Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

EXPLORING THE INTERPRETABILITY OF DEEP LEARNING MODELS IN MEDICAL DIAGNOSIS

Muhammad Zarar, [2]Talha Nadeem, [3]Huzaifah Bin Khawar, [4] Hood Khizer

[1] AI Researcher, Research and Development Department, Trouve Labs, AHOY DMCC and The Superior University Lahore, Pakistan muhammadzarar.ieee@gmail.com

[2] Systems Architect AI/ML, Research and Development Department, Trouve Labs, AHOY DMCC and The Superior University Lahore, Pakistan talhanadeem.ieee@gmail.com

[3] AI Researcher, Research and Development Department, Trouve Labs, AHOY DMCC and The Superior University Lahore, Pakistan huzaifahkhawar.ieee@gmail.com

[4] CEO Trouve Labs, CTO AHOY DMCC, Research and Development Department, Trouve Labs, AHOY DMCC, Dubai hoodkhizer@gmail.com

DOI:

Keywords
Deep learning, interpretability, medical diagnosis, explainable AI, Grad-CAM, LIME, SHAP, healthcare, medical imaging, clinical decision-making.

Article History
Received: 10 Jan, 2026
Accepted: 18 Jan, 2026
Published: 31 Jan, 2026

Copyright @Author
Corresponding Author: * Talha Nadeem*

Abstract

The developments of deep learning technologies have led to a significant enhancement in the accuracy and speed of medical diagnosis systems in medical imaging and disease forecasting, among other applications. The issue with deep learning models, however, is that they are, by design, "black box" systems, which can pose a problem in a medical setting where understanding is essential for clinical decision making. This research delves into some of the methods proposed to improve the interpretability of deep learning models for MC diagnosis. We review some of the most popular methods: feature attribution methods such as Grad-CAM, and methods such as LIME and SHAP, highlighting their strengths and weaknesses in terms of their capacity to provide clinically meaningful explanations from complex model outputs. Proposed mathematical/evaluation framework is discussed, along with illustrative clinical applications. This review shows the need to oversee model performance while preserving model transparency in terms of prediction accuracy and the ethical and practical considerations of implementing the explainable AI system into clinical workflows. Analysis for the future suggests enhanced integration of interpretability in real-time medical diagnostics and regulatory issues for trust and reliability with AI applications to healthcare..

I. INTRODUCTION

In the medical diagnostics field, deep learning (DL) has proven an invaluable asset and has delivered remarkable accuracy across a spectrum of applications, from image analysis to disease prediction [1]. In radiology, pathology, genomics and beyond, the ability of DL to process huge amounts of data and the identification of patterns

that can compliment human assessment has helped to progress the field. While these achievements have been accomplished, lack of transparency remains one of the significant challenges of deep learning models in Healthcare. These kinds of models are frequently referred to as 'black-box' systems, due to the fact that the inner workings of the decision-making process that occurs are not easily interpretable [3]. An opacity is a significant hurdle, especially when using a model in medical settings where understanding how the model reaches a diagnosis is vital to engender patients' confidence and safety of the model's use with them [4].

Interpretability is crucial in medicine since medical professionals have to make decisions which directly affect health outcomes [5]. Predictive accuracy is adequate in many other industries but in the medical field, explaining and trusting the models used is essential [6]. In the context of cancer prediction, for example, the specific features or data points utilized for the diagnosis are identified by the deep learning model, not merely for the sake of it, but to explain why the model has made its prediction [7]. A crucial aspect of this interpretability is that it enables physicians to trust the AI and accept any interpretation generated by the AI as an accurate diagnosis, based on evidence-based reasoning and on their own judgment of medical knowledge [8]. Even if the model is correct and accurate, it can be an unreliable or unsafe model in a clinical setting without these explanations [9].

As black-box systems become more prevalent, a greater emphasis is being placed on the need for greater interpretability.The interpretability of black-box systems is becoming increasingly important and various tools have been developed that yield more insight into AI models. To generate visual explanations, text that support their decision, feature attribution techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) show those portions of the image that have the greatest impact on the model's classification [10]. These methods can prove valuable in medical imaging, where it matters to understand the parts of an image that played a role in the diagnosis, in order to better judge the reasoning in the model. In parallel, other model independent methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) attempt to explain complex models in a simple way that is easier to understand and analyze [4, 11]. The tools transform into explanations that are meaningful to non-experts so that the results can be interpreted by clinicians in the context of their medical knowledge [12].

In real-world applications, the opacity of deep learning models could hinder their use in healthcare due to the difficulty in interpreting their outcomes. AI-based tools are often not trusted by clinicians because they can't explain the logic the model used to make a decision [6]. In medicine, as an example, a model such as a convolutional neural network (CNN) may be better at spotting abnormalities in medical pictures than the human radiologists responsible for reading them, but without the capacity to explain to a clinician which portions of the image were used to make the diagnosis, there may be skepticism about the benefit of adopting the model [9]. This is particularly the case in critical medical situations where the impact of a misdiagnosis can be serious. In addition to fostering trust in AI systems, interpretability can also enable seamless integration and collaboration between AI tools and human expertise,

empowering healthcare professionals to trustfully use the AI's predictions as a second opinion while leveraging their own knowledge to actually make a decision [5].

Accountability and transparency are key to medical decisions made with the help of AI. Representative data, independent training and test sets, clinically relevant testing, the human-AI team, and clear information for users are key principles outlined by the FDA, Health Canada and MHRA in their 2021 principles for good machine learning practice [13]. These rules are not to be interpreted as being a legal requirement for any specific explanation method. An explanation alone is no guarantee for the safety and absence of bias, or regulatory compliance, and does not equate to interpretability.Validation and governance complement interpretability.

Interpretability also plays a key role in patient care, which has increasingly been a component of contemporary healthcare. Patients today are more engaged in their health care decisions and want to know 'how' and 'why' of diagnosis and treatment. [6] To successfully embed AI systems within this paradigm, they must also be able to explain what they did, in such a way that can be understood by the patient. This especially needs to be the case when a diagnosis is not anticipated or complicated, and the patient needs reassurance that the decision has been made in an un-emotional yet rational way [8]. Making AI systems interpretable may create greater trust in the system, leading to better collaboration and satisfaction among patients and clinicians [4].

A. Background of Deep Learning in Healthcare

Deep learning (DL) has introduced advanced analytical capabilities to handle large and complex datasets in the medical field and has thus transformed the healthcare continuum in several aspects, including research on diagnostic accuracy, predictive analytics, and clinical decision support [14]. Machine learning (ML) is a subset of the broader fields of AI and DL, with the aim of learning representations from data by means of multilayer artificial neural networks (ANNs) [1]. The healthcare sector, historically a data heavy but interpretation light industry, has seen remarkable benefits from the processing capabilities of DL to patterns, clinical data including medical images and electronic health records (EHRs), genomics and even the large amounts of data coming in from wearables [15].

Traditionally, medical diagnostics have depended on the skill level and years of experience of clinicians and specialists making diagnoses. [16] But in the era of DL, one can also augment human knowledge with systems that have the ability to process medical data on previously unimaginable scales and at speeds. In the medical field, for instance, DL approaches have been compared with conventional ones for the classification of medical images, including skin cancer from dermoscopy images or lung nodules from computed tomography images [14, 17]. They can be trained on massive amounts of data to recognize minor patterns that might not be easily discernible by a human eye and may improve diagnostic accuracy.

Medical imaging is one area where DL has seen significant success in healthcare. In particular, the images from radiology, pathology, and ophthalmology have been suitable for specific classification using convolutional neural networks (CNNs), which are a particular type of deep learning model [2]. CNNs have proven to have the utmost capabilities in the diagnosis of disease like diabetes retinopathy, pneumonia and breast cancer in many cases even leading the game to the

expert physicians [18] and [19]. This is due to CNNs' hierarchical structure that enables it to automatically capture features at various layers of abstraction from images, which is suitable for understanding complex medical images [1].

DL has come a long way on the field of predictive analytics. Next, the performance of recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which work particularly well with sequences of data, has been assessed for EHRs that pose prediction challenges such as hospital readmission, mortality, and disease progression [20]. These models can be used for temporal data analysis and detect trends and pattern over time that can suggest deterioration or recovery of a patients' status [21]. These have huge potential in preventive care where earlier interventions have the potential to really make a difference to patient outcomes [15].

With all these progresses, there is still a lack of there are still challenges in applying DL to healthcare. Many DL models are "black boxes" that are difficult to interpret. The key problem of these models is the lack of interpretability and trustworthiness on how decisions are made in the clinical settings due to their black-box nature [3]. In a context of high stakes and the life of the patient in danger, clinicians are hesitant in adapting a system that does not offer them a clear justification for their forecasts [5]. Further, the principles of regulation in AI-driven Healthcare are still in flux, and biases, accountability, and transparency challenges must be resolved before safe and ethical use of DL systems is achieved [16]. Although practitioners are experts with large volumes, there are challenges associated with data sharing, data privacy and the silo effect of data spread across various systems in many health facilities [22]. Access to well-structured and well annotated datasets is a key requirement to train strong DL models but finding such datasets in medical applications can be a challenging task because of the nature of medical data and the diversity of the formats and standards [14].

