This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Digital Object Identifier N/A

A Framework for Interpretability in Machine Learning for Medical Imaging

ALAN Q. WANG[1,2], BATUHAN K. KARAMAN[1,2], HEEJONG KIM[2], JACOB ROSENTHAL[2,3], RACHIT SALUJA[1,2], SEAN I. YOUNG[4,5], and MERT R. SABUNCU[1,2]

[1]School of Electrical and Computer Engineering, Cornell University and Cornell Tech, New York, NY 10044, USA
[2]Department of Radiology, Weill Cornell Medical School, New York, NY 10065, USA
[3]Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program, New York, NY 10065, USA
[4]Martinos Center for Biomedical Imaging, Harvard Medical School, Boston, MA 02129, USA
[5]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA

Corresponding author: Alan Q. Wang (e-mail: aw847@cornell.edu).

Funding for this project was in part provided by the NIH grants R01AG053949 and T32GM007739, and the NSF CAREER 1748377 grant.

ABSTRACT Interpretability for machine learning models in medical imaging (MLMI) is an important direction of research. However, there is a general sense of murkiness in what interpretability means. Why does the need for interpretability in MLMI arise? What goals does one actually seek to address when interpretability is needed? To answer these questions, we identify a need to formalize the goals and elements of interpretability in MLMI. By reasoning about real-world tasks and goals common in both medical image analysis and its intersection with machine learning, we identify five core elements of interpretability: localization, visual recognizability, physical attribution, model transparency, and actionability. From this, we arrive at a framework for interpretability in MLMI, which serves as a step-by-step guide to approaching interpretability in this context. Overall, this paper formalizes interpretability needs in the context of medical imaging, and our applied perspective clarifies concrete MLMI-specific goals and considerations in order to guide method design and improve real-world usage. Our goal is to provide practical and didactic information for model designers and practitioners, inspire developers of models in the medical imaging field to reason more deeply about what interpretability is achieving, and suggest future directions of interpretability research.

INDEX TERMS Interpretability, explainability, medical imaging, machine learning.

I. INTRODUCTION

Machine learning (ML) has seen remarkable advancement in recent years. ML's intersection with medical imaging (which we abbreviate as MLMI) is amongst the most promising, offering potential advances to quality of patient care [1–3]. However, the most performant machine learning models like those in computer vision and deep learning are generally regarded as black boxes – they output predictions without revealing to human users how they arrived at those predictions. As such, there has been a surge of papers calling for and proposing methods that make their decision-making interrogable, understandable, or explainable by users. This subfield has gone by names like ''interpretable," ''explainable," and ''transparent" ML. There is a clear need for such methods [4, 5], and the rising interest in this field is a reflection of the safety-critical, high-stakes setting in which medical imaging applications are deployed.

However, in the medical imaging field, there is a general sense of murkiness in how these words are or can be used. Many works claim a certain approach to be ''non-interpretable'' and thus sub-optimal, while others claim to increase interpretability, without a formalism that we can commonly agree on. Our motivation in this paper is to introduce a formal framework to use for considering, motivating, studying, validating, and discussing interpretability in MLMI. Central to our work is answering these questions:

1) Why does the need for interpretability in MLMI arise?
2) What does one actually seek when interpretability is needed? That is, what are the underlying *elements* of interpretability in MLMI?
3) Informed by the above two points, how can we formalize them into a framework for interpretability in MLMI?

This paper attempts to address these three points. In particular, within the domain of medical imaging, we pinpoint elements from a purely applications-based perspective, start-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

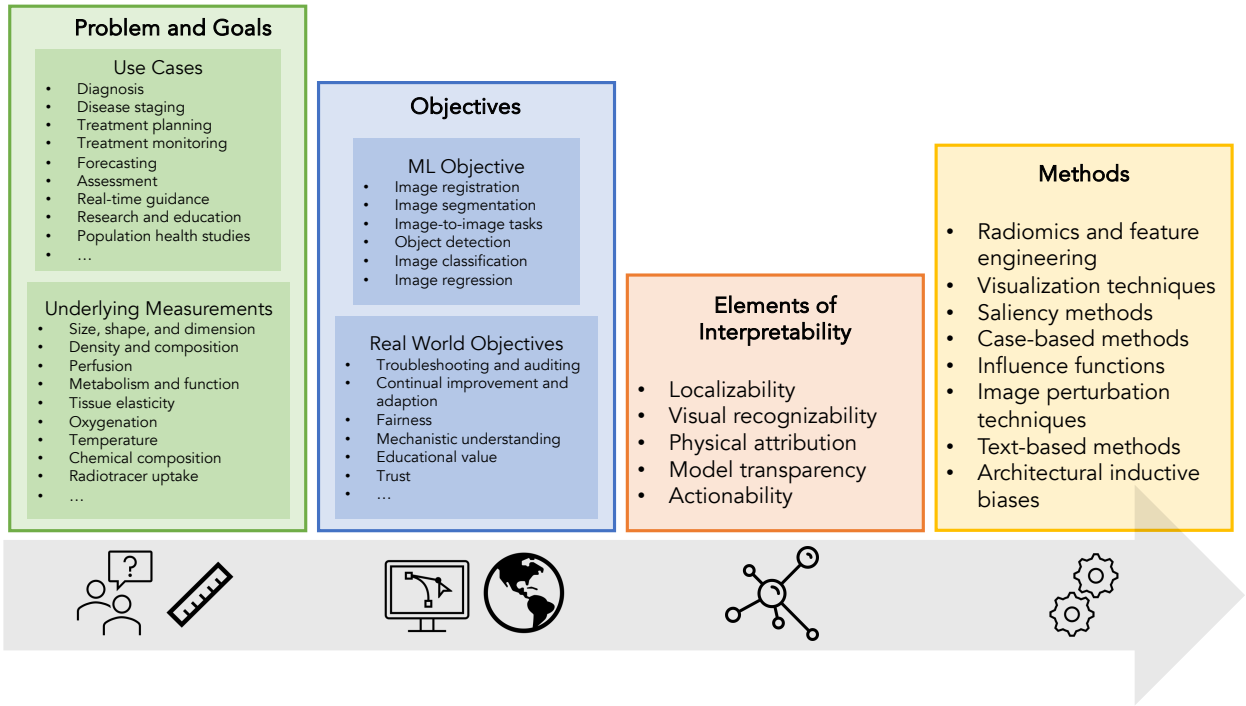


FIGURE 1. **A framework for interpretability in MLMI.**

ing from enumerating common tasks in medical image analysis and its intersection with ML, reasoning about their real world goals and considerations, and distilling from these five underlying elements: **localizability**, **visual recognizability**, **physical attribution**, **model transparency**, and **actionability**. This culminates in a framework for interpretability in MLMI, which serves as a step-by-step guide to approaching interpretability in this context.

We believe this paper can act as a guide for users and designers of interpretability in the medical imaging field by clarifying concepts in this context, suggesting future directions of research, and reasoning more deeply about how to evaluate the efficacy of their proposed methods. In addition, we hope that our work can provide concrete and practical utility for clinicians and researchers who may benefit from these models in their own workflows.

The paper is organized as follows. In Section II, we overview works similar to ours from both general machine learning and medical imaging perspectives, and identify the gap which our work aims to fill. In Section III, we broadly describe the goals and measurements which are sought in medical image analysis. In Section IV, we enumerate real-world tasks which comprise medical image analysis and which ML methods are employed to solve, as well as list possible real-world objectives which might be sought in the context of medical imaging and which cannot be explicitly optimized for. In Section V, we formally describe the elements of interpretable MLMI distilled from these real world tasks and goals. In Section VI, we connect these elements to existing literature by reframing existing methods. These sections culminate in Section VII, which provides a step-by-step guide to practitioners on approaching interpretability for MLMI. Finally, in Section VIII, we discuss further implications, including possible future directions of exploration and limitations of interpretability as a whole.

To summarize, our contributions in this work are as follows:

- We introduce a framework for interpretability in MLMI, starting from goals of medical image analysis and motivated by the need for various real-world considerations in this context, including trustworthiness, continual adaptation, and fairness.
- In doing so, we identify key elements of interpretability in MLMI and tie them to previously-proposed interpretability methods, in an effort to enable practical utility for clinicians and researchers.
- We discuss implications, limitations, and potential future exploration directions, aiming to foster deeper understanding and evaluation of interpretability methods in MLMI.

II. RELATED WORKS

Interpretability in machine learning has been discussed and argued for by many articles [6–8]. Several papers in interpretable machine learning attempt to define desiderata [9–11]. Fundamentally, these surveys identify the need for in-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

terpretability arising from a mismatch [9] or incompleteness [10] between the objectives optimized during training (i.e. test set predictive performance) and the real world costs and benefits in a deployment setting. Lipton enumerates several desiderata for interpretability, including trust, causality, informativeness, transferability, and fair and ethical decision-making [9]. Murdoch et al. view interpretability from the context of the life-cycle of models, including the relationship of the model with the practitioner and the audience [11]. From their vantage point of general machine learning, these above surveys strive to cover all aspects of interpretability in this domain. While broad, they are necessarily vague in order to encompass all facets of ML, sacrificing practical utility in the process.

Our work follows the guidelines of Rudin [6], who argues that interpretability should be defined in a domain-specific manner. However, even in the narrower context of MLMI, there is not a consensus in the literature on what precisely defines interpretability. Many review papers survey the landscape of interpretability methods available to medical imaging practitioners like clinicians and researchers [12–14]. For example, Huff et al. categorize models according to whether they seek to understand model structure and function or whether they seek to understand model output [13]. van der Valden et al. propose a categorization based on model-based vs. post-hoc, model-specific vs. model-agnostic, and global vs. local methods [14]. While comprehensively categorizing available methodologies, these surveys are not focused on obtaining or distilling from these methods any underlying elements.

