A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI

Erico Tjoa, and Cuntai Guan, *Fellow, IEEE*

***Abstract*—Recently, artificial intelligence and machine learning in general have demonstrated remarkable performances in many tasks, from image processing to natural language processing, especially with the advent of deep learning. Along with research progress, they have encroached upon many different fields and disciplines. Some of them require high level of accountability and thus transparency, for example the medical sector. Explanations for machine decisions and predictions are thus needed to justify their reliability. This requires greater interpretability, which often means we need to understand the mechanism underlying the algorithms. Unfortunately, the blackbox nature of the deep learning is still unresolved, and many machine decisions are still poorly understood. We provide a review on interpretabilities suggested by different research works and categorize them. The different categories show different dimensions in interpretability research, from approaches that provide "obviously" interpretable information to the studies of complex patterns. By applying the same categorization to interpretability in medical research, it is hoped that (1) clinicians and practitioners can subsequently approach these methods with caution, (2) insights into interpretability will be born with more considerations for medical practices, and (3) initiatives to push forward data-based, mathematically- and technically-grounded medical education are encouraged.**

***Index Terms*—Explainable Artificial Intelligence, Survey, Machine Learning, Interpretability, Medical Information System.**

I. Introduction

MACHINE LEARNING (ML) has grown large in both research and industrial applications, especially with the success of deep learning (DL) and neural networks (NN), so large that its impact and possible after-effects can no longer be taken for granted. In some fields, failure is not an option: even a momentarily dysfunctional computer vision algorithm in autonomous vehicle easily leads to fatality. In the medical field, clearly human lives are on the line. Detection of a disease at its early phase is often critical to the recovery of patients or to prevent the disease from advancing to more severe stages. While machine learning methods, artificial neural networks, brain-machine interfaces and related subfields have recently demonstrated promising performance in performing medical tasks, they are hardly perfect [1]–[9].

Interpretability and explainability of ML algorithms have thus become pressing issues: who is accountable if things go wrong? Can we explain why things go wrong? If things are working well, do we know why and how to leverage them further? Many papers have suggested different measures and frameworks to capture interpretability, and the topic explainable artificial intelligence (XAI) has become a hotspot in ML research community. Popular deep learning libraries have started to include their own explainable AI libraries, such as Pytorch Captum and Tensorflow tf-explain. Furthermore, the proliferation of interpretability assessment criteria (such as *reliability*, *causality* and *usability*) helps ML community keep track of how algorithms are used and how their usage can be improved, providing guiding posts for further developments [10]–[12]. In particular, it has been demonstrated that visualization is capable of helping researchers detect erroneous reasoning in classification problems that many previous researchers possibly have missed [13].

The above said, there seems to be a lack of uniform adoption of interpretability assessment criteria across the research community. There have been attempts to define the notions of "interpretability", "explainability" along with "reliability", "trustworthiness" and other similar notions without clear expositions on how they should be incorporated into the great diversity of implementations of machine learning models; consider [10], [14]–[18]. In this survey, we will instead use "explainability" and "interpretability" interchangeably, considering a research to be related to interpretability if it does show any attempts (1) to explain the decisions made by algorithms, (2) to uncover the patterns within the inner mechanism of an algorithm, (3) to present the system with coherent models or mathematics, and we will include even loose attempts to raise the credibility of machine algorithms.

In this work, we survey through research works related to the interpretability of ML or computer algorithms in general, categorize them, and then apply the same categories to interpretability in the medical field. The categorization is especially aimed to give clinicians and practitioners a perspective on the use of interpretable algorithms that are available in diverse forms. The trade-off between the ease of interpretation and the need for specialized mathematical knowledge may create a bias in preference for one method compared to another without justification based on medical practices. This may further provide a ground for specialized education in the medical sector that is aimed to realize the potentials that reside within these algorithms. We also find that many journal papers in the machine learning and AI community are algorithm-centric. They often assume that the algorithms used are obviously interpretable without conducting human subject tests to verify their interpretability; see column HSI of table I and II. Note that assuming that a model is obviously interpretable is not necessarily wrong, and, in some cases human tests might be irrelevant (for example pre-defined

Erico T. and Cuntai Guan were with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Erico T. was also affiliated with HealthTech Division, Alibaba Group Holding Limited.

models based on commonly accepted knowledge specific to the content-subject may be considered interpretable without human subject tests). In the tables, we also include a column to indicate whether the interpretability method applies for artificial NN, since the issue of interpretability is recently gathering attention due to its blackbox nature.