Even with these challenges, the future of DL in healthcare is looking bright. Attention mechanism for interpretability, explainable AI (XAI) techniques, DL and more interpretable methods such as decision trees and logistic regression are also being explored to improve the interpretability of DL models [3, 4]. Such advances are contributing to the convergence of the DL's modelling accuracy and transparency requirements in healthcare decision-making [6].

The future of healthcare with DL holds immense promise, promising a more accurate diagnosis, accurate forecasting, and more personalized treatment. But it is crucial that the field moves beyond current hurdles of interpretability, access to data, and ethics to better realize the potential of these technologies in the future, potentially aiding patient care [20].

B. Importance of Interpretability in Medical Diagnosis

The interpretability of medical diagnoses, especially in the context of deep learning (DL) models, is essential. In a therapeutic context where life and death situations can arise, the trust people put in tools is essential and therefore the complete grasp of what, why and how a model predicts is fundamental [7]. Deep learning models, specifically neural networks, are known as "black boxes" and their internal decisions, such as how to process input variables, are opaque, as compared to traditional statistical models or simpler machine learning (ML) model which has better understanding for how such models create decisions based on the input variables. This

opacity is a major problem in the clinical domain, where the meaning of the result is crucial to guarantee the accuracy and trustworthiness of the clinical diagnosis and its decision-making process [24].

Interpretability is key in making a medical diagnosis for a number of reasons. First, doctors and medical personnel have been used to making decisions following tried-and-tested medical ideas, guidelines and experience. They should be able to infer how the DL model made its conclusion when presented with a diagnosis or recommendation to confirm that the model is correct, and to provide valid evidence of it. This is particularly relevant when a decision needs to be made in conditions where consequences are significant, e.g., has a patient access cancer or does they need surgery [6]. Even if the model performs very well, it will not likely be welcomed if it reveals no mechanism for how decisions are made, as physicians are the ones who are held to task for the results.

DL models that are interpretable can be used to detect mistakes or discrepancies in the model. Bias is often an inherent property of medical data caused by the differences in data collection methods, patient populations, or health access [26]). With DB training, a DL model could make wrong or biased predictions that disproportionately impact certain groups of patients. For example, a model based on most information from a single population segment may not perform as well in a more diverse population [20]. An interpretable model facilitates clinical insight into the model output, the potential for bias and acceptable modifications, which can make for improved and fairer diagnostics [27].

A model's interpretability can aid accountability by allowing users to investigate how a model reacts to inputs. A plausible explanation however, does not prove that a tool meets legal or ethical requirements, and guidance varies for jurisdiction and use. Medical AI requires more than explanation methods [13], [28]: explanation methods are not enough to give medical AI the documented validation, information for users, and oversight of the human-AI team it needs.

Trust building between patients and health care providers is another important factor of interpretability. For patients, the process of healthcare decision-making should be fully explained that relies on facts and good reason. Patients, naturally, might have questions about how the diagnostic decision is made should the DL model play a role in that, as in this instance [5]. For instance, when a DL system suggests a certain treatment, a patient and her family would want to understand why this is the recommendation. Making the decision making process understandable with an interpretable model improves trust in the technology and medical team by allowing patients to understand the process [7].

Interpretability plays a role in the ongoing enhancement and education of healthcare professionals. Clinicians can start to enrich their understanding about disease patterns, treatment success, and outcomes when working with interpretable models [29]. The relationship between AI-powered tools and healthcare practitioners is mutually beneficial, forming a cycle in which one group learns while the other refines the system.The partnership between AI tools and healthcare practitioners is symbiotic; the machine learning system gains valuable insights from data and the medical practitioner adds the

valuable real-world clinical experience to the enhancement of this system. But when there's no explainability, that cycle cycles in one direction, making it difficult to make the AI benefactor in furthering medical knowledge and to improve patients' outcomes [3].

Some of these methods have been proposed to make DL models interpretable, such as using attention mechanisms [5] or saliency maps [6] and model-agnostic approaches like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations) [4], but there are ongoing challenges. Many of these methods give "post hoc" explanations, that is to say, information about what the model did predict once it has made the prediction, but not information that made the model bring about this prediction. Another potential problem is that a human-readable explainability might miss the important details of the model's decision-making process [23]; there is also the risk that some interpretability method could over-simplify the DL model's complexity so that it can be expressed in a way that is understandable to humans.

Interpretability of DL models is significant in the medical diagnosis use case because it has several aspects. It can assist healthcare workers to recognize and evaluate these models in practice, to recognize and work out biases, and to stay in line with regulatory and moral guidelines. In addition, interpretability can help foster patient trust and promote the effective integration of AI into the healthcare workflow, which can enhance patient care and improve medical knowledge. To be successful in their integration into clinical practice, models developed will need to be accurate and interpretable as DL becomes an increasingly important role in healthcare.

C. Research Problem and Objectives

Deep learning (DL) models have demonstrated tremendous potential to improve the accuracy and efficiency of medical diagnosis, tremendously benefitting the healthcare sector. Yet, although these achievements have been made, there is still a major missing element – the understanding of the underlying decision making processes of these models. Many are concerned in the medical community about this “black box” component of DL, whereby even those who developed the models don't necessarily know exactly how they came up with their models' "result". Both healthcare providers and patients, as well as regulatory authorities, require transparency, accountability and interpretability in the medical decision-making process [28]. The potential for such power is not going to be embraced by end users in the clinical arena if the models it represents are not easily interpretable, that is, if they cannot provide clear justification for the recommended actions. A research issue then focuses on how to keep models with high predictive accuracy in medical diagnosis important for the diagnosis and at the same time keeping them interpretable for users [23].

This study's primary research goal is to investigate techniques that boost the interpretability of deep learning models without sacrificing predictive accuracy when making medical diagnoses. In healthcare, interpretability is not just a technical requirement, but also a practical one that can aid in building trust with clinicians and boost patient safety [24]. The goal of this study is to partially fill this gap by devising a system that produces more comprehensively interpretable outputs and is able to deliver clinical actionable results that are more readily understood by healthcare professionals and can help them make more confident statements.

II. DEEP LEARNING IN MEDICAL DIAGNOSIS

To pinpoint the challenges to the interpretability of DL models in medical diagnosis [7]

Study and be inspired by recent approaches to enhance interpretability including attention networks, feature importance, and post-hoc explanation approaches like Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) [4; 11].

To inspect the trade-off between interpretability and power in medical diagnosis, specifically to ensure that models' performance levels stay high while they offer interpretable learning [3].

Suggest a framework or best practices for creation of interpretable DL models that can be used in specific medical settings and help successfully integrate them into the clinical workflow of these models [25].

To address the question of how interpretability of models can influence the decision-making process of the clinician, the results of the algorithm, and trust in AI-assisted diagnoses [6].

These goals will facilitate a merging of the latest graphical processing technologies and deep learning with the clinical, moral and regulatory demands of healthcare in order to close this gap. This will not only give insights into how to make DL models more interpretable but will also give guidance on how to develop trustworthy, reliable, up to date models fitting the needs of medical practitioners.

II. Medical diagnosis using deep learning concepts. With its potential to analyze large volumes of intricate medical information and uncover patterns that are often imperceptible to the naked eye, deep learning (DL) has become one of the most revolutionary technologies in the realm of medical diagnosis. The advancement of neural networks gives DL models the potential of providing valuable services in disease diagnosis, predicting disease outcomes, and guiding clinical decision-making [1]. The models are prowessed by performing image recognition, natural language processing, and time-series analysis, all of which are critical for analyzing medical images, interpreting clinical notes, and monitoring patient vital signs over time, respectively [14]. The advantages of DL in health care are large, but there are important difficulties, such as the interpretability and transparency of DL models [28].

Medical imaging is one of the most important and successfully applied areas of deep learning within the field of medical diagnosis, in which convolutional neural networks (CNNs) excel. CNNs have been successfully applied to image-based diagnosis in many health-related fields such as the detection of cancer, cardiovascular disease and neurological disorders, and automatic feature extraction of content from images like X-rays, MRIs and CT scans, demonstrating an impressive level of accuracy. For instance, for the past few years tumor detection in mammograms and classification of skin lesions as either benign or malignant using DL models have been successful in being able to match if not surpass the work of seasoned dermatologists [30]. These models can be quite useful in medical imaging, such as diagnostic radiology and pathology, where the manual interpretation of medical images tends to be tedious and prone to error, especially in the case of high-dimensional datasets [16].

In addition to imaging, DL models are also employed in the field of genomics and personalized medicine, where they can serve as a tool for predicting the risk of a disease based on information about a person's genes. Genetic

sequences have also been analysed in the interest of other applications, including recurrent neural networks (RNNs) and deep generative models, to find associations between gene variants and diseases like rare genetic disorders, diabetes and cancer [31]. This ability to foretell disease susceptibility, using genomic information, is laying the foundation for more personalized medicine, in which treatment can be customized to reflect an individual's distinct genetic composition. The difficulty in these models, however, has raised concerns about its interpretation, as healthcare providers need to comprehend the basis behind these forensics predictions to effectively carry them out in clinical practice [3].

