Chapter 6
Explainable AI and Its Applications in Healthcare

Arjun Sarkar

6.1 Introduction

Due to the lack of high-end graphics or tensor processing units, previously, deep neural networks could not be implemented as state-of-the-art Artificial Intelligence (AI) algorithms. Rather, linear models were preferred, and they were easy to understand and interpret. Things started changing with the advent of more advanced processing units, in the last decade, when the algorithms took on real-world problems. The models began getting bigger and better. While this highly improved the model performances, this also led to a problem of model interpretability. With access to large datasets, such as ImageNet (Krizhevsky et al. 2017), and these larger non-linear models with millions of parameters, AI soon started taking on human performance, on certain tasks. In 2015, a deep learning model called ResNet (He et al. 2016) surpassed human accuracy at the ImageNet challenge. Soon, AI was implemented in real-world tasks, and companies, industries, and research facilities started adopting AI into their workflows.

AI has now become a part of our day-to-day lives. AI algorithms not only help with day to day tasks such as finding outlines in today's phone cameras or recommending movies on Netflix, but also take on more challenging tasks such as beating human beings at strategy games (Lee et al. 2016; Garisto 2019) and surpassing human beings in complex visual recognition tasks (Bartolo et al. 2020; Thompson and Baker 2021). The rise of deep learning algorithms (Lecun et al. 2015) and computational power over the years has led to this extreme advancement in AI.

Healthcare costs are on the rise all around the globe. AI in medicine and healthcare can reduce this cost and, at the same time, improve healthcare (Higgins and Madai

A. Sarkar (✉)
Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, Jena, Germany
e-mail: arjun.sarkar786@gmail.com; arjun.sarkar@leibniz-hki.de

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_6

2020). Even though we read online how an AI beats doctors at predicting a particular disease every other day, the reality of acceptance of AI is still low in healthcare. While there are many reasons for this, one of the most prevalent ones is the explainability of the AI model. The 'black-box' nature of deep learning models is yet to be fully understood, and this causes a lack of trust and transparency. One error by an AI algorithm can be fatal for a patient in a hospital. Thus, the healthcare sector is cautious about implementing AI without completely understanding these algorithms. Most AI software implemented in hospitals today only helps diagnose and aid the doctor in making decisions. The accepted AI software goes through many regulations before being implemented in a hospital environment.

To build trust and reliability on these 'black-box' models, a new research field has emerged in recent years—eXplainable Artificial Intelligence (XAI). This field focuses on interpreting AI models and aims to provide an understandable way to explain AI predictions. At the rise of the deep learning era, most of the research was focused on improving model performance without caring much about explainability. But, that trend is now changing, with many researchers and companies looking to provide high AI accuracies and increased interpretability of the AI models.

A question may arise here: Why can an AI with high accuracy be trusted blindly? Initially, since the AI gives high accuracy, it may seem that the model can be implemented in a real-world situation. But many studies have shown that AI does not always learn the things that humans want it to learn (Lapuschkin et al. 2019). In the PASCAL VOC challenge (Everingham et al. 2010), it was often noticed that the AI was not precisely detecting the object of interest but making its classification based on context (Lapuschkin et al. 2016). For example, a classifier was often noticed to predict images of horses based on the watermark on the pictures and not on the actual horses. Similarly, the algorithm predicted an image as a train, not based on the train itself but the railway tracks (Lapuschkin et al. 2016). So, even though the model gave good accuracy, the correct predictions were often based on some artifacts. Usually, people can't comb through thousands of images on these big data challenges and figure out artifacts. So, these errors mostly go unnoticed. But these overfitting errors occur more often than expected. While this model trained on the PASCAL VOC dataset may perform overwhelmingly well on the test dataset, as the test data also belongs from the same distribution of images as the training data, the same model may fail miserably when tested on real-world data. This is just one of the many examples which fosters the need for the explainability of AI algorithms.

Sometimes, explainability is not about the end results but some intermediate learning. Deep learning algorithms have the power to find interesting patterns from images or text, which may be unknown to a human expert. When Deepmind's AlphaGo AI defeated the rank one human, Lee Sedol, at the game of Go, it played certain moves that other Go experts termed as 'not-human' (Thompson and Baker 2021). This just meant that a human being would not make that move, or that move was previously unknown to humans. Similarly, these algorithms can find patterns in medical images or correlate specific genes with certain diseases previously unknown to health experts. In the scientific and healthcare field, this can prove to be revolutionary. Often scientists and doctors focus more on certain patterns and features than

the final prediction, as those intermediate patterns can lead to new scientific discoveries. Due to the ability of deep learning algorithms to find patterns, these models have had massive success in the field of medicine (Ching et al. 2018; Piccialli et al. 2021), drug discovery (Chen et al. 2018; Gawehn et al. 2016), protein studies (Wang et al. 2017; Xu 2019), neuroscience (Marblestone et al. 2016; Richards et al. 2019), and radiology (Miotto et al. 2017; Kermany et al. 2018; Kuenzi et al. 2020).

The first part of the chapter looks at explainability from different aspects—the multidisciplinary nature of explainable AI in technological, legal, medical, and ethical aspects. Secondly, several explainability algorithms developed over the years which had significant impact on healthcare are explained in the next section. Finally, applications of these algorithms in real world medical tasks are showcased including the use of XAI in the recent COVID-19 pandemic.

6.2 The Multidisciplinary Nature of Explainable AI in Healthcare

The explainability of AI in the healthcare domain is not always a technological issue. It can often be due to combined medical, legal, or ethical issues (Amann et al. 2020).

6.2.1 Technological Outlook

The main issue of XAI is a technological problem: trying to explain an AI algorithm in a human-understandable form. The AI algorithm itself can achieve this explainability, or different models or methods can be used to describe a trained model (Rudin 2019). While the former can be achieved easily for linear models, the latter is necessary for the larger and more complex deep learning models.

Since the inception of XAI, various methods have been developed to try and explain these deep learning models. The explanation of linear models is always very accurate. But these models have severe performance issues compared to the more complex AI models (Esteva et al. 2019). So, there is a tradeoff between the complexity of a model and its explainability. Not only does model understanding help in understanding the final decision, but it also aids developers in tuning the parameters of the model and increasing performance. The problems of overfitting can be reduced or removed altogether. Researchers at Mount Sinai hospital trained a deep learning model to classify safe and high-risk patients based on X-ray images (Zech et al. 2018). The model produced high accuracies on the test set. But when the same model was tested on hospitals other than Mount Sinai, the model performance decreased. When XAI techniques were applied to the model, it was noticed that the model learned from the metadata of the X-ray machine at Mount Sinai hospital rather than on the actual X-ray images. The model was thus able to distinguish the

pictures easily from that particular X-ray machine but failed on images of other X-ray machines, as the metadata no longer matched. XAI techniques help identify and correct these problems before the model is deployed in a real-world scenario. This makes the model more robust, reduces integration costs, and saves time.