Some works go beyond categorizing methodologies and propose higher-level frameworks for interpretability in this narrower context. Schuhmacher et al. propose a framework for falsifiable explanations in medical imaging contexts [15]. Their framework requires that any interpretability method connect to the physical reality of the sample from which the data originate, and also require an explanation to be a falsifiable hypothesis. Their framework can be seen as a middle ground between the inductive nature of machine learning training and the hypothetico-deductive nature of empirical research. Our framework can be seen as an elaboration upon this framework with additional elements which we believe render it more comprehensive.

Common to all aforementioned works is the fact that any core properties which are gleaned from the literature are distilled from previously proposed papers and methods. Typically, these works draw a distinction between models that are inherently interpretable by design, and models that are interpretable post-hoc, possibly by a separate trained model [6, 11, 9]. We believe such a meta-analytical approach runs counter to the applied nature of the field. Instead, in this work, we pinpoint elements by first considering real-world tasks and considerations surrounding medical image analysis and its intersection with ML. We distill elements of interpretability from this applied perspective, without being influenced by current literature. An approach which starts from available methodologies is liable to restrict itself to elements which are currently feasible, instead of those which are truly desired.

III. GOALS OF MEDICAL IMAGE ANALYSIS

Medical image analysis is a multifaceted discipline, comprising many different imaging modalities, anatomies, and end goals. To provide a solid foundation upon which to clarify goals and elements for interpretable MLMI, we start by enumerating the use cases of medical image analysis, as well as the underlying measurements which are derived from medical imaging.

A. USE CASES

Here, we enumerate several real-world use cases of medical image analysis. The ultimate goals for these use cases could be clinical or scientific in nature. On the clinical side, some example goals might include pathology detection and characterization such as in bone fracture assessment, screening high risk groups such as with lung imaging, and assessing treatment outcomes such as with molecular imaging. On the scientific side, example use cases include cell biology, neuroscience, and drug development.

1) Diagnosis

Medical imaging can help in identifying the nature of an illness [16–18], which may require revealing abnormalities such as tumors, fractures, infections, or organ damage that may not be apparent through physical examinations alone.

2) Disease Staging

The stage and severity of diseases can be facilitated by medical imaging [19]. For example, medical images in oncology can show the size and spread of tumors, enabling identification of the cancer severity and furthermore enable treatment planning.

3) Treatment Planning

Treatment planning is concerned with tailoring the application of available treatment resources to each patient's individual goals and needs [20, 21]. Common examples include planning surgical procedures and radiation therapy. With medical imaging, surgeons can visualize the precise location and extent of abnormalities, helping them make informed decisions about the best approach to treatment.

4) Treatment Monitoring

Similar to treatment planning, treatment monitoring seeks to assess the effectiveness of treatments over time [22]. Medical imaging can monitor changes in the size and characteristics of tumors or other disease markers, which can indicate how well a treatment is working.

5) Forecasting

Also known as prognostication, forecasting is concerned with predicting the likely course of a disease and its potential

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

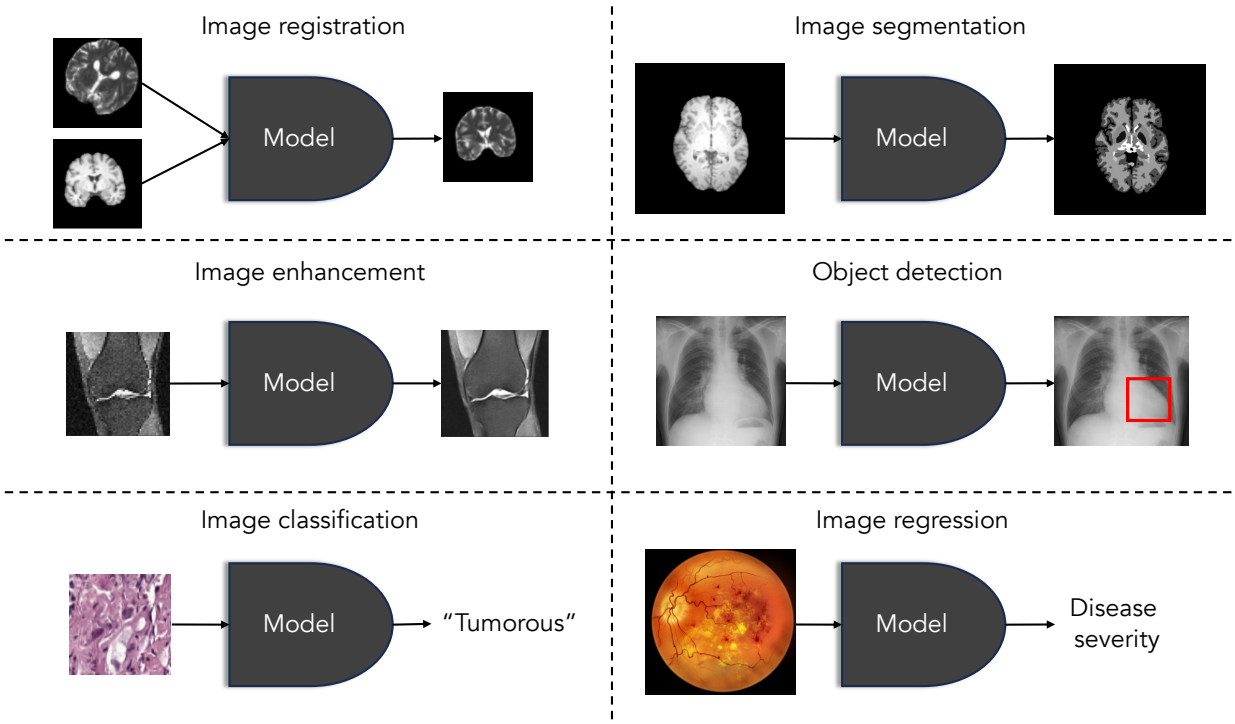


FIGURE 2. Common tasks in MLMI. Tasks are primarily characterized by the structures of their input features and output predictions. In MLMI, inputs are images or features derived from images, sometimes combined with meta-data such as patient information. The structure of output predictions are determined by the task.

outcomes [23, 24]. As a prerequisite, early detection of diseases can be facilitated through routine screening using imaging techniques like mammography or computed tomography (CT) angiography.

6) Assessment

Medical imaging is useful for assessing function of different parts of the body using imaging techniques such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) scans. This is valuable for understanding brain activity, blood flow, and metabolic processes. Medical images also offer ways to interrogate structural health, such as in x-rays and bone density scans (DXA) used to evaluate bone health, detect fractures, and assess the risk of osteoporosis.

7) Real-Time Guidance

Real-time guidance for minimally invasive procedures [25, 26], such as biopsies, catheter insertions, and needle aspirations helps to ensure precision and minimize risks.

8) Research and Education

Medical imaging data is invaluable for clinicians and scientists for purposes of research and education. For example, histopathology imaging is an important tool for researchers studying the human body [27], allowing them to investigate diseases and develop new treatments. Additionally, medical imaging is an important tool for educating medical students and other healthcare professionals.

9) Population Health Studies

Population health studies utilize aggregated medical imaging data to identify trends, risk factors, and patterns of disease within communities [28, 29].

B. UNDERLYING MEASUREMENTS

Fulfilling these use cases may require various physical measurements which are uncovered through medical imaging. In this section, we enumerate several types of underlying signals that are measured in and derived from various imaging modalities. In subsequent sections, we consider how connecting and referencing to such measurements are, amongst other factors, central to interpretable MLMI.

1) Size, Shape and Dimension

Medical imaging can precisely measure the size, shape, and dimensions of organs, tumors, lesions, and other anatomical structures [30–32]. Techniques like ultrasound, X-ray, CT, and magnetic resonance imaging (MRI) are commonly used for this purpose.

2) Density and Composition

Imaging techniques such as X-ray and CT scan provide information about tissue density and composition [33, 34]. Dif-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

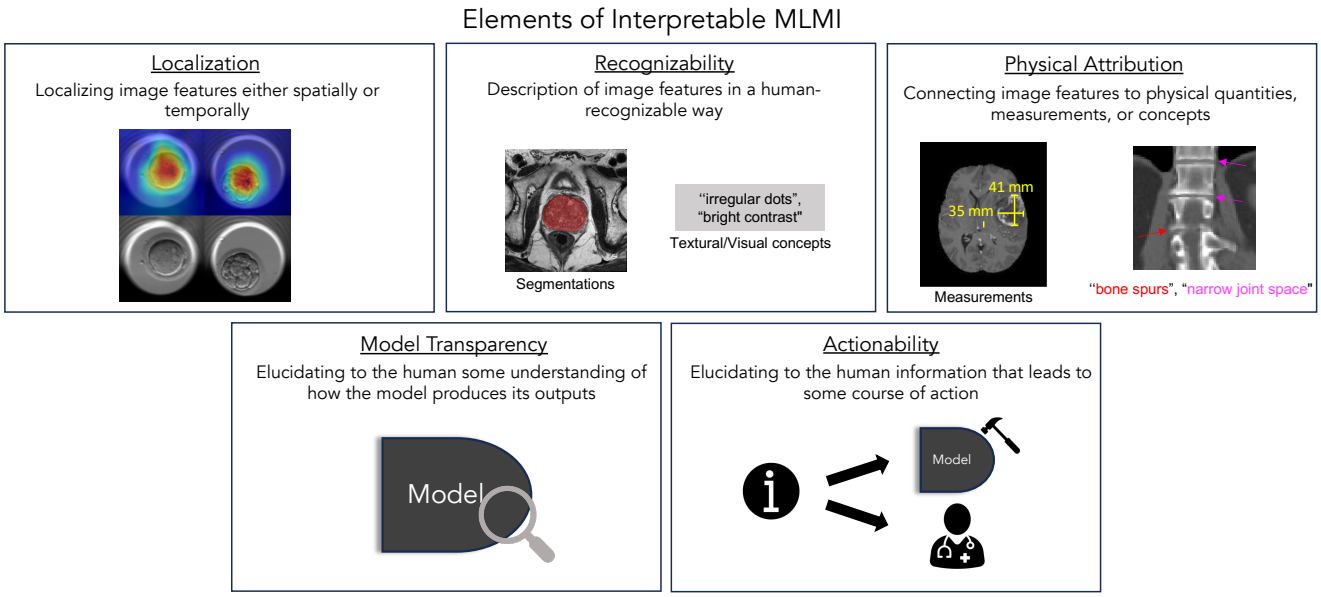


FIGURE 3. **Graphical overview of the elements of interpretable MLMI.**

ferent tissues have varying X-ray absorption characteristics, allowing the differentiation of bone, soft tissue, and air, for example.