We will not attempt to cover all related works many of which are already presented in the research papers and survey we cite [1], [2], [15]–[30]. We extend the so-called *integrated interpretability* [16] by including considerations for subject-content-dependent models. Compared to [17], we also overview the mathematical formulation of common or popular methods, revealing the great variety of approaches to interpretability. Our categorization draws a starker borderline between the different views of interpretability that seem to be difficult to reconcile. In a sense, our survey is more suitable for technically-oriented readers due to some mathematical details, although casual readers may find useful references for relevant popular items, from which they may develop interests in this young research field. Conversely, algorithm users that need interpretability in their work might develop an inclination to understand what is previously hidden in the thick veil of mathematical formulation, which might ironically undermine reliability and interpretability. Clinicians and medical practitioners already having some familiarity with mathematical terms may get a glimpse on how some proposed interpretability methods might be risky and unreliable. The survey [30] views interpretability in terms of extraction of relational knowledge, more specifically, by scrutinizing the methods under *neural-symbolic cycle*. It presents the framework as a sub-category within the interpretability literature. We include it under *verbal interpretability*, though the framework does demonstrate that methods in other categories can be perceived under verbal interpretability as well. The extensive survey [18] provides a large list of researches categorized under *transparent model* and models requiring *post-hoc analysis* with multiple sub-categories. Our survey, on the other hand, aims to overview the state of interpretable machine learning as applied to the medical field.

This paper is arranged as the following. Section II introduces generic types of interpretability and their sub-types. In each section, where applicable, we provide challenges and future prospects related to the category. Section III applies the categorization of interpretabilities in section II to medical field and lists a few risks of machine interpretability in the medical field. Before we proceed, it is also imperative to point out that the issue of accountability and interpretability has spawned discussions and recommendations [31]–[33], and even entered the sphere of ethics and law enforcements [34], engendering movements to protect the society from possible misuses and harms in the wake of the increasing use of AI.

II. Types of Interpretability

There has yet to be a widely-adopted standard to understand ML interpretability, though there have been works proposing frameworks for interpretability [10], [13], [35]. In fact, different works use different criteria, and they are justifiable in one way or another. Reference [36] suggests *network dissection* for the interpretability of visual representations and offers a way to quantify it as well. The interactive websites [37], [38] have suggested a unified framework to study interpretabilities that have thus-far been studied separately. The paper [39] defines a unified measure of *feature importance* in the SHAP (SHapley Additive exPlanations) framework. Here, we categorize existing interpretabilities and present a non-exhaustive list of works in each category.

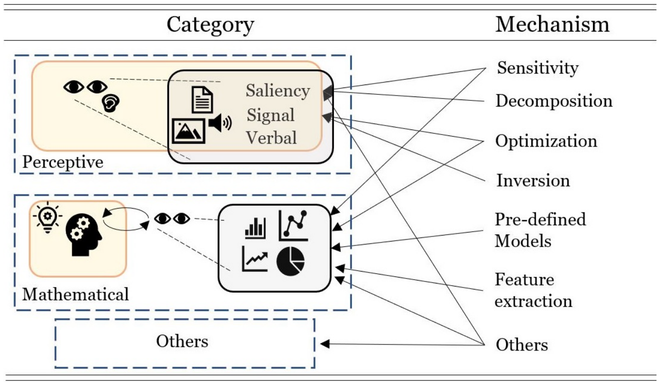


Fig. 1. Overview of categorization with illustration. Orange box: interpretability interface to demarcate the separation between interpretable information and the cognitive process required to understand them. Grey box: algorithm output/product that is proposed to provide interpretability. Black arrow: computing or comprehension process. The perceptive interpretability methods generate items that are usually considered immediately interpretable. On the other hand, methods that provide interpretability via mathematical structure generate outputs that require one more layer of cognitive processing interface before reaching the interpretable interface. The eyes and ear icons represent human senses interacting with items generated for interpretability.

The two major categories presented here, namely *perceptive interpretability* and *interpretability by mathematical structures*, as illustrated in fig. 1, appear to present different polarities within the notion of interpretability. An example of the difficulty with perceptive interpretability is as the following. When a visual “evidence” is given erroneously, the algorithm or method used to generate the “evidence” and the underlying mathematical structure sometimes do not offer any useful clues on how to fix the mistakes. On the other hand, a mathematical analysis of patterns may provide information in high dimensions. They can only be easily perceived once the pattern is brought into lower dimensions, abstracting some fine-grained information we could not yet prove is not discriminative with measurable certainty.