Although recent advances in deep learning have demonstrated its utility in boosting diagnostic accuracy, many deep learning models possess the "black box" characteristics, which is a challenge for the use of deep learning in clinical contexts [23]. Most often, even the model developers will not be able to clarify or explain how a given decision was taken and be persuaded to trust or verify their results without also having to understand how the model was developed. In medicine, for example, this lack of interpretability is very problematic since decisions made in medicine can be life or death decisions. For instance, a DL model can accurately describe the probability of a patient suffering from a disease, but if the doctor is not able to explain what factors led the model to that prediction, he/she might not want to act on it and take the chance [5].

To combat this interpretability problem in medical diagnosis in deep learning there are several attempts underway. For complex models, post hoc explanation methods like Local Interpretable Model-agnostic Explanations (LIME) [4] or Shapley Additive Explanations (SHAP) [11] have been developed to explain the most important features determining a specific prediction. These approaches enable the physician to learn from the image and identify the most significant components (certain regions of an image or certain genetic markers) that assisted in the diagnosis. These approaches help to explain the complexity of deep learning models, but they are limited in their capacity to fully explain the inner workings of a deep model and may have the potential to oversimplify the model [3].

A research possibility to take is to adopt transparent or intrinsic interpretable models when applicable. An attention mechanism reveals attention weights, but does not necessarily give a necessary explanation of a prediction, their connection with the output must be tested. Also, unlike the assumption that these simpler student models replicate their reasoning, surrogate trees and simpler student models demand measured fidelity to the original predictor [28,32].

Another challenge is the ethical and legal implications of using deep learning in medical diagnosis.Another challenge is the legal and ethical issues surrounding the use of deep learning in medical diagnosis. One concern with the application of AI in healthcare is privacy issues, data security, and potential for bias in the outcomes of AI models [26]. It is because medical information like EHRs are frequently incomplete or imbalanced, and if the DL models are trained in accordance with heterogeneous datasets, they can generate biased or unfair outcomes mainly affecting certain populations [27]. Furthermore, the regulatory authorities, such as the U.S. Food and Drug Administration (FDA), have still not learnt from what they should when it comes to

evaluating and approving AI-assisted clinical diagnostics devices, and current evaluation methods might not be appropriate for the ongoing learning cycles of DL models [16].

Ultimately, deep learning has the potential to transform medical diagnosis, delivering more precise, efficient and personalized diagnosis tools. Today, DL models have already been showing their utility in clinical practice, ranging from analyzing medical images to predicting disease risk based on genetic details. But there is still much to be done in terms of interpretability, trust and regulatory approval before such technologies can be widely adopted within health care systems. DL models should interpret themselves, be transparent, and be ethical; ongoing research into making DL models interpretable, transparent, and ethical will be critical to its safe and effective use for medical diagnosis.

For the explanation methods, their location in the diagnostic process is shown in Figure 1. Executing a model output for evaluation of its predictive validity is not a substitute for the evaluation of its quality of explanation and vice versa.

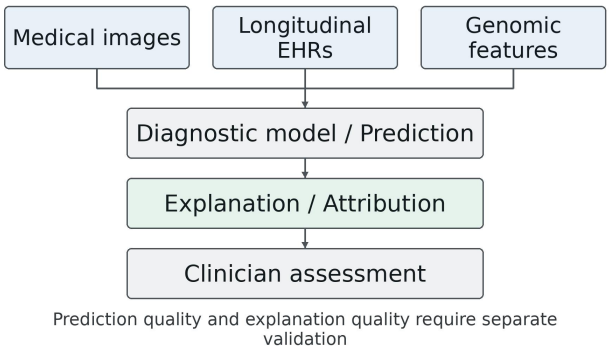


Figure 1. Relationship between medical inputs, diagnostic prediction, explanation, and clinician assessment.

A. Model Families and Architectures

1) Convolutional Neural Networks

CNNs are mainly used to handle grid-related data, like images, since they rely on spatial hierarchical relationships. They are used in the identification of features such as edges, shapes, and textures in images, which is a task that they do automatically. CNNs are widely used in medical diagnosis for various applications, including tumor detection in MRI, CT scans, and X-ray image analysis.

2) Recurrent Neural Networks

RNNs are a family of neural networks specifically adapted to sequential data which are well-suited to problems with time-series data and longitudinal health data. They can store information in their internal memory along input sequences, thus modeling a temporal dependency. In the health sector, RNNs are also prevalent for generating patient status predictions from historical medical records or fitting patients' vital signs through time.

3) Generative Adversarial Networks

GANs have two parts: a generator that generates fake data and a discriminator that attempts to differentiate real data from fake data. They are applied in medical imaging to synthesize datasets for training medical image models, especially when the medical image data are unlabeled. GANs are used to enhance the data augmentation process in medical applications to create realistic images for research and model training purposes.

4) Autoencoders

An autoencoder is an unsupervised deep learning model that is employed to compress and reconstruct high dimension data to reduce the complexity of data. In the field of healthcare, they are put to use in anomaly detection in medical information, or perhaps image denoising in MRIs. A critical function of them is dimensionality reduction which makes complicated medical data more manageable without losing vital information.

5) Long Short Term Memory Networks

The LSTMs are a subclass of the RNNs which help to overcome vanishing gradient problem, and are particularly suitable for sequential data with

long-term dependency. These networks are useful in analyzing patient history, medication records, or in predicting the progression or evolution of disease over time. In the field of electronic health record (EHR) analysis, LSTMs are widely employed in predictive healthcare applications. Electronic health record (EHR) analysis is another area where LSTMs are widely used, such as to predict the volume of hospital admissions.

6) Transformer Networks

Transformers are models based on attention, which enable them to process sequences of data in parallel rather than in sequence like RNNs, making them more efficient for sequential data. These have become popular for use with natural language processing applications, like analyzing clinical notes or medical reports. In medical diagnosis, transformers are used for their ability to large-scale interpret medical texts, such as patient records and pathology reports, for more accurate and context-sensitive diagnoses [33].

7) Deep Belief Networks

DBNs consist of several layers of latent variables, which allows for the unsupervised learning of high level feature representations that are generative models. In the field of healthcare, they can be applied to identify patterns in complex data, including that from functional MRI scans or gene expression data. For example, DBNs can be used to learn hierarchical representations of medical data, which is useful for medical applications such as disease classification or prognosis prediction.

8) Multilayer Perceptrons

The simplest type of feed-forward neural network that has more than one layer of fully connected neurons, thus it is one of the most flexible among feed-forward neural networks for prediction is called a multilayer perceptron (MLP). Less complex than CNNs or RNNs, MLPs are applied to medical diagnostics, such as binary classification of disease presence or patient risk scores. In structured data analysis, for example, generating forecasts on patient outcomes from laboratory measurements, demographic data or clinical features, MLPs are used.

Figure 2 shows that there are generic pathways for imaging and sequential-data pathways. The diagram does not indicate a trained checkpoint, layer numbers, channel width or comparative accuracy.

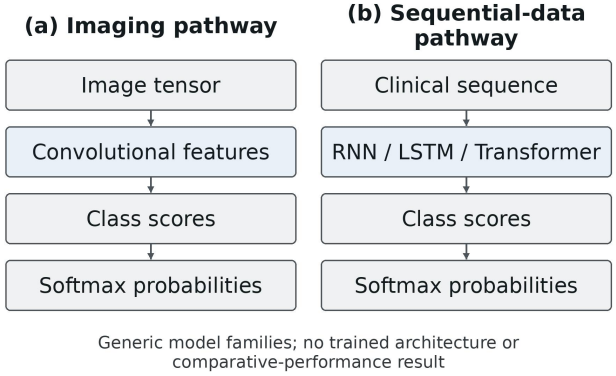


Figure 2. Generic CNN and sequential-model pathways from input to class probabilities.

B. Applications in Medical Imaging and Diagnostics

1) Tumor Detection and Classification

Deep learning, particularly using Convolutional Neural Networks (CNNs), has revolutionized the detection and classification of tumors in medical imaging. CNNs are able to detect multiple cancer types including brain, lung, and breast cancer from different modalities, MRI, CT and mammography. These models are able to detect patterns and features which may be its core power for early detection, that the human radiologists may not pick; in this regard, these models can deliver high accuracy in their early detection. To assess whether model-assisted detection leads to better survival rates a long-term prospective

clinical study is required, not classification performance.

2) Cardiovascular Disease Diagnosis

In healthcare, especially in cardiology, deep learning models are used for the diagnosis of cardiovascular diseases based on medical images such as echocardiograms and angiograms. CNNs and Recurrent Neural Networks (RNNs) identify and classify heart diseases such as arrhythmias, coronary artery disease and heart valve defects. These systems can analyze vast amounts of data from patients and generate predictive outcomes that cardiologists can review along with existing patient information, which can potentially inform diagnostic and treatment decisions as well as aid patients through their care.

3) Diabetic Retinopathy Screening

Deep learning models have proven to be very effective in the diagnosis of diabetic retinopathy (DR), a diabetes-related condition of the eyes. In diabetic retinopathy, the retina is the part of the body that can be examined by CNN-like models to auto detect the signs of the disease, for example in the form of microaneurysm and hemorrhages that can cause blindness if left untreated. In areas where access to ophthalmologists is poor, automated deep learning systems could enable early detection and treatment among those who are at risk.