3) Perfusion

Perfusion imaging measures blood flow to tissues and organs [35, 36]. Techniques like dynamic contrast-enhanced MRI and nuclear medicine scans (e.g. PET) can assess tissue perfusion, helping diagnose conditions like stroke or assess the viability of transplanted organs.

4) Metabolism and Function

PET and fMRI are imaging techniques that can probe metabolic activity and functional processes in the body [37, 38]. They are used to study brain function, detect cancer, and assess organ function.

5) Tissue Elasticity

Elastography is a specialized ultrasound technique that measures tissue elasticity or stiffness [39, 40]. It is valuable in assessing the stiffness of liver tissue in conditions like cirrhosis and for breast cancer diagnosis.

6) Oxygenation

Near-infrared spectroscopy (NIRS) and functional MRI can measure tissue oxygenation and oxygen levels in the blood, aiding in the assessment of brain function and detecting oxygenation problems in tissues [41, 38].

7) Temperature

Infrared thermography uses thermal imaging to measure the temperature of the skin's surface [42]. It has applications in detecting inflammation and circulatory problems [43].

8) Chemical Composition

Magnetic Resonance Spectroscopy (MRS) is a specialized MRI technique that can provide information about the chemical composition of tissues, including the presence of specific metabolites [44].

9) Radiotracer Uptake

PET scans use radiotracers labeled with radioactive isotopes to measure the uptake of specific substances in tissues, such as glucose or neurotransmitters [45, 46].

IV. MLMI OBJECTIVES

A. PRIMARY ML OBJECTIVES

As evidenced by the previous section, medical image analysis represents an array of use cases as well as imaging-derived measurements which are necessary to fulfill these use cases. MLMI is therefore comprised of many different ML tasks. In ML, tasks are primarily characterized by the structures of their input features and output predictions. In MLMI, inputs are images or features derived from images, sometimes combined with meta-data such as patient information. The structure of output predictions are determined by the task. In this section, we detail some of the most common tasks, and visualize them in Fig. 2.

- *Image registration* [47] seeks to align scans of the same anatomy. The output is typically a transformed image aligned to another fixed image.
- *Image segmentation* [48] seeks to delineate different anatomical structures into discrete categories. The output is pixel-level class predictions.
- *Image enhancement, reconstruction, or restoration* predict a clean version of a noisy image [49]. The output is pixel-level scalar predictions.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

- *Object detection* [50, 51] seeks to detect or isolate specific objects in the image, like organs or lesions. The outputs can be image coordinates (e.g. landmarks) or bounding boxes.
- *Image (or patient)-level classification* seeks to categorize an image at the patient-level into one of several distinct categories (e.g. diagnosis) [52, 53]. The output is a single class label.
- *Image (or patient)-level regression* seeks to associate with an image a real number that may represent some quantity (e.g. disease severity or some physical measurement) [54]. The output is a scalar real number.

The task and its corresponding structure of output prediction typically imply a prescribed objective which is optimized (e.g. mean-squared-error or cross-entropy). This objective is mathematically defined and thus can be optimized explicitly. However, in the real world, we often have considerations that go beyond these objectives. This is especially true in the context of medical imaging, where image-based tasks form one component of a larger clinical workflow, and where real world considerations such as safety, trust, fairness, and ease-of-use must be balanced alongside mathematical loss functions [55]. In the next section, we turn to these real world goals.

B. OTHER REAL-WORLD OBJECTIVES

In the previous section, we enumerated common tasks in medical image analysis and characterized them by the structure of their predictions; this structure lends itself to natural objectives that quantify performance and can be explicitly optimized (such as Dice overlap score for image segmentation or accuracy for classification). However, often times in real world deployment, we have goals beyond this natural objective. In this section, we enumerate several such (non-exhaustive) goals.

1) Facilitating Troubleshooting and/or Auditing.

Clinicians and researchers may require that models enable some way of facilitating troubleshooting and/or auditing. This is especially necessary when models are performing poorly, or when the robustness of models are called into question. As an example, Zech et al. used activation maps to uncover models leveraging spurious features (e.g. ‘‘shortcuts") in chest radiographs [56]. In particular, activation maps showed that image artifacts (such as indicators of R and L sides) that were heavily correlated with the outcome of interest were driving the prediction. Such a finding might lead to removing such artifacts or other changes to remove model dependence on this shortcut.

2) Enabling Continual Improvement and Adaptation.

Related but distinct from troubleshooting and auditing, clinicians and researchers may require that models continually improve and adapt over time. Data is a reflection of the real world and is therefore dynamic and constantly evolving. This can stem from availability of more data over time, but also may stem from shifts in the underlying statistics of the images due to changes in prevalence of diseases, changes in imaging protocols, or upgrades in measurement technologies [57]. A model should enable users to identify and correct errors due to these changes.

3) Ensuring Fairness.

Fairness seeks to guarantee that a model’s performance is balanced or invariant across patient groups, and in particular that it does not disadvantage minority or otherwise protected groups [58]. A clinician or researcher should be able to ensure and validate fairness of a model by, for example, interrogating and comparing the model’s performance across groups. As an example, Puyol-Anton et al. identified statistically significant differences in Dice scores between white population and minority ethnic groups for state-of-the-art models that segmented ventricles and the myocardium in cine cardiac magnetic resonance images [59]. As another example in brain MRI segmentation, Ioannou et al. observed a statistically significant drop in Dice scores when models trained on data from white subjects were evaluated on black female subjects [60]. Finally, Abbasi-Sureshjani et al. showed that skin lesion classifiers exhibited large differences in performance between male and female patients even though their training data was relatively balanced [61].

4) Uncovering Scientific Insights and Advancing Mechanistic Understanding.

An exciting value proposition is the possibility of uncovering or revealing novel scientific insights. Science and medicine are often concerned with discovering and characterizing causal and/or mechanistic relationships. These relationships can be characterized on physical, physiological, or biochemical grounds via known mechanisms, phenomena, or pathways [62].

As an example, a model trained to predict survival from H&E histological sections for colorectal cancer was interrogated using a combination of model interpretability approaches and manual review by expert human pathologists, and a prognostic histological feature, tumor adipose feature (TAF), was identified [63]. Human pathologists were then able to learn to independently identify the TAF feature as an effective prognostic marker [64]. Models could play an important role in multiple parts of the scientific process, from discovery of new image-based patterns to improving human performance on critical tasks.

5) Providing Educational Value.

Clinical trainees often need to learn to analyze several kinds of images to extract key clinical imaging features to identify patterns and produce differential diagnosis corresponding to those patterns. However, their training and education may be hampered due to time constraints, limited training opportunities, or a lack of access to educational images. Machine

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

learning based teaching platforms can augment their educational experience and provide key insights to improve them as clinicians. [65] presents an AI-augmented medical education platform to detect hip fractures by showcasing a heatmap to indicate locations of abnormality. They demonstrated that students who were trained by their system showed an improvement in understanding of the key features of a hip fracture in radiographs. Adaptive Radiology Interpretation and Education System (ARIES) is another platform presented by [66] which uses ML models to extract clinical imaging features from various brain MRI sequences and present it to neuroradiology residents to perform differential diagnosis. They observed that the residents performed significantly better with ARIES than without.

6) Engendering Trust.

Perhaps the least tangible but equally important auxiliary objective is engendering trust. The emphasis on trust is most needed in clinical settings, where one of the major barriers to integrating ML models into clinical workflows is a lack of trust among clinicians; from this perspective, interpretability plays a role in MLMI as a means by which to improve trust among humans who are using these models [4, 12, 67, 5]. The concept of trust can further be decomposed into two components: trusting individual predictions produced by the model, and trusting the model itself as an entity for producing predictions [68]. Notably, this implies that explainability is principally valuable as an instrument to achieve other end goals, such as building trust or increasing usability of models, but does not necessarily have intrinsic value in and of itself [69].

V. ELEMENTS OF INTERPRETABLE MLMI

In the previous section, we enumerated various ML tasks and several real world goals for MLMI models, which in conjunction with the explicit objective, can be seen as constituting the combined end goals of the model. From the preceding section, it is apparent that real world models require insights that go beyond a single, aggregated, and quantifiable performance metric which encodes the explicit objective. The need for such insights brings about the need for interpretability [9, 10].