A. Perceptive Interpretability

We include in this category interpretabilities that can be humanly perceived, often one that will be considered obvious. For example, as shown in fig. 2(A2), an algorithm that classifies an image into the cat category can be considered obviously interpretable if it provides segmented patch showing the cat as the explanation. We should note that this alone might on the other hand be considered insufficient, because (1) it still does not un-blackbox an algorithm and (2) it ignores the possibility of using background objects for its decision. The following are the sub-categories to perceptive interpretability. Refer to fig. 3 for the overview of the common sub-categories.

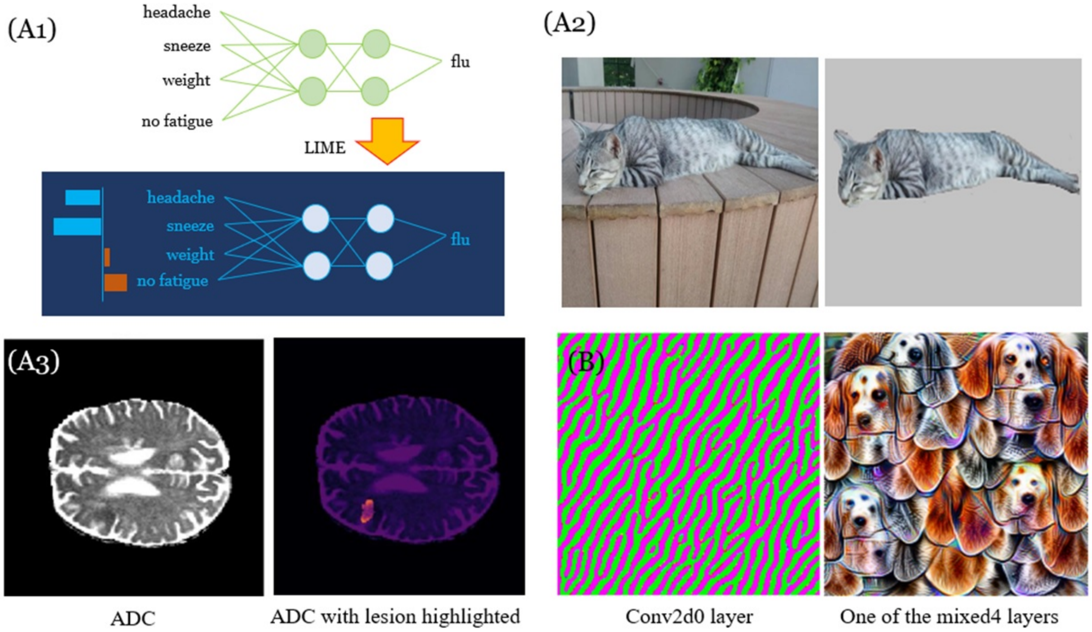


Fig. 2. (A1) Using LIME to generate explanation for text classification. *Headache* and *sneeze* are assigned positive values. This means both factors have positive contribution to the model prediction flu. On the other hand, *weight* and *no fatigue* contribute negatively to the prediction. (A2) LIME is used to generate the super-pixels for the classification cat. (A3) ADC modality of a slice of MRI scan from ISLES 2017 segmentation competition. Reddish intensity region reflects a possible explanation to the choice of segmentation (segmentation not shown) (B) Optimized images that maximize the activation of a neuron in the indicated layers. In shallower layer, simple patterns activate neurons strongly while in deeper layer, more complex features such as dog faces and ears do. Figure (B) is obtained from https://distill.pub/2018/building-blocks/ with permission from Chris Olah.

A.1) Saliency

Saliency method explains the decision of an algorithm by assigning values that reflect the importance of input components in their contribution to that decision. These values could take the forms of probabilities and super-pixels such as heatmaps etc. For example, fig. 2(A1) shows how a model predicts that the patient suffers from flu from a series of factors, but LIME [14] explains the choice by highlighting the importance of the particular symptoms that indicate that the illness should indeed be flu. Similarly, [40] computes the scores reflecting the n-grams activating convolution filters in NLP (Natural Language Processing). Fig. 2(A2) demonstrates the output that LIME will provide as the explanation for the choice of classifications cat and fig. 2(A3) demonstrates a kind of heatmap that shows the contribution of pixels to the segmentation result (segmentation result not shown, and this figure is only for demonstration). More formally, given that model f makes a prediction $y = f(x)$ for input x, for some metric v, typically large magnitude of $v(x_i)$ indicates that the component x_i is a significant reason for the output y.