While certain XAI techniques help developers improve model parameters, other techniques help healthcare professionals without in-depth knowledge in programming understand predictions. Pointing out the position of infection in medical images has immense benefits for doctors and helps provide a second opinion when they are in doubt. AI also can find rare diseases that are not often known to even seasoned experts (Schaefer et al. 2020). During the recent COVID-19 pandemic, while many algorithms were developed to classify whether patients were infected or not, very few were deployed on the field due to a lack of proper explainability (Fuhrman et al. 2022).

6.2.2 Legal Outlook

The explainability of AI in healthcare is a legal need in nearly every country. In different sectors, the legal requirement for XAI is dissimilar. XAI is not a must in logistics, and a few errors are admissible. But in public administration or banking, XAI can play a vital role. A person whose loan has been rejected due to an AI model has the right to know the reason behind the rejection. In no other sector is XAI as mandatory as in the healthcare field (Schönberger 2019). This does not come as a surprise, as in healthcare, even one error has the potential of harming human life.

AI in healthcare is used for many applications, such as disease classification and diagnosis (Qiu et al. 2020), anomaly detection, patient positioning, image segmentation (Aslam et al. 2015), image super resolution (Chaudhari et al. 2018), and image registration (Ma et al. 2017; Wu et al. 2013). AI is meant to improve clinical applications and aid doctors, improve the standard of medical development and save patient lives. But to train a robust model, often sensitive patient data is required. These privacy issues must meet all legal requirements, from image acquisition to final prediction. Similarly, in recent years, anti-discrimination and explainability of Ai models have gained momentum (Deeks 2019).

Hospitals don't use the AI algorithm as a computer program, but the algorithm is wrapped in the form of a software with a user-friendly graphical user interface (GUI). It is a requirement by most regulatory bodies in the USA or the European Union to provide a level of transparency of the AI's output (Smith et al. 2020). Though these regulations are rather vague now, with no solid rules for explainability, these rules will supposedly get stricter as more emphasis is made on XAI.

One more budding question is about the awareness of these AI predictions and the disclosure to patients. That is, how much of the decision would be made by the AI and how much by the doctor, and finally, how the final prediction would be disclosed to the patient (Cohen 2020). One fear is that the legal system is not fast enough to keep with the rising pace of AI development. In healthcare, AI-based decisions need

strict laws such that they do not hamper innovation but also protect patients' rights and privacy. When these laws are clearly defined and AI researchers can overcome the problems with XAI, AI will be fast adopted in all healthcare sectors.

6.2.3 Medical Outlook

The medical outlook aims to bring semblance between the need for laboratory-based testing or replacing it entirely with AI-based algorithms. Laboratory testing and medical imaging are the methods that have been used since time immemorial for the proper diagnosis of a disease. These are methods that are understandable by medical experts. In laboratory testing and imaging, doctors access the results and the images and find meaningful patterns that point to certain complications or diseases. On the other hand, when trained on the images and the corresponding infection types, a deep learning model may predict the condition correctly but does not indicate the patterns it uses to come to that prediction. XAI can be helpful here in showing these patterns and can be much faster than even the most trained experts.

Even though an AI algorithm can provide good overall accuracy and low error rates (Weng et al. 2017), the algorithm cannot be perfect because of data inconsistency due to noise and imaging errors. The trained experts need to look at the false positives and negatives, and they cannot always be heavily reliant on AI. AI bias is another such complication that cannot be removed entirely. For example, suppose the training data is sampled from a large population of people from Europe in skin cancer prediction. In that case, the same algorithm will fail if deployed in Africa due to the differences in skin tone and color (Wen et al. 2022).

XAI can be crucial in deciding the amount of disagreement between a medical expert and an AI. The results of XAI in the medical field are often visual representations or textual explanations. These explanations can be beneficial to the medical experts in making a final decision on the diagnosis. Without XAI, the clinician has to choose blindly whether to trust the AI or not, but with XAI, the person can understand why the AI makes that particular decision. If an algorithm keeps performing poorly, the results can be reported to the developers, and the developers can understand the reason for the poor performance using explainability. In case the algorithm works well, the clinicians can trust it better when they understand the reason for its good performance (Cutillo et al. 2020).

6.2.4 Ethical Outlook

As more and more healthcare institutes adopt AI into their framework, certain ethical aspects need to be examined. One of these issues is protecting the autonomy of the patient and informing the patient about the use of the deep learning models for their diagnosis. If a patient is not informed whether a doctor or an AI algorithm predicted

a particular disease, this can hamper the patient's trust towards the doctor. A more critical situation would be if the prediction is a false positive or false negative, and the patient is mistreated. The patient can challenge the institution, and the healthcare facility will not be able to provide a concrete reason for such a prediction. This is one more reason for introducing a proper XAI before using the AI algorithm blindly. A solution to such a scenario can be first asking the patient for their permission to use AI for the diagnosis and later explaining the results of the AI to them.

One more ethical conundrum can arise with the rise of AI in healthcare. If AI systems start taking over more healthcare positions in the future, it can limit clinicians' decision-making rather than enhance them. This situation should be avoided, as the best outcome for a patient is to be not entirely dependent on an AI, but a decision of AI carefully analyzed by a doctor.

6.2.5 Patient Outlook

The patient outlook is a perspective that focuses on patients and considers them as an active part of the healthcare decision process (Baker 2001). This refers to the treatment process in which each patient is provided with a treatment best suited for that individual. The idea is to provide patient-centered medicine. But deep learning models predicting risk may not be able to provide such treatment. As doctors do not understand the inner workings of the models as well, neither can they inform the patient about the reason for the predicted risk. XAI can prove to be helpful and continue maintaining the patient-centered approach.

Similarly, wearable devices and smartwatches can now predict certain risk factors in patients (Bhattacharya and Lane 2016; Mauldin et al. 2018). Previously these devices provided similar treatment and health plans to all users. But recently, most companies have been trying to give each user a different health plan based on their activity, heart rate, and sleeping patterns (Coutts et al. 2020; Nweke et al. 2018). Risk assessment explained in the form of text or visual data builds trust and increases transparency and continues building towards a patient-centered innovation in healthcare.

6.3 Different XAI Techniques Used in Healthcare

There are various explainability methods for AI in medicine, and there are multiple ways to classify these methods. Some such taxonomies are explained below.

6.3.1 Methods to Explain Deep Learning Models

Since deep learning models are all black-box models and cannot be easily explained, most modern research is focused on trying to explain these models. Saliency maps is one such technique very commonly used to interpret convolutional neural networks (Itti et al. 1998). These saliency maps are pixels of the image that the convolutional neural network considers essential to the final prediction. Saliency is represented on the image as a visual heatmap or topography.

