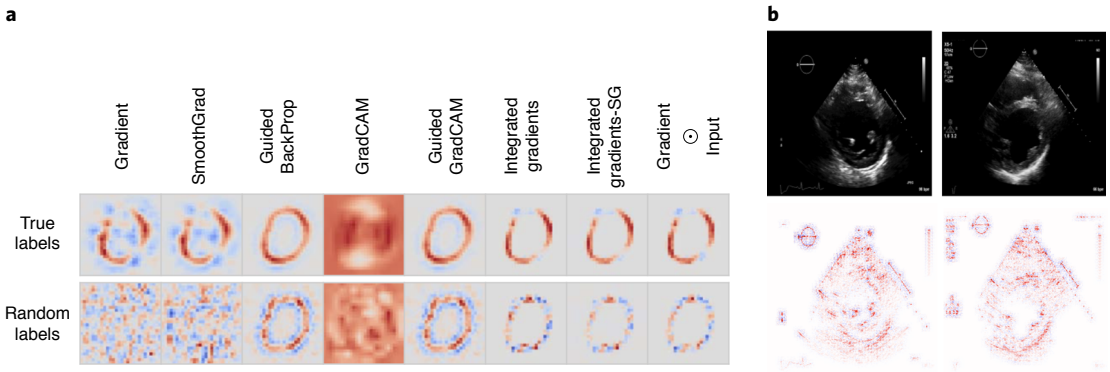

Fig. 3 | Misleading nature of post-hoc model explanations. a, Experiments conducted by Adebayo et al.[54] with models trained on true labels from the MNIST dataset (top) and models trained on random noise (bottom). Models trained on random noise still yield the circular shape of the digit zero when evaluated by the majority of visualization methods. These offer little in terms of true saliency maps, functioning more as class invariant edge detectors. **b**, Detection of echocardiographic view planes: both incorrect classifications (top left) and correct classifications (top right) yield similar saliency maps (bottom). Figure courtesy of the authors of ref. [54].

predictions[52,53]. Adebayo et al. recently showed that relying solely on the visual appearance of saliency maps can be misleading even if at first glance they appear contextually relevant. In a series of extensive tests, they found that instead of deriving true meaning from model weights, many popular methods for generating post-hoc saliency maps are in fact no different from 'edge detectors' (algorithms that simply map sharp transition areas between pixel intensities)[54]. Furthermore, even when these visualization methods work, little can be deciphered beyond 'where' the machine learning algorithms are looking, with numerous examples in which saliency maps look nearly identical for both correct and incorrect predictions[55]. These drawbacks are more pronounced when the difference between a 'diseased' state and a 'normal state' requires attention on the same region of an image or video[56] (Fig. 3).

Clinicians should note that heatmaps alone are insufficient methods for explaining how AI systems function, and care must be taken when attempting to identify failure modes using visualizations such as the ones shown above. A more granular approach may involve serial occlusion tests, where performance is assessed on images after intentional masking of regions that clinicians would otherwise use to make diagnoses or predictions[57]. The idea is quite intuitive: by running the algorithm on images with areas known to be important for diagnosing a certain condition masked off (for example, masking out the left ventricle when attempting to diagnose heart failure), a precipitous decline in performance should be seen. This helps to confirm that the AI system is attending to relevant areas. Specifically in the context of high-dimensional medical imaging studies, activation maps may offer unique insights into the relative importance of certain temporal phases of video-like imaging studies. Certain diseases may show pathognomic features when the heart is contracting, for example, whereas other conditions may require one to focus on when the heart is relaxing. Often such experiments may show that machine learning systems identify potentially informative features from regions of images that clinicians would not traditionally use[6]. In addition to gleaning information on how these machine learning systems generate their outputs, rigorous visualization experiments may offer a unique opportunity to learn biological insights from the machine learning systems being evaluated. On the other hand, deviations of activation from clinically known areas of importance may signal that networks are learning non-specific features, making them unlikely to generalize well to other datasets[58].