4) Lung Disease Diagnosis

Deep learning is used extensively for the detection of pneumonia, tuberculosis, lung cancer, from chest X-rays and CT scans. These models can identify minute abnormal changes in the lungs that might be due to a disease. For lung cancer diagnosis, CNNs can identify discrepancies among nodules, which is crucial for the early diagnosis of cancer. Although validated, the implementation of such systems in routine radiological practice helps to lighten the work burden of the radiologists and can help in timely clinical assessment.

5) Breast Cancer Detection

Deep learning has come a long way in breast cancer screening using mammograms. CNNs are employed to examine mammographic images and spot early indicators of breast cancer like irregularities in the breast tissue, or the presence of plenty of calcium deposits. A false-positive and false-negative rate should be calculated for the target population that was used in the screening test, and compared with a valid reference test. Clinical evaluation of the effect of AI screening on early detection and patient outcomes will need to be carried out separately.

6) Neurological Disorder Detection

Deep learning has come a long way in breast cancer screening using mammograms. CNNs are employed to examine mammographic images and spot early indicators of breast cancer like irregularities in the breast tissue, or the presence of plenty of calcium deposits. A false-positive and false-negative rate should be calculated for the target population that was used in the screening test, and compared with a valid reference test. Clinical evaluation of the effect of AI screening on early detection and patient outcomes will need to be carried out separately.

C. Current Interpretability Challenges

Table 1 summarizes the challenges identified in this review, their possible diagnostic impact, and the corresponding response and research direction.

Table 1. Interpretability challenges, possible diagnostic impact, and research responses.

Challenge	Description	Impact and example	Response and research direction

Challenge	Description	Impact and example	Response and research direction
Black Box Nature of Models	Difficulty in understanding model decisions	Reduces trust among clinicians; AI recommendations unclear in critical cases	Use of explainable AI techniques; Development of interpretable models
Trade-off Between Accuracy and Interpretability	Complexity vs. simplicity in models	May limit adoption of effective models; Simple models may underperform	Hybrid models combining both aspects; Research on balancing accuracy and interpretability
Complexity of Medical Data	High-dimensional and diverse data sources	Challenges in validating AI outputs; Difficulties interpreting complex imaging results	Data preprocessing and feature selection; Exploration of simplified data representations
Lack of Standardized Evaluation Metrics	No universal metrics for interpretability	Inconsistent model assessments; Difficulty comparing model performances	Establishment of standard metrics; Focus on developing comprehensive evaluation frameworks
Bias and Fairness Concerns	Models may learn biases from data	Risk of inaccurate diagnoses for minority groups; Disparities in diagnostic accuracy	Bias detection and correction methods; Research on equitable AI practices

III. TECHNIQUES FOR INTERPRETABILITY

Figure 3 distinguishes methods by their required access to the model. Table 2 summarizes explanation scope and limitations. These are methodological categories, not numerical rankings of diagnostic accuracy.

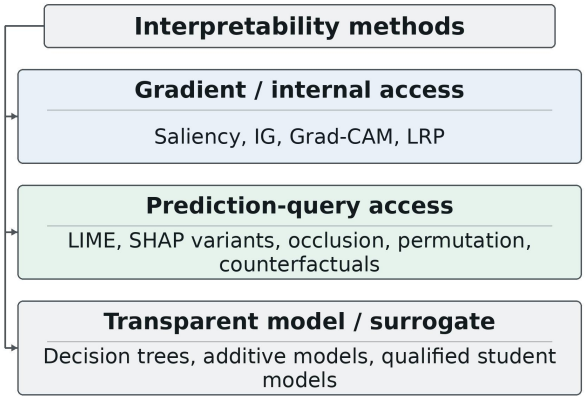


Figure 3. Interpretability methods grouped by gradient or internal access, prediction-query access, and transparent-model design.

A. Method Overview

Table 2. Interpretability techniques, required access, explanation scope, and limitations.

Method	Model access	Scope	Output	Limitation
Saliency	Input gradients	Local	Pixel or feature gradient	Sensitivity is not contribution; may be noisy.
Integrated gradients	Input gradients	Local	Signed baseline-relative attribution	Depends on baseline and integration accuracy.
Grad-CAM	Activations and gradients	Local	Coarse class-specific heatmap	Localization is not lesion segmentation.
LRP	Layer-level access	Local	Input relevance scores	Propagation rules and architecture matter.
LIME	Prediction queries	Local	Weighted interpretable surrogate	Sampling, kernel, and fidelity matter.
SHAP	Variant-dependent	Local; aggregate summaries	Shapley feature contributions	Background distribution and dependence matter.
Occlusion	Prediction queries	Local	Score change under masking	Masks can create implausible inputs.
Permutation importance	Prediction queries and labels	Dataset-level	Held-out performance change	Correlated features can distort importance.
Surrogate / distillation	Teacher prediction queries	Local or global	Simpler predictor	Simplicity and teacher fidelity must be tested.
Counterfactuals	Queries; sometimes gradients	Local	Input change yielding target class	Not a causal treatment recommendation.

Layer-wise relevance propagation redistributes a score of the model towards the inputs while implementing certain propagation rules [34]. Alternatively, a surrogate model comprised of a

decision tree can be fit to model the outputs. Distillation involves training a student to match a teacher, but the simpler is not translatesable and agreeability needs to be evaluated with the model [35]. These are not a replacement for attribution and perturbation methods that are detailed below.

B. Gradient and Feature Attribution Methods

1) Saliency Maps

A saliency map is a visualization technique to identify the most salient/important regions of its input that affect its prediction. Saliency maps measure local score sensitivity to input image changes by computing gradients of the output with respect to the input image. This method can be useful in medical imaging scenarios to facilitate the interpretation of a model, the scope of which is an MRI or CT scan, to see which areas are being marked as important: this helps in the medical imaging assessment of the model. Saliency maps may be able to uncover patterns and features that are not obvious and support the informed decision-making of healthcare professionals based on model outputs.

2) Grad CAM

Grad-CAM is a neat technique for generating visual explanations to deep learning systems, mainly the Convolutional neural network (CNN). Grad-CAM employs the gradients of the output class for the last convolutional layer to produce heatmaps that geometrically highlight image regions the most relevant to the prediction of the model. Grad-CAM could be used in the medical field for diagnostics to focus the areas of interest inside the imaging data, like the lesions inside radiological imaging, in order to learn the focus areas of the model. This could help radiologists study a model's sensitivity, where complex model predictions can be compared to clinical understanding.

3) Integrated Gradients

Integrated gradients allow the direct attribution of the output of a neural network to its input features by measuring the change in the output as the input is shifted along a straight line from baseline to the actual input. This approach combines the gradients over a path between the baseline and input and overcomes saturation issues at the endpoints. For medical use, an integrated gradient could show the way in which particular clinical factors contribute to the prediction, improving the possibility of interpretation and more complex decision making for the individual patient.

A good saliency map doesn't necessarily have to rely on the learned model, as they can be completely visual-plausible. Checks based on randomisation of the parameters and of the data are helpful and visual checks are not enough to guarantee faithfulness [36].

C. Model Agnostic and Perturbation Methods

1) LIME

One of the most commonly used LIME techniques is approximation, which uses a simplified interpretable model to approximate the machine learning model locally and thereby forming interpretable explanations for it. LIME then adapts its input to the target sample and compares how much it changes the output values, which then allows them to construct a linear model that models the behavior of this target sample. LIME can help doctors gain insights to make understanding the decisions made by AI systems in medical diagnosis more transparent, giving better insights in the reason for AI decisions in medical diagnosis and the individual patient predictions.

2) SHAP

SHAP is a game-theoretic, model-agnostic as well as model-specific feature-attribution framework for assigning an importance value (SHAP value) for each feature that reflects its contribution to the model's prediction. SHAP computes the average marginal contribution of a feature to interpretability globally of the way that it contributes to the cost and utility of the value function under consideration. This is especially valuable for medical use where diagnosis is concerned as it is possible to discover the determining attributes in that diagnosis. Healthcare practitioners can utilize SHAP to gain insight into intricate artificial intelligence models, thereby improving healthcare decision-making and boosting trust in AI systems.

3) Occlusion Sensitivity

Occlusion sensitivity involves systematically occluding parts of the input image to observe the impact on the model's predictions. This method, by identifying the importance of the subregion for an accurate classification, gives insights into the regions that are important. For clinical diagnosis, occlusion sensitivity can be used to analyse the changes in scores for masked image regions, thereby improving the comprehensiveness of deep learning diagnostic models.

4) Permutation Feature Importance

The intuition of Permutation feature importance is to measure the effect of a feature on an overall model by shuffling the values in each feature columns, after which the model's performance is evaluated. The approach is simple to perform and model independent and as a result, it is easy to gain insight into what aspects of the model are most important for predicting. Its application in healthcare can aid in prioritizing clinical factors that affect diagnosis, which enhances transparency and aiding physicians and healthcare providers with decision-making.

5) Counterfactual Explanations

To gain insight into what changes in the input data would cause a change to the model output, counterfactual explanations (CFEs) can be used. This approach can be used as a way to gain knowledge about the boundary of a model, through the creation of other scenarios. Counterfactual explanations in a medical setting can be used to show how modifications to the characteristics of patients may impact the model results, but not to establish a treatment effect, encouraging greater interaction between the doctor and the AI tool.