Yet, it is clear that not all these goals are sought after for every task, and furthermore that each one of these goals demands a different "element" of interpretability. That is, there is not one unique element of interpretability that satisfies all real world goals in all scenarios. Thus, we seek to answer the question: what are the core underlying elements of interpretability in MLMI? In particular, we seek to distill these objectives to be as minimal as possible.

We pinpoint five (possibly non-exhaustive) underlying elements: **localizability**, **visual recognizability**, **physical attribution**, **model transparency**, and **actionability**. The first three are concerned with characterizing the features of the image. The fourth element, model transparency, is concerned with the model itself. The fifth element, actionability, is concerned with both the model and influence on the human user. These elements are further detailed below and depicted in Fig. 3.

A. LOCALIZABILITY

Where are the features, either spatially or temporally, that are driving the prediction? Medical images are high-dimensional spatio-temporal data that simultaneously exhibit relevant and redundant information, at various scales and locations. **Localizability** of features which are driving the prediction and conveying this location to the human is a central element of interpretability in MLMI.

B. VISUAL RECOGNIZABILITY

What are the visual features that drive the prediction? We refer to a model which exhibits **visual recognizability** as one that can provide a description of these image features in a human-recognizable way. This might involve conveying to the user features related to brightness, certain pixel intensity patterns, or certain textural patterns.

C. PHYSICAL ATTRIBUTION

This element seeks to reveal to the human salient features of the image which are attributed to some underlying **physical meaning**. This element connects image features to real-world entities of or related to the image formation process. These entities could be quantitative (measurements like those enumerated in Section III-B) or conceptual (e.g. semantic or human-derived concepts).

D. MODEL TRANSPARENCY

Model transparency elucidates to the human some degree of understanding of *how the model produces its outputs*. In contrast to the previous three properties, transparency is concerned with inputs, architectures, and/or algorithms and does not characterize image features.

Prior works in the ML literature have discussed different notions of transparency, including simulatability, decomposability, algorithmic transparency, sparsity, modularity, and intelligibility [9, 11, 70]. These notions are applicable broadly to any ML model and are not special to medical image analysis.

By contrast, a more MLMI-specific notion of transparency are approaches which incorporate information related to known entities or mechanisms with respect to the problem being solved or other domain-specific information. Inductive biases can then be injected which constrain the model to operate within the bounds of that domain. For example, the model designer may have domain knowledge about the data-generating and/or image-formation process [71–73], or have prior knowledge of the mechanisms or anatomy that underlie the problem being studied [74–76]. See Section VI-H for further details.

E. ACTIONABILITY

Actionability elucidates to the human information that *leads to some course of action*. The action may be some form

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

of falsifiability or recourse [15]. We identify two types of actionability.

The first type refers to recourse with respect to the model. Here, actionability refers to interpretable information that makes apparent a solution which involves changing or editing some aspect of the model or algorithm. This is related to model transparency, in that models which are transparent can naturally lead to actionable recourse.

The second type refers to recourse with respect to the human user. Here, actionability refers to interpretable information which leads to actionable insights for the human user. For example, making apparent certain pathologies in the image might correspond to prescribed protocols that the human should undertake. As another example, an interpretation may indicate to the human to run a follow-up experiment to further interrogate the problem in question.

VI. CONNECTING EXISTING METHODS TO ELEMENTS

In this section, we discuss several popular interpretability methods and describe how they relate to our aforementioned elements. For more comprehensive overviews of methodologies, we refer the reader to dedicated surveys [12–14].

A. RADIOMICS AND FEATURE ENGINEERING

Radiomics and other clinically-derived features seek to reduce high dimensional features like shape, intensity and texture into features using image processing or clinically-informed techniques [77–81], and is particularly prevalent in oncological applications. Typical radiomics workflows require a segmentation step to first localize the region of interest (e.g. tumor), and thus may be seen to exhibit **localizability**. Furthermore, since many engineered features are designed to capture intensity or textural information, they may be seen to exhibit some notion of **recognizability**. Features focused on domain-specific and clinically-relevant features like shape, size, or orientation may be seen as exhibiting **physical attribution**. Finally, because many features involve relatively simple mathematical expressions to compute, humans can typically contemplate the computation of these features and can thus be seen to exhibit **model transparency**.

B. VISUALIZATION TECHNIQUES

In this section, we detail several visualization techniques that seek to convey to humans the inner workings of a model via information that can be plotted. Since these methods are all seeking to overcome the black-box nature of models, they can all be seen as exhibiting **model transparency**.

Feature importance techniques such as partial dependence plots (PDPs) [82], individual conditional expectation (ICE) plots [83], and SHapley Additive exPlanations (SHAP) [84–86] have been used for analysing partial dependency of breast tumor malignancy on ultrasound image features [87], and uncovering important features in predicting cognitive decline in Alzheimer's disease [88].

A class of techniques allows for analysis of high-dimensional latent space of deep networks. Techniques like t-SNE [89] and UMAP [90] project high-dimensional embeddings into a low-dimensional space that can be plotted (e.g. in 2 dimensions). Doing so can roughly uncover how models organize datapoints in the latent space.

C. SALIENCY MAPS

Saliency maps reveal salient features within an image such as lesions or changes over time in longitudinal studies by outputting heatmaps which can be overlaid on top of the image(s) [91–93]. Popular techniques include CAM-based methods [94–100] and methods that incorporate attention mechanisms to make predictions [101, 102, 92, 103].

Saliency maps may be seen to exhibit **localizability**, since they highlight salient features locally.

D. CASE-BASED METHODS AND EXPLANATION BY EXAMPLE

This class of methods present evidence for predictions for a query sample by identifying similar samples. This is sometimes referred to as "explanation by example" [9, 104]. Prototypical networks [105–108] classify features or images (e.g. chest X-ray scans) by finding their nearest class prototypes (e.g. images of typical cardiomegaly, nodule, etc) in a learned embedding space. Lamy et al. propose a visual case-based reasoning method for classification and applied it to breast cancer [109].

Case-based methods may be seen to exhibit **model transparency**, since they can reveal to the user which images the model deems most (dis)similar to one another, which in turn reveals some understanding of its inner workings. In some cases, case-based methods may be paired with other techniques which exhibit other elements, like saliency maps for **localizability** [110].

E. INFLUENCE FUNCTIONS

Influence functions quantify the most influential training images (either positive or negative) which lead to the prediction for a given sample [111, 112, 108]. For example, Wang et al. used influence functions in the context of liver tumor diagnosis in order to quantify the radiological features that were most influential for classifying a particular lesion [113].

Influence functions may be seen to exhibit **model transparency**, since they reveal the most "helpful" or "harmful" training images for a model to make a prediction for a given image.

F. IMAGE PERTURBATION TECHNIQUES

Image perturbation techniques directly perturb the input image. These perturbations can identify changes in intensity patterns and deformations that highlight underlying biophysical and biomechanical processes. For example, counterfactuals [114–116] introduce the minimum amount of change to a given image to change its predicted class. Examples of counterfactual image generation for medical image analysis include chest [93, 117], and knee [118] X-ray scans for the prediction of pulmonary disorders, and severity of

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

osteoarthritis, respectively, as well as brain [119], and breast [116] MRI for Alzheimer's disease and breast tumor analyses, respectively.

Other popular perturbation techniques include adversarial attacks [120–124], occlusion sensitivity [125], and local interpretable model-agnostic explanations (LIME) [68].

Perturbation techniques may be seen to exhibit **visual recognizability**, since they may alter features in meaningful ways that humans can recognize. If these perturbations are meaningfully localized, they may also be seen to exhibit **localizability**.

G. TEXT-BASED METHODS

Many methods approach interpretability by incorporating human-derived and annotated concepts. In MLMI, these concepts could be clinically-relevant text or phrases, like ''sclerosis", ''bone spurs", or ''narrow joint space" in the context of arthritis grading. Concept bottleneck models (CBMs) [126–128] aim to decompose end-to-end prediction models into non-linear concept prediction and linear or logistic task prediction from the predicted concept values. Concept whitening (CW) models [129] extend CBMs by additionally othogonalizing and rescaling the concept activation vectors during training for improved interpretability. Applications of CBMs to medical image analysis include skin lesion diagnosis [130, 131], pediatric appendicitis [132], and chest radiology [133]. Quantitative testing with concept activation vectors (TCAV) [134] employs concept activation vectors to interpret the internal state of neural networks in terms of user-defined concepts; the authors apply it to a deep learning model that predicts diabetic retinopathy (DR) level from retinal fundus images.

Other than concepts, text-based methods also may output justifications, reports, and captions. These methods typically provide long-form text along with the predictions. Zhang et al. propose MDNet, which leverages a vision and language model to visualize attention as well as provide justifications of the predicted diagnosis [91]. Lee et al. propose a visual word constraint model to improve the accuracy of a justification generator for diagnostic interpretations [135].

Concept-based methods may be seen to exhibit **physical attribution**, since concepts are typically human-annotated phrases that map to specific real-world characteristics associated with a diagnosis. Additionally, text might point out textural or image-derived information (e.g. ''irregular dots" or ''bright contrast") which may be seen to exhibit **recognizability**.

H. ARCHITECTURAL INDUCTIVE BIASES

In medical imaging contexts, many techniques and methods exist which leverage domain knowledge to inject domain-specific inductive biases into model architectures. These methods use information specific to the problem being solved, for example specific to image content (e.g. anatomy, geometry), or leveraging information about the data-generating process (e.g. a known forward model). Since they directly alter architectures and algorithms used to solve the problem, models of this type may be seen to exhibit **model transparency**.