Saliency methods via decomposition have been developed. In general, they decompose signals propagated within their algorithm and selectively rearrange and process them to provide interpretable information. Class Activation Map (CAM) has been a popular method to generate heat/saliency/relevance-map (from now, we will use the terms interchangeably) that corresponds to discriminative features for classifications [41]–[43]. The original implementation of CAM [41] produces heatmaps using $f_k(x, y)$, the pixel-wise activation of unit k across spatial coordinates (x, y) in the last convolutional layers, weighted by w_k^c, the coefficient corresponding to unit k for class c. CAM at pixel (x, y) is thus given by $M_c(x, y) = \Sigma_k w_k^c f_k(x, y)$.

Similarly, widely used Layer-wise Relevance Propagation (LRP) is introduced in [44]. Some papers that use LRP to construct saliency maps for interpretability include [13], [45]–[50]. It is also applicable for video processing [51]. A short summary for LRP is given in [52]. LRP is considered a decomposition method [53]. Indeed, the importance scores are decomposed such that the sum of the scores in each layer will be equal to the output. In short, the relevance score is the pixel-wise intensity at the input layer $R^{(0)}$ where $R_i^{(l)} = \Sigma_j \frac{a_i^{(l)} w_{ij}^+}{\Sigma_i a_i^{(l)} w_{ij}^+} R_j^{(l+1)}$ is the relevance score of neuron i at layer l with the input layer being at $l = 0$. Each pixel (x, y) at the input layer is assigned the importance value $R^{(0)}(x, y)$, although some combinations of relevance scores $\{R_c^{(l)}\}$ at inner layer l over different channels $\{c\}$ have been demonstrated to be meaningful as well (though possibly less precise; see the tutorial in its website heatmapping.org). LRP can be understood in Deep Taylor Decomposition framework [54]. The code implementation can also be found in the aforementioned website.

Automatic Concept-based Explanations (ACE) algorithm [55] uses super-pixels as explanations. Other decomposition methods that have been developed include, DeepLIFT and gradient*input [56], Prediction Difference Analysis [57] and [40]. Peak Response Mapping [58] is generated by back-propagating peak signals. Peak signals are normalized and treated as probability, and the method can be seen as decomposition into probability transitions. In [59], *Removed*