There are multiple gradient-based techniques used for the explainability of deep learning models (Simonyan et al. 2014). The gradient-based approach shows how much a change in the input would affect the output. Saliency maps are also based on this gradient technique. The Krizhevsky network (Krizhevsky et al. 2017) beat the previous methods and was considered one of the best gradient-based explainability methods. Another algorithm that improves gradient explainability is the DeepLIFT algorithm (Shrikumar et al. 2017). The DeepLIFT algorithm enhances the previous methods by multiplying the input signal to the gradient. The model's superiority was evident when tested on genomic data and natural images. The algorithm assigns a weighted score to the activation of all neurons in the model and shows some crucial connections or features that the previous models failed to identify. DeconvNets or Deconvolution (Zeiler and Fergus 2014) is another method to understand convolutional neural networks (CNN). Unlike regular convolutional layers that extract features from image pixels, deconvolution does the opposite—mapping features to pixel values. Deconvolution is generally used to understand what a convolutional neural network learns in every convolutional layer.

Class Activation Maps (CAM) (Zhou et al. 2016) and its more advanced counterparts Grad-CAM (Selvaraju et al. 2020) and Grad-CAM++ (Chattopadhay et al. 2018) are some of the most famous interpretability methods used to explain the results of convolutional neural networks. CAM helps to identify important locations of the image trained by a model to predict the class of the image. Activations from the final layer of the convolutional model are concatenated to create a feature vector. The weighted sum of this vector is fed into a SoftMax layer to calculate the final result. The result is displayed as a heatmap. While CAM gave good results, it could not be applied to any convolutional model. To overcome this problem, Grad-CAM was developed. Grad-CAM (Selvaraju et al. 2020) can generate the localization maps for any convolutional neural network, regardless of its shape or structure. Grad-CAM++ (Chattopadhay et al. 2018) further improves Grad-CAM by better visualization of the output maps and better object localization for multi-label classification.

The RISE algorithm (Petsiuk et al. 2019) slightly differs from the CAM algorithms. This algorithm considers each pixel of an image to generate a saliency map. Random masks are multiplied elementwise with the images and fed into the network. The model generates a probability score and a saliency map of the input image, which is obtained by combining the masks.

While many algorithms are trying to explain the results of the convolutional neural network models, some studies suggest that none of these techniques are correct in

interpreting the networks (Kindermans et al. 2018). The authors tested some of these explainability methods on simple linear models, but the methods could not correctly interpret the linear models. Hence, the authors argue that if these techniques cannot even explain simple linear models, is their explanation of large complex non-linear models, correct? They further proposed two additional models, PatternNet and PatternAttribution, which work well on linear models and more complex models.

In Natural Language Processing (NLP), a different method is used for explainability (Lei et al. 2016). Small pieces of the input text are added to the model as input, and the model aims to generate the entire text from these small text fragments. Finally, the generated text provides some context and justification for the generated text in terms of the input text.

LIME (Ribeiro et al. 2016) or Local interpretable model-agnostic explanations is an XAI method that can interpret any black-box model. It is also one of the most famous and commonly used interpretability methods for tabular data, text, and images. LIME can interpret individual predictions of a model. It tweaks the feature values of a single data sample and creates an impact of the tweak on the output. Even though LIME can be a simple and powerful interpretable model, it has certain drawbacks. Some studies have shown that choosing poor parameters can cause the model to give different results and miss many essential features completely. This can be a severe problem when the model is deployed in the field. The DLIME algorithm was proposed to overcome this problem. Random sampling used in LIME is replaced in DLIME by choosing clusters of similar data and selecting the most relevant cluster by running k-nearest neighbors (KNN). The authors of DLIME also proved the superiority of DLIME over LIME by testing it on three separate medical datasets.

Shapely Additive explanations (SHAP) (Lundberg and Lee 2017) is another often used interpretability technique. SHAP is a model inspired by game theory. It computes the importance of each feature for all predictions. The SHAP values are a combination of three important properties, namely, accuracy, missingness, and consistency. The authors demonstrate how SHAP is more intuitive and more human interpretable than other XAI methods. Various other model agnostic models such as Anchors (Ribeiro et al. 2018), DeepSHAP (Chen et al. 2021), Protodash (Kim et al. 2016), Permutation Importance (PIMP) (Altmann et al. 2010), and Contrastive Explanation Methods (CEM) (Dhurandhar et al. 2018) are often used for explainability as well.

In deep learning, attention is a trendy concept (Vaswani et al. 2017). The concept of attention was inspired by how humans pay attention to different parts of images or other data sources to analyze them. The MDNet network was created (Xia et al. 2020) to directly map medical imaging and corresponding diagnostic reports. It contained an image model as well as a language model. Attention mechanisms were used to visualize the detection process. This attention mechanism allowed the language model to discover the predominant and distinguishing features used to map the images and diagnostic reports. This was the first study to use the attention mechanism to gain insight from the medical image data. SAUNet, an interpretable U-Net version (Ronneberger et al. 2015), was created (Sun et al. 2020). It also added a secondary

shape stream to capture important shapes-based information in addition to the regular texture features. An attention module was used in the U-Net decoder. SmoothGrad (Hooker et al. 2019) was used to create spatially and shape attention mappings to visualize the high activation area of the images.

These are some commonly used XAI methods for deep learning models in healthcare. All these models have some significance, but there is no one idea to explain all kinds of text and image data or on all sorts of models. SHAP and its advancements are comprehensive and understandable XAI methods of all the methods. Grad-CAM is commonly used for convolutional model interpretation, even for industrial AI software deployed in hospitals.

6.3.2 Explainability by Using White-Box Models

White-box models are transparent models and are easily understandable or interpretable. This category contains mainly linear models, decision trees, and some complex models that are easy to interpret. Some of these complex models used in medical imaging and healthcare are listed here.

Microsoft came up with an interpretable model for predicting pneumonia risk, which also had great accuracy (Caruana et al. 2015). The authors discussed that while interpretable linear models and decision trees could not give good results, neural nets gave much better results, but at the cost of explainability. High-performance generalized additive models with pairwise interactions (GA2Ms) were proposed and tested on two real medical data case studies. The authors also mentioned that the model could be scaled to work on thousands of patient data without losing accuracy and still being highly interpretable.

Another technique that utilizes Boolean rules to create predictive models was proposed—Boolean Rule Column Generation (Dash et al. 2018). This technique uses easy-to-understand Boolean rules with some clauses and conditions. Humans easily understand these clauses and conditions. GLM or Generalized Linear Rule Models (Wei et al. 2019) use an ensemble of rule-based features. GLMs are easy to interpret and complex simultaneously, as the rules can capture non-linear dependencies. TED or Teaching Explanations for Decisions (Hind et al. 2019) is a framework that tries to produce explanations like a human expert rather than explaining the inner workings of an AI model.

Not much research has been done in the complex white-box model development domain. No white-box model can produce the same high accuracy as deep learning models. The white-box models are also very domain-specific, unlike various computer vision and natural language processing neural networks used on various real-world tasks.