In reconstruction, unrolled architectures [71] and physics-based models [72] incorporate knowledge of the underlying forward model, data-generating process, or physical constraints. In classification, mechanistic models [136] and anatomical priors [74] incorporate knowledge of specific objects of interest into the architecture, for example brain midline shift [76], models for tumor growth [75], and cardiac image enhancement and segmentation [137]. In registration, the relative homogeneity in geometry of medical images has been leveraged [138]. Another line of research seeks to apply causal inference to various contexts in MLMI [139, 73, 140]. Works in this vein inject prior information into the model by reasoning about the causal relationships in the data-generating process.

VII. A FRAMEWORK FOR INTERPRETABLE MLMI

In this section, we provide a guide to practitioners of MLMI models for reasoning about interpretability. This guide follows the ordering of sections in the paper, and we reference the corresponding section throughout. A diagram is depicted in Fig. 1.

1) *Problem and goals.* Consider the medical use case being addressed (Section III-A) and any underlying measurements necessary for that use case (Section III-B).
2) *ML objectives.* From the use case, determine the appropriate ML task and its appropriate objective function (Section IV-A).
3) *Real-world objectives.* In addition to the objective function, consider other real-world objectives pertinent to the use case being addressed (Section IV-B).
4) *Elements of interpretability.* From the other real-world objectives, consider the underlying elements which capture these objectives (Section V).
5) *Methods.* Given these elements, select the appropriate interpretability method which addresses it (Section VI).

In the remainder of the section, we provide two case studies to illustrate our proposed framework.

A. CASE STUDY 1: AUTOMATED, INITIAL SCREENING FOR DIABETIC RETINOPATHY

1) *Problem and goals.* Diabetic retinopathy (DR) is a complication of diabetes that can ultimately lead to blindness. Early-stage detection of DR is essential for taking measures to reverse the disease's progression, making it important for people living with diabetes to be screened regularly, with retinal photography to assess for blood vessel damage or retinal lesions being a key imaging modality used for screening [141–143]. However, screening programs often face difficulties due to a shortage of clinical specialists [142]. Therefore, a machine learning solution as an initial triage tool which flags suspicious or high-risk cases and sends them to a clinician for further assessment would reduce clinician

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

burden and enable more widespread screening, which is especially useful in resource-limited environments.

2) *ML objective.* Since this task involves categorization into one of several distinct labels, this is a classification task and the loss function can be a classification loss like cross-entropy.
3) *Real-world objectives.* Beyond classification, as this model will be deployed in a clinical setting, we would demand the ability to troubleshoot and audit the model, and ensure that predictions are fair across patient groups. Additionally, user trust of the model is imperative.
4) *Elements.* Of utmost importance amongst the elements is **localizability** of the lesions or blood vessels which are suspicious. Probing a model's ability to localize and recognize the suspicious parts of the image will satisfy the various real-world objectives. As this tool will be used as an initial screening tool and the nature of the disease being studied is well understood, the value of **physical attribution** and **actionability** is relatively small.
5) *Methods.* Saliency maps may be used to address localizability.

B. CASE STUDY 2: BREAST CANCER DETECTION

1) *Problems and goals.* Early detection of breast cancer is important for increasing the chances of patients' survival [144]. A large part of clinical diagnosis in this area is concerned with characterizing various features of the tumor from mammograms, including masses, calcification, architectural distortion, asymmetries, and other related signs [145]. These and other factors have been standardized in the BI-RADS format [146]. Thus, there exists a list of human-recognizable and physically-understood predictive features.
2) *ML objective.* As the goal of this tool is to make a discrete diagnostic decision, the ML task is classification and the loss function can be a classification loss like cross-entropy.
3) *Real-world objectives.* Similar to Case Study 1, as this model will be deployed in a clinical setting, we would demand the ability to troubleshoot and audit the model, and ensure that predictions are fair across patient groups. Additionally, user trust of the model is imperative.
4) *Elements.* Similar to Case Study 1, **localizability** of the tumors is an important element to consider. As this tool will be used as a diagnostic aid by clinicians, it is important to connect to known features that are widely used for breast cancer diagnosis, e.g. as codified by the BI-RADS format. Thus, **visual recognizability** and **physical attribution** of features are desired.
5) *Methods.* Similar to Case Study 1, saliency maps may be used to address localizability. Visual recognizability and physical attribution may be addressed by incorporating feature engineering and radiomics [147]. The importance of these features can be further interrogated using visualization techniques like ICE and SHAP.

C. CASE STUDY 3: VISUAL ENCODING

1) *Problem and goals.* Visual encoding models that predict functional MRI brain responses to visual stimuli are of scientific interest, as they enable quantification of population heterogeneity in brain stimuli-response mapping, which may allow insight into variability in bottom-up neural systems that can in turn be related to individual's behavior or pathological state [148–150].
2) *ML objective.* Since this task requires predicting real-valued time-series signals, the task is regression and an appropriate ML loss function might be mean-squared error.
3) *Real-world objectives.* The ultimate objective in this case is scientific discovery. As this model will be employed for uncovering aspects of brain function and how it is linked to individual behavior, we are primarily interested in mechanistic understandings.
4) *Elements.* Important elements include **model transparency** and **actionability**, as we are interested in the information and potential scientific understanding which can be uncovered by the model, as well as actionable insights for future experiments.
5) *Methods.* For model transparency, specific architectures can be designed to mimic the brain, like enforcing intermediate layers of the deep network to map to different areas of the visual cortex. These layers may be visualized, for example using t-SNE plots, to understand how the network organizes information in the latent space. This can also address actionability, as being able to diagnose issues in some intermediate layers may uncover issues with certain parts of the network.

VIII. DISCUSSION

A. INHERENT INTERPRETABILITY OF PIXEL-LEVEL TASKS

Pixel-level tasks output predictions directly on the image grid, where the level of prediction is localized to the smallest level of granularity. Often, these outputs represent human-recognizable entities, like region-of-interest labels for segmentation or dense deformation fields for registration. Thus, it may be seen that such pixel-level tasks exhibit an inherent interpretability because the structures of their predictions happen to be localizable and visually recognizable. This may explain why there is a relatively fewer number of works focused on interpretability in this context (to the best of our knowledge, there are no survey papers discussing the subject).

B. INTENDED USERS

Interpretability is a subjective enterprise, and often times a useful interpretation is in the eye of the beholder. Thus, keeping in mind intended users is important when considering interpretability methods. Methods most useful to clinicians and researchers will differ greatly from methods most useful

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

to model developers and auditors, for example. Even amongst clinicians, their expertise might vary and will affect what one is looking for in terms of an interpretation. For example, experienced radiologists might be satisfied with one interpretation, whereas a treating physician might not.

In this work, we have focused on the case where interpretability is sought for deployed systems and where the primary users are clinicians or researchers who may not be experts in ML. Generally, we argue that localizability, visual recognizability, and physical attribution are the most important elements for these users, as it connects the prediction directly to the images and abstracts away modeling-specific choices. By contrast, transparency might be more important for model designers who have knowledge of the ML specifics.

C. FAITHFULNESS

One area of interpretability research not yet mentioned is that of faithfulness of the interpretability method [151–153]. A faithful interpretable method is one that accurately reflects the underlying predictive model, such that the method is internally coherent [152]. Since faithfulness pertains only to the relationship between the explanation and the model, it is relatively removed from connections to the real-world and is thus out-of-scope with respect to this work. However, practitioners are advised to consider this aspect of interpretability in their own use cases.

D. LIMITATIONS OF INTERPRETABLE MLMI

While interpretability is a promising direction for the end goal of improving patient care, it can potentially introduce biases or hinder model performance. Here, we list two potential sources of limitation to consider.

First, interpretable MLMI might induce or contribute to human biases during decision-making processes. When users' levels of trust in algorithmic decision-making aids are miscalibrated, the result manifests as either automation bias (too much trust) or algorithmic aversion (too much distrust) [154–156]. False senses of trust can also be induced when models' high-dimensional internal representations are projected into low-dimensional space for human review (e.g., 2-dimensional heatmaps), leading users to the false impression of true understanding while they in fact only have an approximation of ''ersatz'' understanding of the model's inner workings [157]. Because one of the core objectives of interpretable MLMI is to modulate users' trust in ML-powered decision aids, these biases represent pitfalls which must be considered carefully when deploying these models in practice.

Second and relatedly, interpretability in MLMI might run counter to the goal of ultimately augmenting the human user's capabilities. From this perspective, ''uninterpretability'' might be reflective of a means of performing decision-making orthogonal to the capabilities of humans. As medicine becomes increasingly complex, the amount of information available to be integrated into the decision-making process may exceed the capacities of the human mind [158]. Indeed, constraining ML models to be human-understandable may result in models which overlap with or simply recapitulate patterns already identified by humans. Leveraging the complementary strengths of human perception and machine perception, including incorporating and processing signals which cannot be easily rationalized by humans, may ultimately be necessarily towards the goal of improving decision-making.

E. NOVEL APPROACHES TO INTERPRETABILITY

While there have been many attempts at interpretability in MLMI, there is still a need for further research and new perspectives. For example, a nascent approach to interpretability that seeks to reason about the mechanisms by which neural networks arrive at predictions is called mechanistic interpretability [159]. Works in this line of research interpret models trained on simple tasks like modular arithmetic and decompose it into individual neurons and circuits which compute different parts necessary to arrive at the answer. This field has found promise in understanding the inner workings of transformer-based large language models [160]. While underexplored in the context of medical imaging, mechanistic interpretability is directly aligned with the real-world goal of connecting to and revealing mechanisms and mechanistic understanding. We believe works along this vein will be important in the future.