The features learned by an machine learning system can depend on architectural design choices. More importantly, machine learning systems will learn and perpetuate systemic inequities on the basis of the training data and targets provided to it[59,60]. As healthcare AI systems move towards future prediction of disease, greater care must be taken in accounting for the extensive disparities in access to healthcare and the outcomes across these groups. In a recent review, Chen et al. gave an in-depth overview of potential sources of bias from problem selection to the post-deployment phase[61]. Here we focus on potential solutions early in the development of machine learning systems. There have been demands for methods to explain otherwise 'black box' predictions of modern machine learning systems, while others have advocated restricting ourselves to more explainable models to begin with[55]. An intermediate approach involves training medical imaging neural networks using black box models in addition to incorporating inputs for structured data when training the overall AI system. This can be achieved by building 'fusion networks' in which tabular data are incorporated into image- or video-based neural networks, or other more advanced methods with the same fundamental goal (autoencoders that generate a low-dimensional representation of the combined data)[14,62,63]. Even without the incorporation of demographic inputs into high-dimensional vision networks, it is critical that research groups audit their models by comparing performance across genders, ethnicities, geographies and income groups. Machine learning systems may inadvertently learn to further perpetuate and discriminate against minorities and people of colour, and it is essential to understand this kind of bias early in the model development process[59,61]. Trust in machine learning systems is critical for wider adoption, as is exploring how and why specific features or variables lead to predictions via a combination of saliency maps and model agnostic approaches of estimating feature importance[64–66]. An alternative approach is constraining a machine learning algorithm within the training logic, ensuring that optimization steps occur to control for demographic variables of interest. This is analogous to a multivariable regression model wherein the effect of risk factors of interest can be studied independently of baseline demographic variables. From a technical standpoint, this would involve inserting an additional penalty loss in the training loop, keeping in mind the potential trade-offs with slightly lower model performance[67]. Fairlearn, for example, is popular toolkit for assessing fairness in traditional machine learning models, and constrained optimizations based on the Fairlearn algorithms (FairTorch) have been developed that are a promising exploratory foray into incorporating bias adjustments within the training process[68]. Numerous open-source toolkits exist to help researchers determine the relative importance different variables and input streams (image predictions, and variables such as

gender and race). These techniques may allow the development of more equitable machine learning systems, and even uncover hidden biases where none are anticipated[69].

Conclusion

Although computational architectures and access to high-quality data are key to building good models, developing translatable machine learning systems for high-dimensional imaging modalities requires proactive efforts to better represent the 'video-like' nature of the data, in addition to building in features that help address bias, uncertainty and explainability at the earliest stages of model development. The scepticism surrounding medical imaging and AI is healthy and, for the most part, warranted. We hope that meaningful steps towards improving the delivery of AI will be made possible by building in features that allow researchers to assess clinical performance, integration within hospital workflows, interactions with clinicians and the downstream risk of socio-demographic harm. We hope that researchers will find this Perspective useful, for both the overview of potential challenges that await them in terms of clinical deployment and the tacit guidance towards how some of these issues may be addressed.

Received: 23 March 2021; Accepted: 7 September 2021;
Published online: 16 November 2021