6.3.3 *Explainability Methods to Increase Fairness in Machine Learning Models*

AI models are not just theoretical analysis techniques anymore, but with every passing day, more and more models are adopted in real-world applications. Any discrimination or inequality in these models can potentially impact human lives. The fairness of these AI models is another part of XAI that tackles ethical and social aspects. Usually, bias in the models is checked by implementing the model in a different setting, such as a different demographic, and evaluated. Many techniques developed in recent years focus on tackling the bias and discrimination in these models.

The method of disparate impact testing (Feldman et al. 2015) is a model-evaluation tool that can assess the fairness and accuracy of a model but does not provide any details or insight into the causes of bias. It uses simple experiments to highlight differences between model predictions and errors for different demographic groups. It can also detect biases in terms of ethnicity, gender, marital status, or demographics. Another data preprocessing technique was suggested to remove bias from machine-learning models (Calmon et al. 2017). The authors developed a convex optimization to learn a data representation that meets the three stated goals: controlling discrimination, limiting distortion in individual instances, and preserving utility. Adversarial debiasing (Zhang et al. 2018) is an approach to tackling biases regarding demographic segments in machine-learning systems. It involves selecting a feature about the element of interest and then simultaneously training both the primary and adversarial models. The main model is trained to predict the label. The adversarial model, based on the primary model's prediction for each instance, attempts to predict the segment. The goal is to maximize the main machine learning system's accuracy in correctly predicting the label while minimizing the adversarial ability. Adversarial biasing can be used for both classification and regression tasks.

Many methods to make classifiers aware of discriminatory biases need data modifications or algorithm tweaks (Kamiran et al. 2012). They are also not flexible regarding multiple sensitive features handling and control over performance versus discrimination tradeoff. Two new methods, Reject Option-based Classification and Discrimination-Aware Ensemble were developed to solve these problems.

Counterfactual fairness (Kusner et al. 2017) captures the intuition that a decision that affects an individual is fair if it affects the same person in both the real and counterfactual worlds. The individual would then be part of a different demographic. It was also argued that causality in fairness must be addressed. Consequently, a framework was developed to model fairness using tools of causal inference. The authors state that any measure of causality in fairness should not be based on counterfactuals. It is also essential to ensure that counterfactual causal guarantees can be used. Based on the concept of counterfactual fairness, the proposed framework allows users to create models that can take sensitive attributes that could reflect social biases towards people and compensate accordingly. A recent study (Kearns et al. 2018) found that most machine-learning fairness notions and definitions only focus on predefined social segments. It was also pointed out that while such simple

constraints can force classifiers at the segment level to attain fairness, they could lead to discrimination against sub-segments that contain specific combinations of sensitive feature values. The authors suggested that fairness be defined across an infinite or exponential number of sub-segments. These were then determined using the space of sensitive feature values. An algorithm was developed to produce the fairest distribution of sub-segments over classifiers.

One study (Elisa Celis et al. 2019) pointed out that while recent research has attempted to attain fairness regarding a particular metric, specific metrics have been overlooked. Furthermore, some proposed algorithms lack solid theoretical support. The authors developed a meta-classifier that could handle multiple fairness constraints concerning multiple non-disjoint sensitive elements. Another work pointed out that many existing notions about fairness regarding treatment and impact are too restrictive and strict. This can lead to poor model performance. The authors suggested notions of fairness that were based on the collective preferences of different demographic groups to address this issue. Their concept of fairness, more specifically, tries to define which outcome or treatment the various demographic groups would prefer if given a choice.

Fairness is still a new area of machine learning interpretability. However, the incredible progress made over the past few years has been remarkable. Many methods can ensure fair resource allocation and protect the most vulnerable demographics. Several techniques can be used to manipulate data before training models, make algorithmic changes during training, and adjust post-hoc. However, these methods tend to focus too heavily on group fairness. They often overlook individual-level factors at both the local and global levels, leading to the mistreatment of individuals. A small portion of scientific literature deals with fairness in images or text. This gap is still a significant one that needs to be explored in the future.

6.3.4 Explainability Methods to Analyze Sensitivity of a Model

Interpretability methods are used to evaluate and challenge machine learning models to ensure they are reliable and trustworthy. These methods use some form of sensitivity analysis. Models are evaluated for their stability and their ability to predict the impact of subtle but intentional changes in inputs. Sensitivity analysis may interpret changes in output across a range of examples or just one.

The sensitivity index is a traditional method of sensitivity analysis that represents each input variable using a numerical value. First-order indices measure the contribution of one input variable to the output variation. Second, third, and higher-order indexes measure the interaction contribution between two, three, or more input variables to that output variance. The total-effect indices combine the contributions of higher-order and first-order interactions with the output variance.

Sobol (2001) proposed an output variance sensitivity analysis based on ANOVA decomposition. He suggested using Monte-Carlo methods to approximate sensitivity indices higher and first order. Fourier Amplitude Sensitivity Test (FAST) (Cukier et al. 1973) is a method to improve the approximation of Sobol's indexes. The Fourier transformation converts a multi-dimensional integral to a one-dimensional integrated. These algorithms were further enhanced to an RBD-FAST (random balance designs-FAST) algorithm (Plischke 2010), which improved computational efficiency. Morris's method (Morris 1991) of global sensitivity analysis, also known as the one-step-at-a-time (OAT) method, is another option. Although the Morris method is complete, it can be very computationally expensive. Fractional factorial designs (Saltelli et al. 2008) needed to be developed and used in practice to perform sensitivity analysis more efficiently.

6.4 Application of XAI in Healthcare

There are two main types of explanations for deep neural networks in medical images: those that use standard attribution-based approaches and those that use novel, often domain-specific or architecture-specific methods. Many attribution methods can be used to assign an attribution value, contribution, or relevance to each network input feature. An attribution method determines the importance of an input element to the target neuron, which is often the output neuron for a classification problem. Heatmaps show the arrangement of all input features according to the shape of the input samples. Non-attribution is a methodology that is validated on an issue rather than using separate analyses using pre-existing methods. These included concept vectors, attention maps, return of a similar image, and text justifications.

Some applications of XAI in healthcare are explained in this section. While each healthcare domain has hundreds of studies where XAI has been used, only a few examples from each domain are listed.

6.4.1 Medical Diagnostics

One study (Kavya et al. 2021) proposed a computer-aided framework for allergy diagnosis. They evaluated several ML algorithms and then chose the most effective one using k-fold cross-validation. They developed a rule-based approach to the XAI method by creating a random forest. If-Then rules and explanations representing each path within a tree are extracted using medical data. The computer-aided framework was also deployed on the mobile app by the authors, which can be used to assist junior clinicians in verifying the diagnostic predictions. Another study (Dindorf et al. 2021) suggests an explanation-independent classifier for spinal positions. SVM and radiofrequency were used as ML classifiers. Then, they applied LIME to predict the classification. The authors of another study (El-Sappagh et al. 2021) suggested an

RF model to diagnose and detect Alzheimer's progression. The authors also used SHAP to identify the essential features of the classifier. Next, they used a fuzzy rule-based method. SHAP could provide a local explanation for specific patient diagnosis/progression prediction explanations about feature impacts. The fuzzy rule-based system could also generate natural language forms that can aid patients and doctors in understanding the AI model. One paper suggested an XAI framework to assist doctors in diagnosing hepatitis patients (Peng et al. 2021). The authors compared intrinsic XAI methods such as logistic regression, decision trees, and kNN to the more complex models SVM, XGBoost, and RF. The authors also used the post-hoc methods SHAP and LIME and partial dependence plots (PDP). For chronic wound classification, a CNN model was proposed (Sarp et al. 2021). For XAI, the authors used LIME, which aided clinicians in better diagnosis.