While this work is focused on interpretable ML in medical imaging, we do recognize that similar ideas may be applicable to other domains like law and finance, and hope that our work can inspire others to reason similarly in a domain-specific manner about interpretability.

IX. CONCLUSION

We arrived at five elements of interpretability in machine learning for medical imaging through reasoning about real-world goals, tasks, and objectives in medical image analysis and its intersection with machine learning. These elements are localizability, visual recognizability, physical attribution, model transparency, and actionability. Furthermore, we connect our elements to the current literature by summarizing various methods and approaches in interpretable MLMI and attributing each to our elements. Taken together, these components constitute a framework for interpretability in MLMI, a guide for practitioners on how to reason and utilize interpretability in their own use cases.

X. ACKNOWLEDGEMENTS

We thank Kayhan Batmanghelich for insightful discussions and suggesting the actionability element.

REFERENCES

[1] Karin Dembrower, Alessio Crippa, Eugenia Colón, Martin Eklund, and Fredrik Strand. Artificial intelligence for breast cancer detection in screening mammography in sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*, 2023/09/11 2023.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

[2] Bradley J. Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, Mar-Apr 2017.

[3] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017. PMID: 28301734.

[4] Pranav Rajpurkar and Matthew P. Lungren. The current and future state of ai interpretation of medical images. *New England Journal of Medicine*, 388(21): 1981–1990, 2023. PMID: 37224199.

[5] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020. PMID: 32510054.

[6] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

[7] Zachary C. Lipton. The doctor just won't accept that!, 2017.

[8] Shinjini Kundu. Ai in medicine must be explainable. *Nature Medicine*, 27(8):1328–1328, Aug 2021.

[9] Zachary C. Lipton. The mythos of model interpretability, 2017.

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[11] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116 (44):22071–22080, oct 2019.

[12] Zohaib Salahuddin, Henry C. Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, 140:105111, 2022.

[13] Daniel T Huff, Amy J Weisman, and Robert Jeraj. Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in Medicine and Biology*, 66(4):04TR01, feb 2021.

[14] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

[15] David Schuhmacher, Stephanie Schörner, Claus Küpper, Frederik Großerueschkamp, Carlo Sternemann, Celine Lugnier, Anna-Lena Kraeft, Hendrik Jütte, Andrea Tannapfel, Anke Reinacher-Schick, Klaus Gerwert, and Axel Mosig. A framework for falsifiable explanations of machine learning models with an application in computational pathology. *Medical Image Analysis*, 82:102594, 2022.

[16] Juan Manuel Fernández Montenegro, Barbara Villarini, Anastasia Angelopoulou, Emmanouil Kapetanios, Juan Garcia-Rodriguez, and Vasilis Argyriou. A survey of alzheimer's disease early diagnosis methods for cognitive assessment. *Sensors*, 20(24):7292, December 2020.

[17] Reem Nooreldeen and Horacio Bach. Current and future development in lung cancer diagnosis. *International Journal of Molecular Sciences*, 22(16):8661, August 2021.

[18] Rossana Bussani, Matteo Castrichini, Luca Restivo, Enrico Fabris, Aldostefano Porcari, Federico Ferro, Alberto Pivetta, Renata Korcova, Chiara Cappelletto, Paolo Manca, Valentina Nuzzi, Riccardo Bessi, Laura Pagura, Luca Massa, and Giuseppe Sinagra. Cardiac tumors: Diagnosis, prognosis, and treatment. *Current Cardiology Reports*, 22(12):169, October 2020.

[19] James E Conklin, Joseph V Lieberman, Charles A Barnes, and David Z Louis. Disease staging: implications for hospital reimbursement and management. *Health Care Financ Rev*, Suppl:13–22, 1984.

[20] Sharon L. Johnson. Chapter 3 - treatment planning. In Sharon L. Johnson, editor, *Therapist's Guide to Pediatric Affect and Behavior Regulation*, Practical Resources for the Mental Health Professional, pages 131–236. Academic Press, San Diego, 2013.

[21] Nora E. Noel. Chapter twenty-one - evidence-based treatment planning for substance abusetherapy. In Peter M. Miller, editor, *Evidence-Based Addiction Treatment*, pages 399–417. Academic Press, San Diego, 2009. ISBN 978-0-12-374348-0.

[22] Jeffrey K Aronson. Monitoring therapy. *British Journal of Clinical Pharmacology*, 60(3):229–230, 2005.

[23] Christina Chu, A Nicola White, and Patrick Stone. Prognostication in palliative care. *Clinical Medicine*, 19(4):306–310, Jul 2019.

[24] David Hui, Carlos Eduardo Paiva, Egidio G Del Fabbro, and et al. Prognostication in advanced cancer: update and directions for future research. *Supportive Care in Cancer*, 27:1973–1984, 2019.

[25] Maria Antico, Fumio Sasazawa, Liao Wu, Anjali Jaiprakash, Jonathan Roberts, Ross Crawford, Ajay K. Pandey, and Davide Fontanarosa. Ultrasound guidance in minimally invasive robotic procedures. *Medical Image Analysis*, 54:149–167, 2019.

[26] Justin D Stull, John J Mangan, Alexander R Vaccaro, and Gregory D Schroeder. Robotic guidance in minimally invasive spine surgery: a review of recent literature and commentary on a developing technology. *Current Reviews in Musculoskeletal Medicine*, 12(2): 245–251, 2019.

[27] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67: 101813, 2021.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

[28] Marc N. Gourevitch, Lesley H. Curtis, Maureen S. Durkin, Angela Fagerlin, Annetine C. Gelijns, Richard Platt, Belinda M. Reininger, Judith Wylie-Rosett, Katherine Jones, and William M. Tierney. The Emergence of Population Health in US Academic Medicine: A Qualitative Assessment. *JAMA Network Open*, 2(4): e192200–e192200, 04 2019.

[29] Mohammad Shahzad, Ross Upshur, Peter Donnelly, and et al. A population-based approach to integrated healthcare delivery: a scoping review of clinical care and public health collaboration. *BMC Public Health*, 19:708, 2019.

[30] Yanxi Liu, Pingkun Yan, Xianbin Cao, Yonggang Shi, Shuo Li, Qian Wang, Liang Wang, and Xiaowei Zhuang, editors. *Shape in Medical Imaging*. Springer, 2020.

[31] Yashbir Singh, Colleen M. Farrelly, Quincy A. Hathaway, Tim Leiner, Jaidip Jagtap, Gunnar E. Carlsson, and Bradley J. Erickson. Topological data analysis in medical imaging: current state of the art. *Insights into Imaging*, 14(1):58, 2023.

[32] Sinh Van Nguyen, Ha Manh Tran, and Truong Son Le. Application of geometric modeling in visualizing the medical image dataset. *SN Computer Science*, 1(8): 254, 2020.

[33] Oskui S Mesbah, HC Bhakta, G Diamante, H Liu, D Schlenk, and Grover WH. Measuring the mass, volume, and density of microgram-sized objects in fluid. *PLoS One*, 12(4):e0174068, Apr 2017.

[34] M.J. Yaffe. Mammographic density. measurement of mammographic density. *Breast Cancer Res*, 10(3): 209, 2008.

[35] Jelle Demeestere, Anke Wouters, Soren Christensen, Robin Lemmens, and Maarten G. Lansberg. Review of perfusion imaging in acute ischemic stroke. *Stroke*, 51(3):1017–1024, 2020.

[36] Marije E Kamphuis, Marcel J W Greuter, Riemer H J A Slart, and Cornelis H Slump. Quantitative imaging: systematic review of perfusion/flow phantoms. *Eur Radiol Exp*, 4(1):15, Mar 2020.

[37] A Judge and MS Dodd. Metabolism. *Essays Biochem*, 64(4):607–647, Oct 2020.

[38] Gary H Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America*, 22(2):133–139, vii, 2011.

[39] Rosa M.S. Sigrist, Joy Liau, Ahmed E. Kaffas, Maria Cristina Chammas, and Juergen K. Willmann. Ultrasound elastography: Review of techniques and clinical applications. *Theranostics*, 7(5):1303–1329, Mar 2017.

[40] A Sarvazyan, TJ Hall, MW Urban, M Fatemi, SR Aglyamov, and BS Garra. An overview of elastography - an emerging branch of medical imaging. *Curr Med Imaging Rev*, 7(4):255–282, Nov 2011.

[41] DR Dantzker. Adequacy of tissue oxygenation. *Crit Care Med*, 21(2 Suppl):S40–S43, Feb 1993.

[42] Akshay A. Sarawade and Nadir N. Charniya. Infrared thermography and its applications: A review. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pages 280–285, 2018.

[43] B.B. Lahiri, S. Bagavathiappan, T. Jayakumar, and John Philip. Medical applications of infrared thermography: A review. *Infrared Physics & Technology*, 55 (4):221–235, 2012.

[44] JM Tognarelli, M Dawood, MI Shariff, VP Grover, MM Crossey, IJ Cox, SD Taylor-Robinson, and MJ McPhail. Magnetic resonance spectroscopy: Principles and techniques: Lessons for clinicians. *J Clin Exp Hepatol*, 5(4):320–328, Dec 2015.

[45] Gerard J O'Sullivan, Fiona L Carty, and Carmel G Cronin. Imaging of bone metastasis: An update. *World J Radiol*, 7(8):202–211, Aug 2015.