It is important to understand that, unlike model-agnostic black-box approaches, integrated gradients does not use derivatives of the value that it is trying to learn. There are multiple implementations of SHAP, some are restricted to solving only prediction queries, others take advantage of the model structure. Individual SHAP explanations are local, and it is known that a set of local values can be aggregated and summarized to capture the behavior of a model [37] [11].

D. Comparative Analysis of Interpretability Techniques

1) Transparency and Accuracy

There should be a distinction between transparency and prediction.Transparency and predictive performance should be assessed independently. A post-hoc explanation alone does not necessarily make the model more accurate and an accuracy degradation for explainable models does not follow when measured against the accuracy of their black-box counterpart. The comparison to be used should be the same task, with the same held-out population under the

same reference standard and both should be measured on a model of explanation fidelity and usability [28].

2) Local and Global Interpretability

Local explanations refer to an explanation for an isolated prediction and global summaries include a description of behavior through a series of examples. LIME is a local program. Additionally, SHAP does local attributions that can be aggregated into a dataset-level summary. The absence of a specific evaluation design of these two types renders them invalid to establish clinical validity [4, 11].

3) Computational Efficiency

Different interpretability techniques have varying computational costs. Some methods such as saliency map can be computed in short time while other methods such as SHAP calculation can take more resources as the method is combinatorial. The choices of efficient and effective but interpretable techniques are very important in clinical settings, especially when time and compute resources are limited and it is critical to practically implement the choices in medical workflow.

4) Ease of Understanding

Explanations given by different techniques can affect acceptance of these techniques for their use in clinical practice. Some techniques such as decision trees and feature importance are more intuitive and easier to explain for the clinicians, while some other methods such as integrated gradients might need more explanation and training. It is essential to ensure that interpretability techniques are easily accessible and understandable for healthcare professionals, with the aim of establishing trust and improving collaboration with AI systems.

5) Domain Adaptability

An important aspect to consider is the versatility of interpretability methods in diverse medical fields. Some methods might be appropriate for particular applications or purposes, e.g., imaging analysis, or patient risk prediction, etc., and others may be suitable for more generic applications or purposes. The applicability of various techniques in different clinical situations, their use and optimization for the improvement of the interpretability of medical diagnostics, is significant.

E. Challenges in Clinical Implementation

1) Clinical Workflow Integration

The challenges of adopting interpretability techniques into the existing clinical workflow are great. Time is often short on hand and may lack the technical skills to properly use complex interpretability methods by clinicians. These techniques need to be easily implemented (adoptable, useful), while being smoothly woven into the practice.

2) Stakeholder Needs

The needs of interpretability differ among users in the healthcare sector, whether they are clinicians, patients or governmental decisions makers. A clinician might give detailed explanations of disease mechanisms for specific diagnoses while a patient might be looking for simpler explanations he can understand. Curating for this variability involves creating flexible interpretability solutions that work across a range of audiences, achieving effective communication and understanding across the spectrum.

3) Trust and Acceptance

The ability of AI systems to provide explanations of their predictions is key to achieving trust in AI. If the decision-making process is "black box" and it is unclear, that could be a problem with clinicians relying on an AI-generated diagnosis. To achieve

confidence and uptake of AI technologies in healthcare, there is a need for the reliability and robustness of AI interpretability methods to be demonstrated, so that they can help the clinical workflow as a whole without negatively impacting patient services.

4) Regulatory Standards

Similarly, the regulatory landscape regarding interpretability and transparency of AI-driven healthcare solutions is continually evolving.Regulatory requirements for the interpretability and transparency of AI-based healthcare solutions are also evolving. This can be a tricky task to follow in order to keep effective and clinically relevant techniques that are interpretable for the evolving guidelines. Navigating the regulatory environment is a process that stakeholders must undertake to create interpretability ~ or viable solutions ~ that are compliant with the relevant regulations and will yield effective insights.

5) Model Complexity

Deep learning models are complex, and interpretability can be challenging. The more sophisticated a model is, the more difficult it will be to understand what operations it is using to make a decision. One of the research and practical challenges to get these models to work effectively without losing interpretability is quite a complex task to undertake and improve the use of AI for medical diagnostics.

IV. MATHEMATICAL METHODOLOGY AND EVALUATION FRAMEWORK

In this section the methods presented in Section 3 are formalized and a manner of assessment is proposed. It doesn't add a trained diagnostic model or a done clinical trial. Prediction scores, baseline- relative contributions, gradients, and explanation fidelity maintained separately to gauge their role without accreditating an appealing pictographic view as a clinical merits.

A. Mathematical Notation and Methodological Relevance

The following notation is used in the equations and algorithms (defined in Table 3). Each input x and class score f_c is linked to a specific prediction target. Each medical modality has its unique choices of baseline x_0, background distribution Z, and perturbation rules that must be specified.

Table 3. Mathematical notation and relevance to the interpretability methodology.

Symbol	Definition	Methodological relevance
x, d, x_0	Input vector, feature count, and reference baseline.	Represent patient features or flattened images; define the attribution reference.
θ, f_θ, f_c	Fixed model parameters, class-score vector, and score for class c.	Keep the model and explained target fixed during interpretation.
$C, p_c, \hat{y}$	Class count, Softmax probability, and predicted class.	Separate score explanations from classification decisions.
S_i, IG_i	Signed input gradient and integrated-gradients attribution for feature i.	Distinguish endpoint sensitivity from baseline-relative contribution.
m, τ	Integration-step count and path coordinate in [0, 1].	Control numerical approximation along the baseline-to-input path.
$\mathcal{F}, T, Z$	Feature-index set, coalition, and background random vector.	Specify which features are present and how absent features are represented.
$v(T), \phi_i, \phi_0$	SHAP value function, feature attribution, and reference score.	Define an additive explanation under a stated background distribution.
z_q, z_q', Q	Perturbed input, its interpretable representation, and perturbation count.	Fit a local surrogate with paired model and interpretable inputs.
π_x, κ, g, Ω	Locality weights, kernel width, surrogate, and complexity penalty.	Balance local fidelity and simplicity in LIME.

Symbol	Definition	Methodological relevance
A^k, H, W, K	Convolutional channel k, spatial height/width, and channel count.	Define the feature-map dimensions used by Grad-CAM.
$\alpha_k{}^c, L^c$	Class-specific channel weight and raw Grad-CAM map.	Aggregate class-score gradients and weighted activation maps.
$\widehat{L^c}, \varepsilon$	Display-normalized heatmap and positive numerical safeguard.	Rescale for visualization only; avoid a zero normalization denominator.
B, Δ_B	Masked feature/region set and resulting class-score change.	Assess perturbation response without interpreting it as a causal effect.
$\mathcal{D}, N, R$	Held-out evaluation data, sample count, and permutation repetitions.	Evaluate performance loss from shuffling one input feature.
x_n, y_n	Input and integer class label for held-out example n.	Select the probability assigned to the observed class in cross-entropy.
Err, PI_i	Mean classification loss and permutation importance.	A positive loss increase indicates reliance under the chosen perturbation.
$x^{\mathrm{cf}}, c^\star, \mathcal{V}$	Counterfactual input, desired class, and clinically feasible input set.	Constrain the model-output change without recommending treatment.
TP, TN	True-positive and true-negative counts.	Define binary metrics for a prespecified positive diagnosis.
FP, FN	False-positive and false-negative counts.	Report error types and document any undefined metric denominators.
$y_{n,c}, p_{n,c}$	One-hot outcome and predicted probability for sample n and class c.	Assess probability calibration with a multiclass Brier score.
$e(x), \delta$	Explanation vector and a valid input perturbation.	Assess stability with the same target and explanation settings.
$r(t), X, a$	Analytical scalar function, illustration grid, and its four-value axis vector.	Demonstrate attribution mathematics without using patient data.

Scalar indices are local to their equations; illustration variables are not clinical measurements.

B. Prediction Target and Gradient Attribution

Let the fixed predictor map a d-dimensional input to C class scores:

$$x \in \mathbb{R}^d, \quad f_\theta(x) = [f_1(x), \ldots, f_C(x)] \in \mathbb{R}^C \qquad (1)$$

Softmax converts these scores to probabilities and selects the predicted class with a prespecified tie-breaking rule. Explanations below target the prespecified pre-Softmax score f_c, rather than silently switching between a score and a probability:

$$p_c(x) = \frac{\exp(f_c(x))}{\sum_{j=1}^{C} \exp(f_j(x))}, \quad \hat{y}(x) = \arg\max_c p_c(x) \qquad (2)$$

Signed saliency is the derivative of the target score with respect to each input feature. Its magnitude is often displayed as a saliency map; magnitude alone loses the direction of the response:

$$S_i(x) = \frac{\partial f_c(x)}{\partial x_i}, \quad \mathrm{Saliency}_i(x) = |S_i(x)| \qquad (3)$$

Integrated gradients (IG) instead measures baseline-relative contribution by integrating along the straight path from x_0 to x [37]:

$$\mathrm{IG}_i(x;\, x_0) = (x_i - x_{0,i}) \int_0^1 [\nabla f_c(x_0 + \tau(x - x_0))]_i \, \mathrm{d}\tau \qquad (4)$$

For the exact integral and a suitable differentiable path, the sum of feature attributions equals the score difference. Numerical implementations should report the completeness residual:

$$\sum_{i=1}^{d} \mathrm{IG}_i(x;\, x_0) = f_c(x) - f_c(x_0) \qquad (5)$$

A midpoint approximation with m integration steps is

$$\mathrm{IG}_i(x;\, x_0) \approx \frac{x_i - x_{0,i}}{m} \sum_{j=1}^{m} [\nabla f_c(x_0 + \frac{j - \frac{1}{2}}{m}(x - x_0))]_i \qquad (6)$$

The baseline must be meaningful for the input modality. A zero-valued image or clinical vector is a mathematical reference, not necessarily a healthy patient, an absence of disease, or a plausible clinical state. Baseline sensitivity and integration convergence should therefore be tested alongside explanation validity.