References

1. Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).
2. Rajpurkar, P. et al. AppendiXNet: deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Sci. Rep.* **10**, 3958 (2020).
3. Huang, S.-C. et al. PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. *npj Digit. Med.* **3**, 61 (2020).
4. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* https://doi.org/10.1038/s41586-020-2145-8 (2020).
5. Ghorbani, A. et al. Deep learning interpretation of echocardiograms. *npj Digit. Med.* **3**, 10 (2020).
6. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
7. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
8. Yim, J. et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat. Med.* **26**, 892–899 (2020).
9. Beede, E. et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–12 (ACM, 2020); https://doi.org/10.1145/3313831.3376718
10. Allen, B. et al. A road map for translational research on artificial intelligence in medical imaging: from the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop. *J. Am. Coll. Radiol.* **16**, 1179–1189 (2019).
11. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. Preprint at https://arxiv.org/abs/1912.01703 (2019).
12. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/abs/1603.04467v2 (2016).
13. Langlotz, C. P. et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology* **291**, 781–791 (2019).
14. Ulloa Cerna, A. E. et al. Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-020-00667-9 (2021).
15. Raghunath, S. et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat. Med.* **26**, 886–891 (2020).
16. Oren, O., Gersh, B. J. & Bhatt, D. L. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digit. Health* **2**, e486–e488 (2020).
17. Mildenberger, P., Eichelberg, M. & Martin, E. Introduction to the DICOM standard. *Eur. Radiol.* **12**, 920–927 (2002).
18. Mesterhazy, J., Olson, G. & Datta, S. High performance on-demand de-identification of a petabyte-scale medical imaging data lake. Preprint at https://arxiv.org/abs/2008.01827 (2020).
19. Mason, D. et al. pydicom/pydicom: pydicom 2.1.0. *Zenodo* https://doi.org/10.5281/ZENODO.4197955 (2020).
20. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
21. Rubin, D. L. et al. Automated tracking of quantitative assessments of tumor burden in clinical trials. *Transl. Oncol.* **7**, 23–35 (2014).
22. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
23. Chang, K. et al. Distributed deep learning networks among institutions for medical imaging. *J. Am. Med. Inform. Assoc.* **25**, 945–954 (2018).
24. Balachandar, N., Chang, K., Kalpathy-Cramer, J. & Rubin, D. L. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. *J. Am. Med. Inform. Assoc.* **27**, 700–708 (2020).
25. Xu, Y. et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. Preprint at *medRxiv* https://doi.org/10.1101/2020.05.10.20096073 (2020).
26. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**, 473–484 (2021).
27. Warnat-Herresthal, S. et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
28. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
29. Anwar, S., Barnes, N. & Petersson, L. A systematic evaluation: fine-grained CNN vs. traditional CNN classifiers. Preprint at https://arxiv.org/abs/2003.11154 (2020).
30. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. Preprint at https://arxiv.org/abs/1603.05027 (2016).
31. Hara, K., Kataoka, H. & Satoh, Y. Learning spatio-temporal features with 3D residual networks for action recognition. Preprint at https://arxiv.org/abs/1708.07632 (2017).
32. Tan, M. & Le, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. Preprint at https://arxiv.org/abs/1905.11946 (2019).
33. Carreira, J. & Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. Preprint at https://arxiv.org/abs/1705.07750 (2018).
34. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at http://arxiv.org/abs/1409.1556 (2014).
35. Marcel, S. & Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proc. International Conference on Multimedia - MM '10* 1485 (ACM, 2010); https://doi.org/10.1145/1873951.1874254
36. Zhang, J. et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation* **138**, 1623–1635 (2018).
37. Taleb, A. et al. 3D self-supervised methods for medical imaging. Preprint at https://arxiv.org/abs/2006.03829v3 (2020).
38. Shad, R. et al. Predicting post-operative right ventricular failure using video-based deep learning. *Nat. Commun.* **12**, 5192 (2021).
39. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C. & Zisserman, A. A short note about Kinetics-600. Preprint at https://arxiv.org/abs/1808.01340 (2018).
40. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: understanding transfer learning for medical imaging. Preprint at https://arxiv.org/abs/1902.07208 (2019).
41. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. Preprint at https://arxiv.org/abs/2010.00747 (2020).
42. Real, E., Aggarwal, A., Huang, Y. & Le, Q. V. Regularized evolution for image classifier architecture search. Preprint at https://arxiv.org/abs/1802.01548 (2019).
43. Piergiovanni, A., Angelova, A., Toshev, A. & Ryoo, M. Evolving space-time neural architectures for videos. In *2019 IEEE/CVF International Conf. Computer Vision (ICCV)* 1793–1802 (IEEE, 2019); https://doi.org/10.1109/ICCV.2019.00188
44. Yamashita, R., Long, J., Saleem, A., Rubin, D. L. & Shen, J. Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images. *Sci. Rep.* **11**, 2047 (2021).
45. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
46. Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-event prediction with neural networks and Cox regression. Preprint at https://arxiv.org/abs/1907.00825 (2019).
47. Sensoy, M., Kaplan, L. & Kandemir, M. Evidential deep learning to quantify classification uncertainty. Preprint at https://arxiv.org/abs/1806.01768 (2018).
48. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588**, 203–204 (2020).
49. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021).
50. Abdar, M. et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inform. Fusion* **76**, 243–297 (2021).

51. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **19**, 121–127 (2012).
52. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140 (2015).
53. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
54. Adebayo, J. et al. Sanity checks for saliency maps. Preprint at https://arxiv.org/abs/1810.03292 (2020).
55. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
56. Arun, N. et al. Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. Preprint at https://arxiv.org/abs/2008.02766 (2020).
57. Hughes, J. W. et al. Deep learning prediction of biomarkers from echocardiogram videos. Preprint at *medRxiv* https://doi.org/10.1101/2021.02.03.21251080 (2021).
58. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-021-00338-7 (2021).
59. Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S. & Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat. Med.* **27**, 136–140 (2021).
60. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
61. Chen, I. Y. et al. Ethical machine learning in health care. Preprint at https://arxiv.org/abs/2009.10576 (2020).
62. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3**, 136 (2020).
63. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
64. Esteva, A. et al. Deep learning-enabled medical computer vision. *npj Digit. Med.* **4**, 5 (2021).
65. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. Preprint at https://arxiv.org/abs/1704.02685 (2019).
66. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
67. Pfohl, S. R., Foryciarz, A. & Shah, N. H. An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.* **113**, 103621 (2021).
68. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. & Wallach, H. A Reductions approach to fair classification. Preprint at https://arxiv.org/abs/1803.02453 (2018).
69. Shapley, L. S. A value for n-person games. *Contrib. Theory Games* **2**, 307–317 (1953).

Acknowledgements

R.S. was supported in part by the American Heart Association Postdoctoral Fellowship Award (grant number 834986).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to William Hiesinger.

Peer review information *Nature Machine Intelligence* thanks Pearse Keane, Yipeng Hu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.