[46] Emilia Fortunati, Giulia Argalia, Lucia Zanoni, Stefano Fanti, and Valentina Ambrosini. New pet radiotracers for the imaging of neuroendocrine neoplasms. *Curr Treat Options Oncol*, 23(5):703–720, May 2022.

[47] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21 (11):977–1000, 2003.

[48] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, Jul 2022.

[49] Emmanuel Ahishakiye, Martin Bastiaan Van Gijzen, Julius Tumwiine, Ruth Wario, and Johnes Obungoloch. A survey on deep learning in medical image reconstruction. *Intelligent Medicine*, 1(3):118–127, 2021.

[50] Amandeep Kaur, Yadvinder Singh, Neeru Neeru, et al. A survey on deep learning approaches to medical images and a systematic look up into real-time object detection. *Archives of Computational Methods in Engineering*, 29(4):2071–2111, Oct 2022.

[51] Nilay Ganatra. A comprehensive study of applying object detection methods for medical image analysis. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 821–826, 2021.

[52] Samir S. Yadav and Shivajirao M. Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6 (1):113, 2019.

[53] K Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph*, 31(4-5):198–211, Jun-Jul 2007.

[54] Alexander D. Cohen, Ziyi Chen, Oisin Parker Jones, Chenhui Niu, and Yang Wang. Regression-based machine-learning approaches to predict task activation using resting-state fmri. *Hum Brain Mapp*, 41(3):815–826, Feb 2020.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

[55] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5:48, 2022.

[56] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, November 2018.

[57] Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*, 385(3):283–286, July 2021.

[58] Simon Caton and Christian Haas. Fairness in machine learning: A survey, 2020.

[59] Esther Puyol-Antón, Bram Ruijsink, Jorge Mariscal Harana, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, Phil Chowienczyk, and Andrew P King. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. *Frontiers in cardiovascular medicine*, 9:859310, 2022.

[60] Stefanos Ioannou, Hana Chockler, Alexander Hammers, Andrew P King, and Alzheimer's Disease Neuroimaging Initiative. A study of demographic bias in cnn-based brain mr segmentation. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 13–22. Springer, 2022.

[61] Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt EJ Michels, Gerard Schouten, and Veronika Cheplygina. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pages 183–192. Springer, 2020.

[62] Lauren N. Ross. Causal concepts in biology: How pathways differ from mechanisms and why it matters. *The British Journal for the Philosophy of Science*, 72 (1):131–158, 2021.

[63] Ellery Wulczyn, David F. Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, Narayan Hegde, Apaar Sadhwani, Robert MacDonald, Benny Ayalew, Greg S. Corrado, Lily H. Peng, Daniel Tse, Heimo Müller, Zhaoyang Xu, Yun Liu, Martin C. Stumpe, Kurt Zatloukal, and Craig H. Mermel. Interpretable survival prediction for colorectal cancer using deep learning. *npj Digital Medicine*, 4(1):71, April 2021.

[64] Vincenzo L'Imperio, Ellery Wulczyn, Markus Plass, Heimo Müller, Nicolò Tamini, Luca Gianotti, Nicola Zucchini, Robert Reihs, Greg S. Corrado, Dale R. Webster, Lily H. Peng, Po-Hsuan Cameron Chen, Marialuisa Lavitrano, Yun Liu, David F. Steiner, Kurt Zatloukal, and Fabio Pagni. Pathologist Validation of a Machine Learning–Derived Feature for Colon Cancer Risk Stratification. *JAMA Network Open*, 6(3): e2254891, March 2023.

[65] Chi-Tung Cheng, Chih-Chi Chen, Chih-Yuan Fu, Chung-Hsien Chaou, Yu-Tung Wu, Chih-Po Hsu, Chih-Chen Chang, I-Fang Chung, Chi-Hsun Hsieh, Ming-Ju Hsieh, et al. Artificial intelligence-based education assists medical students' interpretation of hip fracture. *Insights into Imaging*, 11:1–8, 2020.

[66] Jeffrey D Rudie, Jeffrey Duda, Michael Tran Duong, Po-Hao Chen, Long Xie, Robert Kurtz, Jeffrey B Ware, Joshua Choi, Raghav R Mattay, Emmanuel J Botzolakis, et al. Brain mri deep learning and bayesian inference system augments radiology resident performance. *Journal of Digital Imaging*, 34(4):1049–1058, 2021.

[67] Max W. Shen. Trust in ai: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient, 2022.

[68] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco California USA, August 2016. ACM.

[69] Liam G. McCoy, Connor T.A. Brenna, Stacy S. Chen, Karina Vold, and Sunit Das. Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, 142:252–257, February 2022.

[70] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks, 2018.

[71] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing, 2020.

[72] Manuel Arias Chao, Chetan Kulkarni, Kai Goebel, and Olga Fink. Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, 217:107961, 2022.

[73] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11:3673, 2020.

[74] Adrian V. Dalca, John Guttag, and Mert R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[75] Nathan Gaw, Andrea Hawkins-Daarud, Leland S Hu, Hyunsoo Yoon, Lujia Wang, Yanzhe Xu, Pamela R Jackson, Kyle W Singleton, Leslie C Baxter, Jennifer

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

Eschbacher, et al. Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric mri. *Scientific reports*, 9(1):10063, 2019.

[76] Maxim Pisov, Mikhail Goncharov, Nadezhda Kurochkina, Sergey Morozov, Victor Gombolevskiy, Valeria Chernina, Anton Vladzymyrskyy, Ksenia Zamyatina, Anna Chesnokova, Igor Pronin, Michael Shifrin, and Mikhail Belyaev. Incorporating task-specific structural knowledge into cnns for brain midline shift detection, 2019.

[77] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[78] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

[79] Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12): 749–762, 2017.

[80] Andreas M. Rauschecker, Jeffrey D. Rudie, Long Xie, Jiancong Wang, Michael Tran Duong, Emmanuel J. Botzolakis, Asha M. Kovalovich, John Egan, Tessa C. Cook, R. Nick Bryan, Ilya M. Nasrallah, Suyash Mohan, and James C. Gee. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain mri. *Radiology*, 295(3):626–637, 2020.

[81] Kaustav Bera, Nathaniel Braman, Amit Gupta, Vamsidhar Velcheti, and Anant Madabhushi. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nature Reviews Clinical Oncology*, 19(2):132–146, 2022.

[82] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:1189–1232, 2001.

[83] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24:44–65, 01 2015.

[84] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[85] Minji Bang, Jihwan Eom, Chansik An, Sooyon Kim, Yae Won Park, Sung Soo Ahn, Jinna Kim, Seung-Koo Lee, and Sang-Hyuk Lee. An interpretable multiparametric radiomics model for the diagnosis of schizophrenia using magnetic resonance imaging of the corpus callosum. *Translational psychiatry*, 11(1): 462, 2021.

[86] Haiyan Zhang, Hongyi Chen, Chao Zhang, Aihong Cao, Qingqing Lu, Hao Wu, Jun Zhang, and Daoying Geng. A radiomics feature-based machine learning models to detect brainstem infarction (rmebi) may enable early diagnosis in non-contrast enhanced ct. *European Radiology*, 33(2):1004–1014, 2023.

[87] Su Zhang. Partial dependence of breast tumor malignancy on ultrasound image features derived from boosted trees. *Journal of Electronic Imaging*, 19: 023004, 04 2010.

[88] Batuhan K. Karaman, Elizabeth C. Mormino, and Mert R. Sabuncu. Machine learning based multi-modal prediction of future decline toward alzheimer's disease: An empirical study. *PLOS ONE*, 17:e0277322, 11 2022.

[89] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[90] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[91] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

[92] Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10029–10038, 2021.

[93] Susu Sun, Stefano Woerner, Andreas Maier, Lisa M Koch, and Christian F Baumgartner. Inherently interpretable multi-label classification using class-specific counterfactuals. *arXiv preprint arXiv:2303.00500*, 2023.

[94] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.

[95] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 02 2020.

[96] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv:2011.08891 [cs, eess]*, 11 2021.

[97] Xinyang Feng, Frank A. Provenzano, and Scott A. Small. A deep learning mri approach outperforms other biomarkers of prodromal alzheimer's disease. *Alzheimer's Research & Therapy*, 14, 03 2022.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

[98] Yan-Wei Lee, Sheng-Kai Huang, and Ruey-Feng Chang. Chexgat: A disease correlation-aware network for thorax disease diagnosis from chest x-ray images. *Artificial Intelligence in Medicine*, 132:102382, 10 2022.

[99] Rachel Lea Draelos and Lawrence Carin. Explainable multiple abnormality classification of chest ct volumes, 11 2021.

[100] Heejong Kim and Mert R Sabuncu. Learning to compare longitudinal images. *arXiv preprint arXiv:2304.02531*, 2023.

[101] Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. Melanoma recognition via visual attention. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 793–804. Springer, 2019.

[102] Heechan Yang, Ji-Ye Kim, Hyongsuk Kim, and Shyam P Adhikari. Guided soft attention network for classification of breast cancer histopathology images. *IEEE transactions on medical imaging*, 39(5):1306–1315, 2019.

[103] Chenzhong Yin, Phoebe Imms, Mingxi Cheng, Anar Amgalan, Nahian F Chowdhury, Roy J Massett, Nikhil N Chaudhari, Xinghe Chen, Paul M Thompson, Paul Bogdan, et al. Anatomically interpretable deep learning of brain age captures domain-specific cognitive impairment. *Proceedings of the National Academy of Sciences*, 120(2):e2214634120, 2023.