6.4.2 Medical Imaging

Due to their simplicity, attribution-based methods were used in most medical imaging literature. Researchers can efficiently train a neural network architecture that is suitable without making it difficult to explain. They also have access to an attribution model. A pre-existing deep learning model or a custom model can obtain the best results on a given task. The existing model implementation is more straightforward and can leverage transfer learning techniques. In comparison, custom models can concentrate on specific data and avoid overfitting with fewer parameters. Both are useful for medical imaging datasets.

Analyzing the post-model data using attributions can show if the model is learning the right features or if it's learning the wrong features. This allows researchers to adjust the hyperparameters and architecture of the model to get better results with test data and potentially in a real-world setting.

6.4.2.1 Brain Imaging

Different methods were analyzed in a study to compare their robustness in CNN's Alzheimer's classification using brain MRI. The methods that were compared were LRP (Bach et al. 2015) and Guided backpropagation (GBP). The L2 norm was calculated between the average attribution maps for multiple runs to check the repeatability of heatmaps of identically trained models. Because occlusion covers more area, it was an order of magnitude lower than the baseline occlusion. LRP performed better than all other methods, indicating a fully attribution-based method. LRP also had the highest similarity in the sum, density, and gain (sum/density), for the top 10 regions across all attributions. Another study (Pereira et al. 2018) used GradCAM and GBP to examine the clinical coherence between the features learned from a CNN for automatic grading brain tumors using MRI. Both methods activated the tumor and surrounding ventricles, which could indicate malignancy. They were both correctly

graded in cases. This focus on non-tumor areas and spurious patterns in GBP maps can lead to errors that indicate unreliability.

6.4.2.2 Breast Imaging

SmoothGrad and IG were used to visualize features in a CNN for classifying estrogen receptor status using breast MRI (Papanastasopoulos et al. 2020). The model learned relevant features from both dynamic and spatial domains, with each contribution. Visualizations showed that the model had learned some irrelevant features from pre-processing artifacts. These observations led us to make changes in our pre-processing and training methods. A previous study to classify breast mass from mammograms (Hassan et al. 2020) used two different CNNs, AlexNet (Szegedy et al. 2015) or GoogleNet (Krizhevsky et al. 2017)—and used saliency maps for visualizing image features. Both CNNs were able to detect the contours of the mass, which is the essential clinical criteria. They also showed sensitivity to context. In another study (Amoroso et al. 2021), the authors also presented an XAI framework to help breast cancer patients. The framework was used to identify a patient's most important clinical feature.

6.4.2.3 Skin Imaging

GradCAM and KernelSHAP were used to compare the features of a set of 30 CNN models trained for melanoma detection (Young et al. 2019). GradCAM and Kernel SHAP were used to compare the features of a suite of 30 CNN models trained for melanoma detection. It was found that even high-accuracy models would sometimes focus on features that were not relevant to the diagnosis. The attribution maps of both methods showed differences in the models' explanations. This demonstrated that different neural network architectures learn various features. A further study (Molle et al. 2018) showed how CNN features were used to classify skin lesions. By scaling the feature maps of activations to the input size, the features for the two last layers were visualized. The layers looked for indicators such as lesion borders, color irregularities, and risk factors such as lighter skin or pink textures. However, some spurious features such as hair and artifacts had no significance.

6.4.2.4 X-ray Imaging

Some studies have used attribution-based diagnostic methods in addition to the more popular imaging modalities. These include both image inputs and non-image inputs. One study used CNNs to perform uncertainty and interpretability analyses on colorectal polyps (Wickstrøm et al. 2020). This is a precursor to rectal cancers. CNN used GBP to create heatmaps. They were found to use the shape and edge information to make predictions. The uncertainty analysis also revealed higher levels of

uncertainty in samples that were misclassified. The authors (Lundberg et al. 2018) presented a model that uses SHAP attributions for hypoxemia. This study was done to analyze preoperative and in-surgery factors. The resulting attributions were consistent with known factors such as BMI, physical state (ASA), tidal volumes, inspired oxygen, and others.

Attribution-based methods were the first method of visualizing neural networks. They have evolved from simple gradient-based class activation maps to more advanced techniques such as Deep SHAP. These visualizations show that models are learning relevant features in most cases. Any spurious features were flagged and corrected by the readers. The identification of relevant features can be improved by smaller models and custom variants to the attribution methods.

6.4.2.5 CT Imaging

DeepDreams inspired attribution method (Mordvintsev et al. 2015) was presented in (Couteaux et al. 2019) to explain the segmentation and classification of tumors from liver CT images. Based on the DeepDreams concepts, this innovative method can be applied to black-box neural networks. The algorithm performed a sensitivity assessment of the features and maximized the activation of target neurons by performing gradient ascent. Comparing networks trained on synthetic and real tumors showed that the former was more sensitive than the latter to clinically relevant features. At the same time, the latter was also more focused on other features. The network was sensitive to both intensity and sphericity with domain knowledge.

6.4.2.6 Retina Imaging

As a diabetic retinopathy (DR) tool, grading by ophthalmologists, a system that produced IG heatmaps and model predictions was investigated (Sayres et al. 2019). The assistance provided by the system was shown to improve the accuracy of the grading over that of an expert without any help or the model predictions. Although the initial grading process was slower, users soon found that it improved their grading experience. This is especially true when heatmaps and predictions are used. Patients without DR saw a decrease in accuracy when model assistance was used. Expressive gradients (EG) were proposed as an extension to IG for weakly supervised segmentation (Yang et al. 2019). Compact CNNs performed better than larger ones, and EG highlighted regions of interest more effectively than traditional IG or GBP methods. EG extends IG by enriching input-level attribution maps with high-level attribution maps. A comparative analysis of various explainability models, including DeepLIFT, DeepSHAP, IG, etc., was performed for a model for detection of choroidal neovascularization (CNV), and diabetic macular edema (DME), and drusens from optical coherence tomography (OCT) scans (Singh et al. 2020).

6.4.3 Surgery

In one study (Kletz et al. 2019), the authors presented a CNN-based medical app to learn the representations of instruments in laparoscopy. They validated their model using different datasets. To help explain how the model classified an instrument, they also provided activation maps from different CNN layers. XAI-CBIR (Chittajallu et al. 2019) was proposed to explain surgical training. XAI-CBIR provides an example post hoc explanation of XAI methods. It extracts representative examples to offer explanations. It uses a self-supervised deep learning model to extract semantic descriptions from MIS video frames. It also used a saliency map to explain visually why it believes the image retrieved is similar to the query. Minimally invasive surgery (MIS) videos can be retrieved using the XAI CBIR system.