C. SHAP and Local Surrogate Explanations

For feature set $\mathcal{F}$ = {1, ..., d}, SHAP assigns a feature its average marginal contribution over

coalitions T that omit that feature. The value function must be specified [11]:

$$\phi_i = \sum_{T \subseteq \mathcal{F} \setminus \{i\}} \frac{|T|!\,(d - |T| - 1)!}{d!} [v(T \cup \{i\}) - v(T)] \quad (7)$$

One explicit choice is an interventional background value function, which combines the observed features in T with the corresponding remaining features of a background vector Z:

$$v(T) = \mathbb{E}_Z[f_c(x_T, Z_{\mathcal{F} \setminus T})] \quad (8)$$

Under this definition, the full coalition retains the complete input, and efficiency gives the additive decomposition

$$f_c(x) = \phi_0 + \sum_{i=1}^{d} \phi_i, \quad \phi_0 = v(\emptyset) \quad (9)$$

This choice is not interchangeable with conditional SHAP, which conditions the missing-feature distribution on observed values. Correlated medical variables can create implausible combinations under marginal substitution. The implementation variant and background cohort must be documented, and SHAP values must not be interpreted as effects of treatment or causal mechanisms.

LIME fits an interpretable surrogate g near a selected input. Each perturbed model input z_q is paired with an interpretable representation. A weighted squared-error formulation balances local agreement and surrogate complexity [4]:

$$g^{\star} = \arg\min_g [\sum_{q=1}^{Q} \pi_x(z_q)[f_c(z_q) - g(z_q')]^2 + \Omega(g)] \quad (10)$$

A commonly used distance kernel is

$$\pi_x(z) = \exp\left(-\frac{\mathrm{dist}^2(x,z)}{\kappa^2}\right) \quad (11)$$

The distance metric, kernel width κ, valid perturbation generator, interpretable representation, and complexity penalty Ω control the explanation. Figure 4 shows why querying a fixed model and checking local fidelity are separate from fitting a simple explanation.

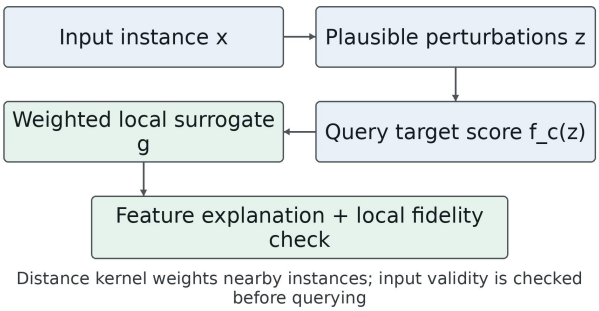


Figure 4. LIME workflow with plausible perturbations, target-score queries, distance weights, and local surrogate validation.

D. Class Specific Image Explanations

Grad-CAM uses a convolutional layer with K activation channels, each of spatial size H by W. For fixed class c, averaging score gradients over spatial positions yields channel weights [10]:

$$\alpha_k^c = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} \frac{\partial f_c(x)}{\partial A^k_{uv}} \quad (12)$$

The raw class-specific map is the rectified weighted sum of activation channels:

$$L^c = \mathrm{ReLU}\left(\sum_{k=1}^{K} \alpha_k^c A^k\right) \quad (13)$$

If a normalized display is needed, min-max scaling with a positive safeguard can be applied to the raw map:

$$\widehat{L^c} = \frac{L^c - \min(L^c)}{\max(L^c) - \min(L^c) + \varepsilon}, \quad \varepsilon > 0 \quad (14)$$

Figure 5 and Algorithm 1 show the computation. The target class stays fixed during gradient calculation, and training-only stochastic layers are disabled. Upsampling and normalization are display operations: the map is not a probability, a calibrated risk estimate, or a lesion segmentation.

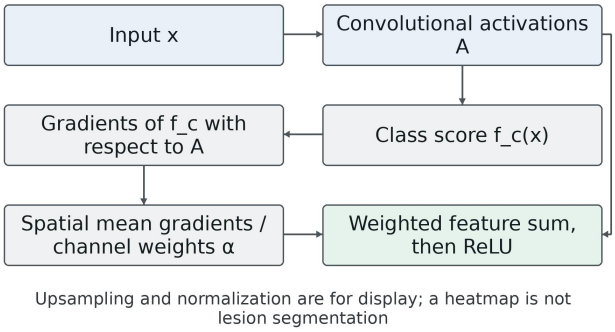


Figure 5. Grad-CAM calculation from class-score gradients, spatially averaged channel weights, and rectified activation sums.

E. Perturbation and Predictive Evaluation

For occlusion, replace features in a selected set or region B with their baseline values, leaving all other features unchanged and keeping the explained class fixed. The score change is

$$\Delta_B(x) = f_c(x) - f_c(x^{\text{mask(B)}}), \quad x^{\text{mask(B)}}{}_i = x_{0,i} \; for \; i \in B \qquad (15)$$

This is a response to a chosen artificial perturbation, not an effect of removing a disease or intervening on a patient. Masks must be checked for input validity. Perturbation-based image explanations should be tested rather than accepted solely because they look plausible [38].

For held-out data $\mathcal{D}$ with N labeled examples, x_n is the input and y_n is its observed integer class label. A mean cross-entropy error provides an explicit loss convention:

$$\text{Err}(\mathcal{D}) = -\frac{1}{N}\sum_{n=1}^{N} \log p_{y_n}(x_n) \qquad (16)$$

Permutation importance averages the loss increase after independently shuffling feature i across examples, using R repeated permutations of the held-out data [39]:

$$\text{PI}_i = \frac{1}{R}\sum_{r=1}^{R} [\text{Err}(\mathcal{D}^{\text{perm}(i,r)}) - \text{Err}(\mathcal{D})] \qquad (17)$$

The model is not retrained between these evaluations. Conditional or grouped shuffling may be preferable for dependent clinical features; the chosen policy must be stated. A negative estimate can occur from finite-sample variability or model behavior and must not be silently discarded.

A counterfactual explanation searches for a nearby feasible input that changes the predicted class. Its mathematical target is the model output, not a treatment benefit [40]:

$$x^{\text{cf}} \in \arg\min_{z \in \mathcal{V}:\, \hat{y}(z) = c^{\star}} \text{dist}(x, z) \qquad (18)$$

The feasible set $\mathcal{V}$ should constrain immutable attributes, units, clinically plausible ranges, and allowed changes. The desired class must differ from the original prediction; if no feasible input attains that class, report failure rather than an unconstrained solution. A class-changing input is not a recommendation to alter a patient characteristic; causal and clinical assessments are separate.

For a binary diagnostic task with a prespecified positive condition, standard confusion-matrix counts define accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (19)$$

Sensitivity and specificity describe different error types at a stated operating threshold:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (20)$$

The positive-class F1 score is

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \qquad (21)$$

For multiclass predictions, report the one-versus-rest and averaging conventions. Undefined ratios need an explicit zero-division policy. A multiclass Brier score measures squared probability error against one-hot outcomes:

$$\text{Brier} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} (p_{n,c} - y_{n,c})^2 \qquad (22)$$

Explanation quality needs separate measurements. For a local surrogate, a normalized weighted mean-squared error describes fidelity on independently assessed perturbations, provided the total locality weight is positive:

$$\text{FidelityError}_x(g) =$$

$$\frac{\sum_{q=1}^{Q} \pi_x(z_q)[f_c(z_q) - g(z_q')]^2}{\sum_{q=1}^{Q} \pi_x(z_q)} \quad (23)$$

For explanation vector e, a finite-difference ratio can summarize stability under a valid input change:

$$\text{StabilityRatio}(x, \delta) = \frac{\|e(x+\delta) - e(x)\|_2}{\|\delta\|_2 + \varepsilon} \quad (24)$$

Stability comparisons keep the class target, baseline or background, representation, and settings fixed. A genuinely changing prediction can justify a changing explanation, so the ratio is not a standalone validity test. Predictive calibration, faithfulness, clinician usability, and subgroup safety should be evaluated separately. Figure 6 and Table 4 summarize a prospective protocol.

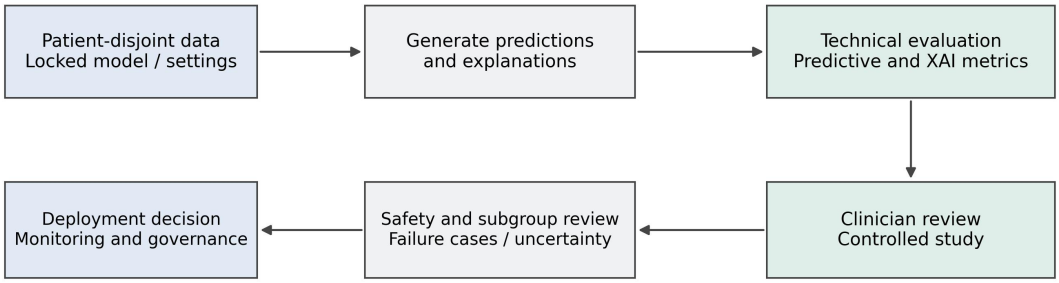


Figure 6. Proposed evaluation workflow separating held-out predictive performance, explanation checks, clinician study, and governance.