[104] Janet Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, 06 2014.

[105] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[106] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[107] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15719–15728, 2021.

[108] Alan Q. Wang and Mert R. Sabuncu. A flexible nadaraya-watson head can offer explainable and calibrated classification. *Transactions on Machine Learning Research*, 2023.

[109] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, 94:42–53, 03 2019.

[110] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 440–450, 2022.

[111] R. D. Cook and S Weisberg. Residuals and influence in regression, 1982.

[112] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2017.

[113] Chien Jung Wang, Chadrick A Hamm, LJubomir Savic, Maria Ferrante, Ivo Schobert, Todd Schlachter, Ming Lin, Jeffrey C Weinreb, James S Duncan, Julius Chapiro, and Brian Letzen. Deep learning for liver tumor diagnosis part ii: convolutional neural network interpretation using radiologic imaging features. *European Radiology*, 29(7):3348–3357, 2019.

[114] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *SSRN Electronic Journal*, 2017.

[115] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating realistic counterfactuals with residual generative adversarial nets, 05 2021.

[116] Siqiong Zhou, Upala J. Islam, Nicholaus Pfeiffer, Imon Banerjee, Bhavika K. Patel, and Ashif S. Iquebal. Scgan: Sparse countergan for counterfactual explanations in breast cancer prediction. *medRxiv*, 2023.

[117] Susu Sun, Lisa M Koch, and Christian F Baumgartner. Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? *arXiv preprint arXiv:2307.12344*, 2023.

[118] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images. *arXiv preprint arXiv:2101.07563*, 2021.

[119] Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8309–8319, 2018.

[120] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.

[121] Xinyi Du, Callum Arthurs, Clarisse Ganier, R T Woolf, Zainab Laftah, Manpreet Lakhan, Amr Salam, Bo Wan, Fiona M Watt, Nicholas M Luscombe, and Magnus D Lynch. Clinically relevant vulnerabilities of deep machine learning systems for skin cancer diagnosis. *Journal of Investigative Dermatology*, 141: 916–920, 04 2021.

[122] Rahul Paul, Matthew Schabath, Robert Gillies, Lawrence Hall, and Dmitry Goldgof. Mitigating adversarial attacks on medical image understanding systems. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1517–1521, 2020.

[123] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial exam-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

ples, 2014.

[124] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23:828–841, 10 2019.

[125] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.

[126] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[127] Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules. *Advances in Neural Information Processing Systems*, 34:23359–23373, 2021.

[128] Cristiano Patrício, João C. Neves, and Luís F. Teixeira. Coherent concept-based explanations in medical image and its application to skin lesion diagnosis, 2023.

[129] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[130] Cristiano Patrício, João C Neves, and Luis F Teixeira. Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3798–3807, 2023.

[131] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

[132] Ugne Klimiene, Ričards Marcinkevičs, Patricia Reis Wolfertstetter, Ece Özkan Elsen, Alyssia Paschke, David Niederberger, Sven Wellmann, Christian Knorr, and Julia E Vogt. Multiview concept bottleneck models applied to diagnosing pediatric appendicitis. In *2nd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*. ETH Zurich, Institute for Machine Learning, 2022.

[133] Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5948–5955, 2023.

[134] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 07 2018.

[135] Hyebin Lee, Seong Tae Kim, and Yong Man Ro. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis, 2019.

[136] Ruth E. Baker, Jose-Maria Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, 14(5): 20170660, 2018.

[137] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[138] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training, 2023.

[139] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6.

[140] Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. A review of causality for learning algorithms in medical image analysis, 2022.

[141] Kangrok Oh, Hae Min Kang, Dawoon Leem, Hyungyu Lee, Kyoung Yul Seo, and Sangchul Yoon. Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. *Scientific Reports*, 11(1):1897, 2021.

[142] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080.

[143] Wejdan L. Alyoubi, Wafaa M. Shalash, and Maysoon F. Abulkhair. Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, 20: 100377, 2020.

[144] Mohammed Nasser and Umi Kalthum Yusof. Deep learning based methods for breast cancer diagnosis: A systematic review and future direction. *Diagnostics (Basel)*, 13(1):161, Jan 2023. .

[145] Mengdi Ma, Rui Liu, Chengyuan Wen, and et al. Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms. *Eur Radiol*, 32: 1652–1662, 2022. . URL https://doi.org/10.1007/s00330-021-08271-4.

[146] D.A. Spak, J.S. Plaxco, L. Santiago, M.J. Dryden, and B.E. Dogan. Bi-rads® fifth edition: A summary of changes. *Diagnostic and Interventional Imaging*, 98(3):179–190, 2017. ISSN 2211-5684. . URL https://www.sciencedirect.com/science/article/pii/S2211568417300013.

[147] P. Karatza, K. Dalakleidi, M. Athanasiou, and K.S.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702

Nikita. Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 2310–2313, 2021. .

[148] Zijin Gu, Keith Jamison, Mert Sabuncu, and Amy Kuceyeski. Personalized visual encoding model construction with small data. *Communications Biology*, 5 (1):1382, 2022.

[149] Subba Reddy Oota, Manish Gupta, Raju S. Bapi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. Deep neural networks and brain alignment: Brain encoding and decoding (survey), 2023.

[150] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, 28(12):4136–4160, 10 2017.

[151] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

[152] Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations, 2022.

[153] José P. Amorim, Pedro H. Abreu, João Santos, Marc Cortes, and Victor Vila. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing & Management*, 60(2):103225, 2023.

[154] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.

[155] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, January 2012.

[156] Raja Parasuraman and Dietrich H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3): 381–410, June 2010.

[157] Boris Babic, Sara Gerke, Theodoros Evgeniou, and I. Glenn Cohen. Beware explanations from AI in health care. *Science*, 373(6552):284–286, July 2021.

[158] Ziad Obermeyer and Thomas H. Lee. Lost in Thought — The Limits of the Human Mind and the Future of Medicine. *New England Journal of Medicine*, 377(13): 1209–1211, September 2017.

[159] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in.

[160] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.



ALAN Q. WANG is a PhD candidate at Cornell University in the School of Electrical and Computer Engineering. He is also affiliated with Cornell Tech and the Department of Radiology at Weill Cornell Medical School. Previously, he obtained his B.S. degree in Computer Engineering at the University of Illinois at Urbana-Champaign (UIUC).

His research interests include machine learning and computer vision, especially applied to medical imaging. More specifically, he is interested in bridging the gap between deep learning models and human experts like doctors and clinicians, which leads him to be interested in research in interpretability, explainability, interactivity, and robustness of deep learning models.



BATUHAN K. KARAMAN is a PhD candidate at Cornell University in the School of Electrical and Computer Engineering. He is also affiliated with Cornell Tech and the Department of Radiology at Weill Cornell Medicine. Previously, he obtained his MS degree in Electrical and Computer Engineering at Cornell University.

His research interests are in machine learning and biomedical data analysis. Focusing on Alzheimer's, his current research develops interpretable and explainable deep learning techniques for the early prediction of future cognitive decline toward dementia, and analysis of associated biomarkers.



HEEJONG KIM is a postdoctoral researcher in Radiology at Weill Cornell Medicine. She completed her Ph.D. in computer science at New York University, New York, NY. Her research interest lies in discovering machine learning techniques and their application to biomedical images.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3387702



JACOB ROSENTHAL is an M.D.-Ph.D. candidate in the Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional M.D.-Ph.D. Program, New York, NY, USA. Previously, he obtained his M.Sc. degree in health data science from the Harvard T.H. Chan School of Public Health, Boston, MA, USA and his B.A. in biology from Oberlin College, Oberlin, OH, USA.

His research interests lie in the domain of machine learning for medicine, specifically in the development and deployment of machine learning methods to improve accuracy, efficiency, and fairness of clinical workflows.



RACHIT SALUJA is a PhD candidate at Cornell University in the School of Electrical and Computer Engineering. He is also affiliated with Cornell Tech and the Department of Radiology at Weill Cornell Medical School. Previously, he obtained his M.S. degree in Electrical Engineering from the University of Pennsylvania and obtained his B.E. degree in Electrical and Electronics Engineering from PES Institute of Technology, India. His research interests lie in building Multi-modal AI models for Clinical Radiology and developing robust Radiology AI systems.



SEAN I. YOUNG is an Instructor at the Martinos Center for Biomedical Imaging, Harvard Medical School, and a Resesarch Affiliate in CSAIL, MIT. Previously, he was a postdoctoral scholar at Stanford University, Stanford, CA. He received his PhD degree in electrical engineering from the University of New South Wales, Sydney, Australia. He received the APRS/IAPR best paper award at DICTA 2018, together with David Taubman.



MERT R. SABUNCU received a PhD degree in Electrical Engineering from Princeton University in 2006, where his dissertation focused on entropy-based approaches to image registration. Mert then moved to the Massachusetts Institute of Technology for a post-doc at the Computer Science and Artificial Intelligence Lab, where he worked on a range of biomedical image analysis problems, including the segmentation of brain MRI scans. After his post-doc at MIT, he spent a few years at the A.A Martinos Center for Biomedical Imaging (Massachusetts General Hospital and Harvard Medical School) as a junior faculty member, where he built a research program on algorithmic tools for integrating genetics and medical imaging. Today, he is a Professor in Electrical and Computer Engineering at Cornell University and Cornell Tech, in New York City. He also holds a faculty appointment in Radiology at Weill Cornell Medicine. His research group develops machine learning based computational tools for biomedical imaging applications. He is a recipient of an NSF CAREER Award (2018) and an NIH Early Career Development Award (2011).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4