6.4.4 Detection of COVID-19

Understanding the COVID-19 data associated with COVID-19 is necessary to fully understand the clinical applications of explainable AI for COVID-19 assessment (Fuhrman et al. 2022). While reverse transcription-polymerase chain reaction (RT-PCR) tests are the most common tool for COVID-19 detection, radiography, and CT scans can supplement RT-PCR testing to improve detection accuracy and throughput. For COVID-19 diagnosis, only X-ray or CT finding may not be sufficient. Therefore, differential diagnosis is difficult because of the subtle differences in COVID-19 and non-COVID-19 pneumonia (Cleverley et al. 2020). For improved differential diagnosis and detection accuracy, explainable AI can be helpful (Dong et al. 2021; Salehi et al. 2020). A standard language for describing COVID-19 can also be used. These datasets are publicly available and can meet data requirements. This includes large-scale projects such as the NIH-funded Medical Imaging and Data Resource Center.

This case study is mainly focused on radiography and chest CT. However, other modalities such as PET/CT, lung ultrasound, and MRI may also play a part in COVID-19 patient care. The development of AI systems to assess COVID-19 is similar to other disease evaluations in many ways. The most common use of explainable AI in COVID-19 assessment is to ensure the model accurately focuses on regions of concern in the input image that indicate disease presence. This is usually done through heatmap visualization. Some studies have had mixed success (Mei et al. 2020; Xiong et al. 2020; Wehbe et al. 2021). One study (Wehbe et al. 2021) shows heatmaps for both negative and positive COVID-19 cases. They note that the negative examples have a low influence on the lungs. Another study (Xiong et al. 2020) shows heatmaps that accurately highlight COVID-19 in lung segmentation and identify regions without significant COVID-19 content to aid the classification decision. This finding is limited in understanding the model's performance and should be investigated further before clinical implementation. COVID-19 can be easily confused

with other diseases such as viral pneumonia. One study differentiates between these, and the results of this study are precise (Jin et al. 2020). The authors divided CT scans into four types of pneumonia and COVID-19. They also identified phenotypic mistakes that were common for humans and AI readers. Grad-CAM and Guided GradCAM were used to visualize the most critical image regions. The authors also provided segmentation for diseased areas. Like previous works, Grad-CAM indicates that the model identifies high-value regions within and outside of the lungs. However, Guided GradCAM does not capture all of the diseased lung tissue. They also use t-SNE to visualize feature embeddings from the various disease classes and identify image features that may be problematic in the classification decision. Another study (Zhang et al. 2020) presents another unique use of explainable AI in the COVID-19 assessment. In this case, the authors use clinical metadata and quantitative lesion features to create classifiers that can predict patient prognosis. They use Shapley numbers to assess how each feature impacts the risk classifier. This includes whether it increases or decreases a prediction output. They also evaluate the effectiveness of different drug administrations and the patient's response to treatment. This type of analysis is beneficial for understanding images indicative of high risk. It is helpful when combined with clinical metadata.

A method called GSInquire was used in a recent study to detect COVID-19 using chest X-ray images (Wang et al. 2020). It produced heatmaps that were used to verify the features of the COVID-net model. GSInquire was created to be an attribution method that performed better than other methods such as SHAP or Expected gradients. It uses the new metrics impact score and coverage. The impact score was the percentage of features that strongly impacted the model decision or confidence. Impact coverage was determined in relation to the inclusion of factors that could be adversely affected. While these studies use python programming to create the deep learning models, the Cognex VisionPro Deep Learning Software classified Covid-19 X-ray images using their deep learning-based graphic user interface (GUI) (Sarkar et al. 2021). The software has built-in Grad-CAM for interpretability, highlighting the regions of interest. A trained medical expert can then look at the Grad-CAM and judge the efficacy of the software.

6.5 Conclusion

Despite its rapid growth, explainable AI is still not a mature field. It often suffers from a lack of formality and poorly defined definitions. Although many machine learning interpretability methods and studies have been developed in academia and other institutions, they do not often form an integral part of machine-learning workflows or pipelines.

This chapter examines the role of explicable AI in clinical decision-support systems from technological, legal, and ethical perspectives.

There are many applications of XAI within the healthcare industry. The concept of explainability has many implications for all stakeholders. Developers, doctors,

and legislators face challenges regarding medical AI. Combining multiple modalities, such as medical images and patient records, is possible to make decisions and attribute model decisions to each one. This could simulate a physician's workflow where images and patient parameters are used to make a diagnosis. This can improve accuracy and provide more detailed explanations. Despite making impressive strides in explaining the diagnosis, there are still many steps to meet regulators and end-users needs.

There is still much to be done in machine learning interpretability methods. Many studies have been conducted over the years and demonstrated many opportunities for improvement. They also highlighted the potential benefits and enhancements these methods bring to existing machine-learning workflows. However, they also revealed their weaknesses and performance limitations. However, we believe that explainable AI still has many unexplored areas and great potential to be explored in the future.

References

Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26** (2010). https://doi.org/10.1093/bioinformatics/btq134

Amann, J., Blasimme, A., Vayena, E., et al.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. **2** (2020). https://doi.org/10.1186/s12911-020-01332-6

Amoroso, N., Pomarico, D., Fanizzi, A., et al.: A roadmap towards breast cancer therapies supported by explainable artificial intelligence. Appl. Sci. (Switzerland) **11** (2021). https://doi.org/10.3390/app11114881

Aslam, A., Khan, E., Beg, M.M.S.: Improved edge detection algorithm for brain tumor segmentation. Procedia Comput. Sci. (2015)

Bach, S., Binder, A., Montavon, G., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE (2015). https://doi.org/10.1371/journal.pone.0130140

Baker, A.: Book: crossing the quality chasm: a new health system for the 21st century. BMJ **323** (2001). https://doi.org/10.1136/bmj.323.7322.1192

Bartolo, M., Roberts, A., Welbl, J., et al.: Beat the AI: investigating adversarial human annotation for reading comprehension. Trans. Assoc. Comput. Linguist. **8** (2020). https://doi.org/10.1162/tacl_a_00338

Bhattacharya, S., Lane, N.D.: From smart to deep: Robust activity recognition on smartwatches using deep learning. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2016 (2016)

Calmon, F.P., Wei, D., Vinzamuri, B., et al.: Optimized pre-processing for discrimination prevention. In: Advances in Neural Information Processing Systems (2017)

Caruana, R., Lou, Y., Gehrke, J., et al.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015)

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings—2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018 (2018)

Chaudhari, A.S., Fang, Z., Kogan, F., et al.: Super-resolution musculoskeletal MRI using deep learning. Magn. Reson. Med. (2018). https://doi.org/10.1002/mrm.27178