Table 4. Proposed evidence required to assess predictive and explanation quality.

Assessment	Measure or check	Required design control
Predictive performance	Accuracy, sensitivity, specificity, F1, threshold-based error counts	Patient-disjoint held-out cohort; fixed reference standard; stated threshold and averaging.
Probability calibration	Brier score and reliability analysis	Independent outcomes and probability predictions; subgroup assessment.
Explanation fidelity	Independent local surrogate error; plausible occlusion/permutation effects	Fixed target and model; medically valid perturbations; held-out explanation test points.
Explanation robustness	IG completeness/convergence; baseline sensitivity; randomization checks; stability	Prespecified settings and seeds; interpret changes alongside score changes.
Clinical usefulness and safety	Controlled clinician study; blinded outcome assessment, error review, usability and subgroup analysis	Clinician-only / clinician + AI / clinician + AI + explanations; ethics approval and uncertainty reporting.

F. Algorithms for Explanation and Evaluation

Algorithm 1: Class-specific Grad-CAM explanation.

Input: Fixed differentiable CNN; input x; target class c; convolutional layer with K channels; positive ε.

Output: Raw class map L^c and optional display-normalized map.

1: Set the model to evaluation mode and disable training-only dropout.

2: Run the forward pass; retain activations for the selected convolutional layer.

3: Select the fixed class score f_c; keep the same class during explanation.

4: Differentiate f_c with respect to the retained activation tensor.

5: **for** each channel k in {1, ..., K} do

6: Average its spatial score gradients to obtain α_k^c using Eq. (12).

7: **end for**

8: Compute the rectified weighted activation sum L^c using Eq. (13).

9: **if** a display map is required then

10: Upsample and normalize for display using Eq. (14).

11: **end if**

12: **return** the raw map and optional display map; do not label either as segmentation.

Algorithm 2: Prospective multimethod interpretability evaluation.

Input: Patient-disjoint data; trained fixed model; prespecified methods; targets; baselines/backgrounds; perturbation rules; evaluation criteria.

Output: Predictive metrics, explanation-check records, failure cases, and a separately designed clinician-study report.

1: Lock the model, test cohort, preprocessing, operating threshold, and random seeds.

2: Compute held-out probabilities, labels, and predictive/calibration measurements.

3: **for** each selected test input x do

4: Fix the explained class c and record the target score f_c.

5: **for** each prespecified explanation method do

6: Check required model access and document the reference and method settings.

7: **if** the method is IG then

8: Integrate with Eq. (6) and test completeness with Eq. (5).

9: **else if** the method is Grad-CAM then

10: Compute the raw explanation with Algorithm **1**.

11: **else if** the method is SHAP or LIME then

12: State background semantics or perturbation/kernel settings.

13: **end if**

14: Generate the explanation using its defined mathematical formulation.

15: Generate independent, clinically plausible explanation-test perturbations.

16: Record applicable fidelity, robustness, and score-response measurements.

17: Record unsupported, unstable, or misleading explanations as failure cases.

18: **end for**

19: **end for**

20: Summarize results by method, patient subgroup, and failure category with uncertainty.

21: Evaluate clinician usefulness in a separate controlled and ethically approved study.

22: **return** the technical records and clinician-study findings without mixing them.

G. Analytical Scatter Plots and Heatmaps

To distinguish gradient sensitivity from baseline-relative attribution without assigning unsupported values to clinical methods, consider the differentiable scalar function r(t) = t cubed with zero baseline. The derivative and exact attribution are

$$r(t) = t^3, \quad r'(t) = 3t^2, \quad \mathrm{IG}(t; 0) = \Delta(t) = t^3 \quad (25)$$

Figure 7 contains 21 analytical inputs evenly spaced from -1 to 1. Panel (a) compares derivative magnitude with exact IG: they have different meanings and are not interchangeable importance scales. Panel (b) compares exact IG with the baseline deletion score difference; their equality follows algebraically from the function, not from clinical validation or an empirical correlation test.

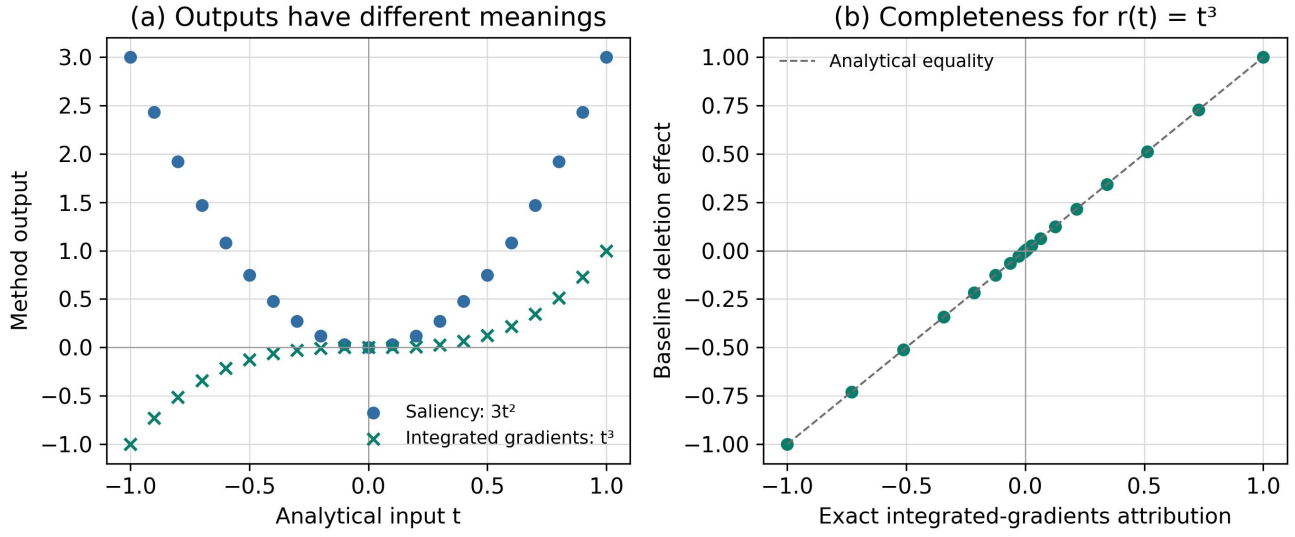


Figure 7. Analytical scatter plots for the cubic function with zero baseline; all points are computed, non-clinical illustrations.

Table 5. Selected exact values used in the analytical scatter illustration.

Input t	Score r(t)	Saliency 3t squared	Exact IG	Deletion effect
-1.00	-1.000	3.000	-1.000	-1.000
-0.50	-0.125	0.750	-0.125	-0.125
0.00	0.000	0.000	0.000	0.000
0.50	0.125	0.750	0.125	0.125
1.00	1.000	3.000	1.000	1.000

These are analytical function values, not patient measurements, diagnostic performance scores, or results of a newly trained model.

The heatmap illustration uses a four-by-four grid constructed from four evenly spaced axis values:

$$X_{uv} = \frac{a_u + a_v}{2}$$
$$a = (-1, \frac{-1}{3}, \frac{1}{3}, 1)$$
$$u, v \in \{1,2,3,4\} \quad (26)$$

For the scalar grid score obtained by summing the cubes of its entries, a zero-grid baseline gives exact pixelwise IG and saliency:

$$\text{GridScore}(X) = \sum_{u=1}^{4}\sum_{v=1}^{4} X_{uv}^{3},$$
$$\text{IG}_{uv} = X_{uv}^{3},$$
$$\text{Saliency}_{uv} = 3X_{uv}^{2} \quad (27)$$

Figure 8 displays the input, signed attribution, and nonnegative derivative magnitude on their stated color scales. Signed IG separates negative and positive contributions; saliency magnitude does not. The grid is not a medical image, and its maps are not Grad-CAM outputs or learned clinical evidence.

V. CLINICAL APPLICATIONS AND EVIDENCE CONSIDERATIONS

A. Interpretability in Medical Imaging

A domain where deep learning models are intensively used without compromising on their interpretability is the medical imaging domain. How a model responds to some parts of an image can be analyzed using Grad-CAM and saliency maps. For instance, if a single system, such as a lung-CT classifier, is evaluated, a class-specific heatmap can be used to explore this, but a highlighted area does not indicate malignancy or changes in treatment. These visualizations enhance AI outcomes and radiological analysis and aid radiologists in evaluating AI-assisted predictions. The interpretability embedding within imaging can not only help to understand the models' predictions but also it helps to make AI "work collaboratively" with healthcare professionals such that it can support their work, in cases where appropriate.

Examples of medical model applications in published skin-lesion classification and screening - for retinal imaging - have shown to help when used for diagnostic purposes, but not when used to explain improved results for a particular use case [14], [18]. If a comparison is to be an explanation-specific, then it requires its own evaluation design.

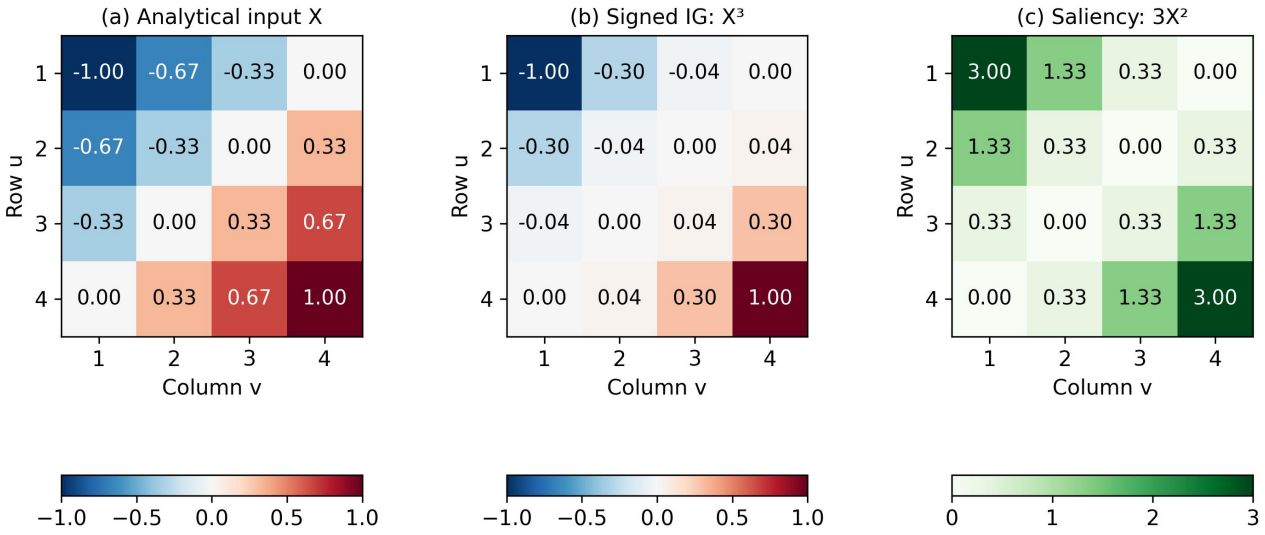


Figure 8. Analytical four-by-four input and attribution heatmaps; values are formula-derived and are not medical-image or trained-model results.

B. Disease Prediction and Patient Risk

The interpretability of a model for diagnosis and/or disease prediction is crucial for successful application in clinical practice. For instance, in a setting where there is a specified background distribution of the recorded patient features, SHAP can elucidate how a risk model relates these features with a prediction. Such visualisation can make it possible for the clinician to have a justified conversation with the patients on health and interventions to be given. LIME can also give a local surrogate explanation for an individual risk prediction under valid perturbations and measured fidelity, and under the same circumstances. These tools not only help to understand the AI predictions that are made but also boost the trustworthiness of the AI model prediction, which can be used to provide discussion on model applicability.

One of the applications involves making a medical prediction based on the longitudinal records, as done in the prediction of heart failure with an RNN [21]. Predicting an outcome is not causalising an explanation of an outcome of how the feature would change to obtain a different outcome.

C. Clinical Impact and Decision Making

While interpretability can be used to dissect what happens with AI-driven decisions, the benefit in improving diagnostic accuracy and development of trust from the clinicians, treatment options, and outcomes for the patient are other hypotheses. A controlled study should involve comparing clinician-only decisions, AI prediction displays and predictions with explanations, while capturing errors, decision times, and calibrated reliance. This is parting ways from the added value of explanations, and shows the importance of AI assistance. An appropriate clinician need for the task must be identified not inferred from a heatmap or feature ranking [6]. Before the deployment, prospective validation, subgroup assessment and governance is necessary [13].

VI. CONCLUSION AND FUTURE DIRECTIONS

A. Summary of the Review

One of the key discoveries in the field of interpretability of deep learning models in medical diagnosis is the significance of visualizing and understanding the decisions these models make.Visualizing and gaining insight into what a medical deep learning model decides is a crucial aspect. In particular, attribution techniques like saliency maps, Grad-CAM, LIME and SHAP, offer significant understanding of model predictions, which can aid in questioning the model result. Second, using these interpretable techniques in medical imaging and disease prediction necessitates the evaluation of their predictive accuracy, the fidelity of the explanation and the clinical utility separately. Moreover, a comparative study of interpretability methods indicated the importance of addressing complexity and transparency to meet different healthcare stakeholders' needs. The overall goal is to balance interpretability with external validation and clinically appropriate human-AI working together so that it does not replace evidence of patient outcomes.

B. Ethical Implications in Healthcare

There's a lot of ethics to it when it comes to healthcare and interpretability. As artificial intelligence (AI) systems become more integral to medical decision making, it is important to understand how these systems make decisions, to support patient autonomy and informed consent. The importance of transparency and avoiding misinterpretation of the model predictions to avoid adverse consequences in AI systems.

Moreover, there is a strong emphasis on ethics in relation to bias and fairness, with the potential for unclear models to exacerbate current healthcare inequalities. When building interpretability, it is important to do so to reveal bias as well as evaluate fairness independently, so as to make biased healthcare solutions inferred by interpretable methods more fair and equitable. Therefore, the application of interpretable AI in healthcare needs to be complemented by a thorough framework of ethics, ensuring patient safety and a trustworthy and controlled use of such technology.

C. Future Research

Future research in the field of explainable AI in medicine needs to further focus on increasing the interpretability of the model and preserving its performance. Perhaps hybrid approaches such as using complex deep learning architectures and simpler components with broader, understandable interpretability would be fruitful lines of inquiry. Moreover, creation of standardized evaluation measures of interpretability will assist the comparisons in methods and applications and assist the practitioner in selection of the most appropriate method. Even in the diversified medical science domain-specific interpretability requirements will facilitate solutions which will be customized to make them usable and effective. Technical, ethical, and clinical advances in EXAI will require AI researchers, clinicians and ethicists to work together closely.

D. Final Remarks

Interpretability in medical diagnosis for deep learning is of paramount importance. Understandability of these systems and their evaluation of trustworthiness is crucial for their successful integration into clinical practice, and as AI technologies evolve, it is an issue that should be highly relevant. Overall, the findings of this exploration highlight the importance of continued research and collaboration for creating and thoroughly testing interpretable AI tools that can enhance healthcare delivery. Addressing

challenges of interpretability and incorporating medical ethics demands clinical utility along with transparency, accountability and safety in the future of AI in medicine.

REFERENCES

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539
2. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005
3. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv:1708.08296. https://arxiv.org/abs/1708.08296
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778
5. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730. https://doi.org/10.1145/2783258.2788613
6. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR 106, 359-380. https://proceedings.mlr.press/v106/tonekaboni19a.html
. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. https://arxiv.org/abs/1702.08608
. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. Published December 25, 2020. https://doi.org/10.3390/e23010018
. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics, 80-89. https://doi.org/10.1109/DSAA.2018.00018
0. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618-626. https://arxiv.org/abs/1610.02391
1. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1705.07874
2. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. Computer Vision - ECCV 2014, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53
3. Food and Drug Administration (FDA), Health Canada, & Medicines and Healthcare products Regulatory Agency (MHRA). (2021). Good machine learning practice for medical device development: Guiding principles.

https://content.govdelivery.com/accounts/USFDA/bulletins/2f919ed
14. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115-118. https://doi.org/10.1038/nature21056
15. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094. https://doi.org/10.1038/srep26094
16. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25, 44-56. https://doi.org/10.1038/s41591-018-0300-7
17. Ardila, D., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25, 954-961. https://doi.org/10.1038/s41591-019-0447-x
18. Gulshan, V., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410. https://doi.org/10.1001/jama.2016.17216
19. Rajpurkar, P., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225. https://arxiv.org/abs/1711.05225
20. Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1, 18. https://doi.org/10.1038/s41746-018-0029-1
21. Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370. https://doi.org/10.1093/jamia/ocw112
2. Rieke, N., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1
3. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36-43. https://doi.org/10.1145/3233231
4. Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850
5. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923. https://arxiv.org/abs/1712.09923
6. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342
7. Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist - it's time to make it fair. Nature, 559, 324-326. https://doi.org/10.1038/d41586-018-05707-8
8. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206-215. https://doi.org/10.1038/s42256-019-0048-x
9. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314

30. Haenssle, H. A., et al. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8), 1836-1842. https://doi.org/10.1093/annonc/mdy166
31. Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular Systems Biology, 12(7), 878. https://doi.org/10.15252/msb.20156651
32. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of NAACL-HLT, 3543-3556. https://doi.org/10.18653/v1/N19-1357
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762
34. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7), e0130140. https://doi.org/10.1371/journal.pone.0130140
35. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531. https://arxiv.org/abs/1503.02531
6. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31. https://arxiv.org/abs/1810.03292
7. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 3319-3328. https://proceedings.mlr.press/v70/sundararajan17a.html
8. Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. Proceedings of the IEEE International Conference on Computer Vision, 3429-3437. https://arxiv.org/abs/1704.03296
9. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324
0. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887. https://arxiv.org/abs/1711.00399