Chen, H., Engkvist, O., Wang, Y., et al.: The rise of deep learning in drug discovery. Drug Discov. Today **23** (2018)
Chen, H., Lundberg, S., Lee, S.I.: Explaining models by propagating shapley values of local components. In: Studies in Computational Intelligence (2021)
Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., et al.: Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface **15** (2018). https://doi.org/10.1098/rsif.2017.0387
Chittajallu, D.R., Dong, B., Tunison, P., et al.: XAI-CBIR: explainable AI system for content based retrieval of video frames from minimally invasive surgery videos. In: Proceedings—International Symposium on Biomedical Imaging (2019)
Cleverley, J., Piper, J., Jones, M.M.: The role of chest radiography in confirming covid-19 pneumonia. BMJ **370** (2020)
Cohen, I.G.: Informed consent and medical artificial intelligence: what to tell the patient? SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3529576
Couteaux, V., Nempont, O., Pizaine, G., Bloch, I.: Towards interpretability of segmentation networks by analyzing deepDreams. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2019)
Coutts, L.V., Plans, D., Brown, A.W., Collomosse, J.: Deep learning with wearable based heart rate variability for prediction of mental and general health. J. Biomed. Inform. **112** (2020). https://doi.org/10.1016/j.jbi.2020.103610
Cukier, R.I., Fortuin, C.M., Shuler, K.E., et al.: Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. J. Chem. Phys. **59** (1973). https://doi.org/10.1063/1.1680571
Cutillo, C.M., Sharma, K.R., Foschini, L., et al.: Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit. Med. **3** (2020)
Dash, S., Günlük, O., Wei, D.: Boolean decision rules via column generation. In: Advances in Neural Information Processing Systems (2018)
Deeks, A.: The judicial demand for explainable artificial intelligence. C. Law Rev. **119** (2019)
Dhurandhar, A., Chen, P.Y., Luss, R., et al.: Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Advances in Neural Information Processing Systems (2018)
Dindorf, C., Konradi, J., Wolf, C., et al.: Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (Xai). Sensors **21** (2021). https://doi.org/10.3390/s21186323
Dong, D., Tang, Z., Wang, S., et al.: The role of imaging in the detection and management of COVID-19: a review. IEEE Rev. Biomed. Eng. **14** (2021). https://doi.org/10.1109/RBME.2020.2990959
Elisa Celis, L., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: a meta-algorithm with provable guarantees. In: FAT* 2019—Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (2019)
El-Sappagh, S., Alonso, J.M., Islam, S.M.R., et al.: A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci. Rep. **11** (2021). https://doi.org/10.1038/s41598-021-82098-3
Esteva, A., Robicquet, A., Ramsundar, B., et al.: A guide to deep learning in healthcare. Nat. Med. **25** (2019)
Everingham et al. 2010Everingham, M., van Gool, L., Williams, C.K.I., et al.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88** (2010). https://doi.org/10.1007/s11263-009-0275-4
Feldman, M., Friedler, S.A., Moeller, J., et al.: Certifying and removing disparate impact. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015)
Fuhrman, J.D., Gorre, N., Hu, Q., et al.: A review of explainable and interpretable AI with applications in COVID-19 imaging. Med. Phys. **49** (2022)

Garisto, D.: Google AI beats top human players at strategy game StarCraft II. Nature (2019). https://doi.org/10.1038/d41586-019-03298-6
Gawehn, E., Hiss, J.A., Schneider, G.: Deep learning in drug discovery. Mol. Inform. **35** (2016)
Hassan, S.A., Sayed, M.S., Abdalla, M.I., Rashwan, M.A.: Breast cancer masses classification using deep convolutional neural networks and transfer learning. Multimed. Tools Appl. **79** (2020). https://doi.org/10.1007/s11042-020-09518-w
He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
Higgins, D., Madai, V.I.: From bit to bedside: a practical framework for artificial intelligence product development in healthcare. Adv. Intell. Syst. **2** (2020). https://doi.org/10.1002/aisy.202000052
Hind, M., Wei, D., Campbell, M., et al.: TED: teaching AI to explain its decisions. In: AIES 2019—Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019)
Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: Advances in Neural Information Processing Systems (2019)
Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998). https://doi.org/10.1109/34.730558
Jin, C., Chen, W., Cao, Y., et al.: Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat. Commun. **11** (2020). https://doi.org/10.1038/s41467-020-18685-1
Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: Proceedings—IEEE International Conference on Data Mining, ICDM (2012)
Kavya, R., Christopher, J., Panda, S., Lazarus, Y.B.: Machine learning and XAI approaches for allergy diagnosis. Biomed. Signal Process. Control **69** (2021). https://doi.org/10.1016/j.bspc.2021.102681
Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: 35th International Conference on Machine Learning, ICML 2018 (2018)
Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172** (2018). https://doi.org/10.1016/j.cell.2018.02.010
Kim, B., Khanna, R., Koyejo, O.: Examples are not enough, learn to criticize! Criticism for interpretability. In: Advances in Neural Information Processing Systems (2016)
Kindermans, P.J., Schütt, K.T., Alber, M., et al.: Learning how to explain neural networks: PatternNet and PatternAttribution. In: 6th International Conference on Learning Representations, ICLR 2018—Conference Track Proceedings (2018)
Kletz, S., Schoeffmann, K., Husslein, H.: Learning the representation of instrument images in laparoscopy videos. Healthc. Technol. Lett. (2019)
Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM (2017). https://doi.org/10.1145/3065386
Kuenzi, B.M., Park, J., Fong, S.H., et al.: Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell **38** (2020). https://doi.org/10.1016/j.ccell.2020.09.014
Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems (2017)
Lapuschkin, S., Binder, A., Montavon, G., et al.: Analyzing classifiers: fisher vectors and deep neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
Lapuschkin, S., Wäldchen, S., Binder, A., et al.: Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. **10** (2019). https://doi.org/10.1038/s41467-019-08987-4
Lecun, Y., Bengio, Y., Hinton. G.: Deep learning. Nature (2015)
Lee, C.S., Wang, M.H., Yen, S.J., et al.: Human versus computer go: review and prospect [Discussion Forum]. IEEE Comput. Intell. Mag. **11** (2016). https://doi.org/10.1109/MCI.2016.2572559
Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. In: EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Proceedings (2016)
Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (2017)

Lundberg, S.M., Nair, B., Vavilala, M.S., et al.: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. **2** (2018). https://doi.org/10.1038/s41551-018-0304-0

Ma, K., Wang, J., Singh, V., et al.: Multimodal image registration with deep context reinforcement learning. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2017)

Marblestone, A.H., Wayne, G., Kording, K.P.: Toward an integration of deep learning and neuroscience. Front. Comput. Neurosci. **10** (2016). https://doi.org/10.3389/fncom.2016.00094

Mauldin, T.R., Canby, M.E., Metsis, V., et al.: Smartfall: a smartwatch-based fall detection system using deep learning. Sensors (Switzerland) **18** (2018). https://doi.org/10.3390/s18103363

Mei, X., Lee, H.C., Diao, K.Y., et al.: Artificial intelligence—enabled rapid diagnosis of patients with COVID-19. Nat. Med. **26** (2020). https://doi.org/10.1038/s41591-020-0931-3

Miotto, R., Wang, F., Wang, S., et al.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. **19** (2017). https://doi.org/10.1093/bib/bbx044

Mordvintsev, A., Tyka, M., Olah, C.: Inceptionism: going deeper into neural networks, google research blog. In: Google Research Blog (2015)

Nweke, H.F., The, Y.W., Al-garadi, M.A., Alo, U.R.: Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Syst. Appl. **105** (2018)

Papanastasopoulos, Z., Samala, R.K., Chan, H.-P., et al.: Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI (2020)

Peng, J., Zou, K., Zhou, M., et al.: An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. J. Med. Syst. **45** (2021). https://doi.org/10.1007/s10916-021-01736-5

Pereira, S., Meier, R., Alves, V., et al.: Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018)

Petsiuk, V., Das, A., Saenko, K.: RisE: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference 2018, BMVC 2018 (2019)

Piccialli, F., di Somma, V., Giampaolo, F., et al.: A survey on deep learning in medicine: why, how and when? Inf. Fusion **66** (2021). https://doi.org/10.1016/j.inffus.2020.09.006

Plischke, E.: An effective algorithm for computing global sensitivity indices (EASI). Reliab. Eng. Syst. Saf. **95** (2010). https://doi.org/10.1016/j.ress.2009.11.005

Qiu, S., Joshi, P.S., Miller, M.I., et al.: Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain **143** (2020). https://doi.org/10.1093/brain/awaa137

Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)

Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 (2018)

Richards, B.A., Lillicrap, T.P., Beaudoin, P., et al.: A deep learning framework for neuroscience. Nat. Neurosci. **22** (2019)

Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2015)

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1** (2019)

Salehi, S., Abedi, A., Balakrishnan, S., Gholamrezanezhad, A.: Coronavirus disease 2019 (COVID-19) imaging reporting and data system (COVID-RADS) and common lexicon: a proposal based on the imaging data of 37 studies. Eur. Radiol. **30** (2020). https://doi.org/10.1007/s00330-020-06863-0

Saltelli, A., Ratto, M., Andres, T., et al.: Global sensitivity analysis: the primer (2008)
Sarkar, A., Vandenhirtz, J., Nagy, J., et al.: Identification of images of COVID-19 from chest X-rays using deep learning: comparing COGNEX VisionPro deep learning 1.0^{TM} software with open source convolutional neural networks. SN Comput. Sci. **2** (2021). https://doi.org/10.1007/s42979-021-00496-w
Sarp, S., Kuzlu, M., Wilson, E., et al.: The enlightening role of explainable artificial intelligence in chronic wound classification. Electronics (Switzerland) **10** (2021). https://doi.org/10.3390/electronics10121406
Sayres, R., Taly, A., Rahimy, E., et al.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology **126** (2019). https://doi.org/10.1016/j.ophtha.2018.11.016
Schaefer, J., Lehne, M., Schepers, J., et al.: The use of machine learning in rare diseases: a scoping review. Orphanet J. Rare Dis. **15** (2020)
Schönberger, D.: Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. Int. J. Law Inf. Technol. **27** (2019). https://doi.org/10.1093/ijlit/eaz004
Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128** (2020). https://doi.org/10.1007/s11263-019-01228-7
Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: 34th International Conference on Machine Learning, ICML 2017 (2017)
Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014—Workshop Track Proceedings (2014)
Singh, A., Mohammed, A.R., Zelek, J., Lakshminarayanan, V.: Interpretation of deep learning using attributions: application to ophthalmic diagnosis (2020)
Smith, J.A., Abhari, R.E., Hussain, Z., et al.: Industry ties and evidence in public comments on the FDA framework for modifications to artificial intelligence/machine learning-based medical devices: a cross sectional study. BMJ Open **10** (2020). https://doi.org/10.1136/bmjopen-2020-039969
Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55** (2001). https://doi.org/10.1016/S0378-4754(00)00270-6
Sun, J., Darbehani, F., Zaidi, M., Wang, B.: SAUNet: shape attentive U-net for interpretable medical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2020)
Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015)
Thompson, B., Baker, N.: Google AI beats humans at designing computer chips. Nature (2021). https://doi.org/10.1038/d41586-021-01558-y
van Molle, P., de Strooper, M., Verbelen, T., et al.: Visualizing convolutional neural networks to improve decision support for skin lesion classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018)
Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
Wang, L., Lin, Z.Q., Wong, A.: COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci. Rep. (2020). https://doi.org/10.1038/s41598-020-76550-z
Wang, S., Li, Z., Yu, Y., Xu, J.: Folding membrane proteins by deep transfer learning. Cell Syst. **5** (2017). https://doi.org/10.1016/j.cels.2017.09.001
Wehbe, R.M., Sheng, J., Dutta, S., et al.: DeepCOVID-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. clinical data set. Radiology **299** (2021). https://doi.org/10.1148/RADIOL.2020203511

Wei, D., Dash, S., Gao, T., Günlük, O.: Generalized linear rule models. In: 36th International Conference on Machine Learning, ICML 2019 (2019)
Wen, D., Khan, S.M., Xu, A.J., et al.: Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit. Health **4** (2022)
Weng, S.F., Reps, J., Kai, J., et al.: Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE **12** (2017). https://doi.org/10.1371/journal.pone.0174944
Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. **60** (2020). https://doi.org/10.1016/j.media.2019.101619
Wu, G., Kim, M., Wang, Q., et al.: Unsupervised deep feature learning for deformable registration of MR brain images. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2013)
Xia, H., Sun, W., Song, S., Mou, X.: Md-net: multi-scale dilated convolution network for CT images segmentation. Neural Process. Lett. **51** (2020). https://doi.org/10.1007/s11063-020-10230-x
Xiong, Z., Wang, R., Bai, H.X., et al.: Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology **296** (2020). https://doi.org/10.1148/radiol.2020201491
Xu, J.: Distance-based protein folding powered by deep learning. Proc. Natl. Acad. Sci. U. S. A. **116** (2019). https://doi.org/10.1073/pnas.1821309116
Young, K., Booth, G., Simpson, B., et al.: Deep neural network or dermatologist? In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2019)
Zafar, M.B., Valera, I., Rodriguez, M.G., et al.: From parity to preference-based notions of fairness in classification. In: Advances in Neural Information Processing Systems (2017)
Zech, J.R., Badgeley, M.A., Liu, M., et al.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. **15** (2018). https://doi.org/10.1371/journal.pmed.1002683
Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2014)
Zhang, K., Liu, X., Shen, J., et al.: Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell (2020). https://doi.org/10.1016/j.cell.2020.04.045
Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating Unwanted Biases with Adversarial Learning. In: AIES 2018—Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)
Zhou, B., Khosla, A., Lapedriza, A., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)