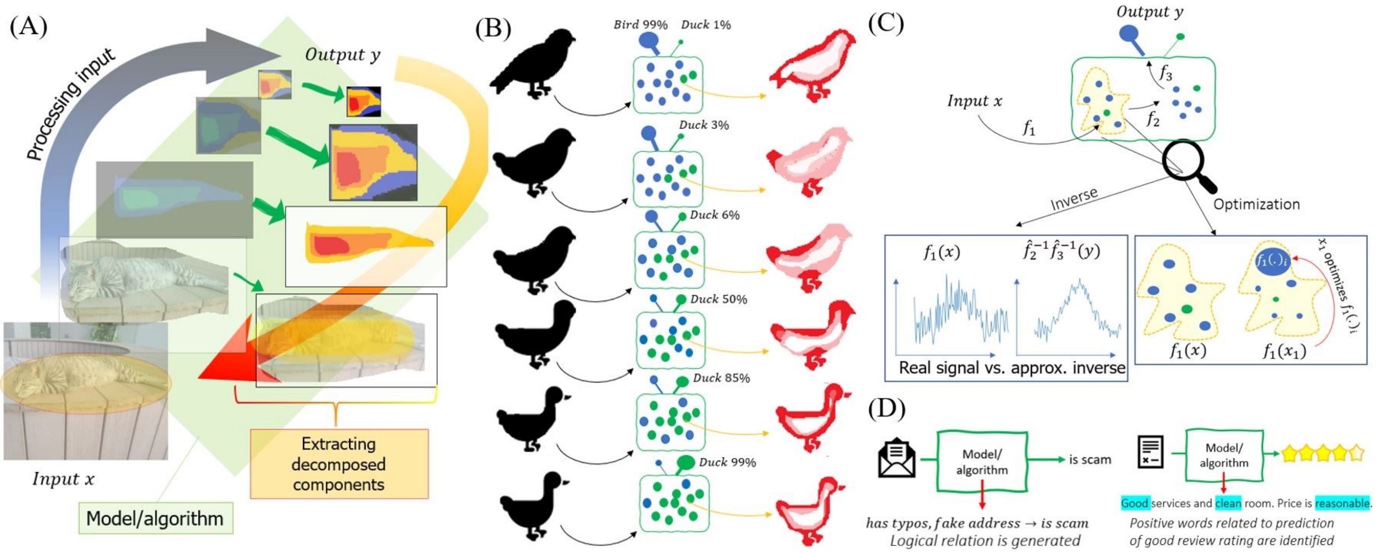

(A)
Processing input
Output y
Input x
Extracting decomposed components
Model/algorithm
(B)
Bird 99%
Duck 1%
Duck 3%
Duck 6%
Duck 50%
Duck 85%
Duck 99%
(C)
Output y
Input x
f_1
f_2
f_3
Inverse
Optimization
$f_1(x)$
$\hat{f}_2^{-1}\hat{f}_3^{-1}(y)$
Real signal vs. approx. inverse
$f_1(.)_i$
x_1 optimizes $f_1(.)_i$
$f_1(x)$
$f_1(x_1)$
(D)
Model/ algorithm
is scam
has typos, fake address → is scam
Logical relation is generated
Model/ algorithm
Good services and clean room. Price is reasonable.
Positive words related to prediction of good review rating are identified

Fig. 3. Overview on perceptive interpretability methods. (A) *Saliency method with decomposition mechanism*. The input which is an image of a cat is fed into the model for processing along the blue arrow. The resulting output and intermediate signals (green arrows) are decomposed and selectively picked for processing, hence providing information for the intermediate mechanism of the model in the form of (often) heatmappings, shown in red/orange/yellow colors. (B) *Saliency method with sensitivity mechanism*. The idea is to show how small changes to the input (black figures of birds and ducks) affect the information extracted for explainability (red silhouette). In this example, red regions indicate high relevance, which we sometimes observe at edges or boundary of objects, where gradients are high. (C) *Signal method by inversion and optimization*. Inverses of signals or data propagated in a model could possibly reveal more sensible information (see arrow labeled "Inverse"). Adjusting input to optimize a particular signal (shown as the i-th component of the function f_1) may provide us with x_1 that reveals explainable information (see arrow labeled "Optimizatin"). For illustration, we show that the probability of correctly predicting duck improves greatly once the head is changed to the head of a duck which the model recognizes. (D) *Verbal interpretability* is typically achieved by ensuring that the model is capable of providing humanly understable statements, such as the logical relation or the positive words shown.

correlation ρ is proposed as a metric to measure the quality of signal estimators. And then it proposes PatternNet and PatternAttribution that backpropagate parameters optimized against ρ, resulting in saliency maps as well. SmoothGrad [60] improves gradient-based techniques by adding noises. Do visit the related website that displays numerous visual comparison of saliency methods; be mindful of how some heatmaps highlight apparently irrelevant regions.

For natural language processing or sentiment analysis, saliency map can also take the form of heat scores over words in texts, as demonstrated by [61] using LRP and by [62]. In the medical field (see later section), [6], [43], [63]–[69] have studied methods employing saliency and visual explanations. Note that we also sub-categorize LIME as a method that uses optimization and sensitivity as its underlying mechanisms, and many researches on interpretability span more than one sub-categories.

Challenges and Future Prospects. As seen, the formulas for CAM and LRP are given on a heuristic: certain ways of interaction between weights and the strength of activation of some units within the models will eventually produce the interpretable information. The intermediate processes are not amenable to scrutiny. For example, taking one of the weights and changing its value does not easily reveal any useful information. How these prescribed ways translate into interpretable information may also benefit from stronger evidences, especially evidences beyond visual verification of localized objects. Signal methods to investigate ML models (see later section) exist, but such methods that probe them with respect to the above methods have not been attempted systematically, possibly opening up a different research direction.

A.2) Signal Method

Methods of interpretability that observe the stimulation of neurons or a collection of neurons are called signal methods [70]. On the one hand, the activated values of neurons can be manipulated or transformed into interpretable forms. For example, the activation of neurons in a layer can be used to reconstruct an image similar to the input. This is possible because neurons store information systematically [71]: *feature maps* in the deeper layer activate more strongly to complex features, such as human face, keyboard etc while feature maps in the shallower layers show simple patterns such as lines and curves. Note: an example of *feature map* is the output of a convolutional filter in a Convolutional Neural Network (CNN). On the other hand, parameters or even the input data might be optimized with respect to the activation values of particular neurons using methods known as *activation optimization* (see a later section). The following are the relevant sub-categories.

Feature maps and Inversions for Input Reconstructions. A feature map often looks like a highly blurred image with most region showing zero (or low intensity), except for the patch that a human could roughly discern as a detected feature. Sometimes, these discernible features are considered interpretable, as in [71]. However, they might be too distorted.

Then, how else can a feature map be related to a humanly-perceptible feature? An inverse convolution map can be defined: for example, if feature map in layer 2 is computed in the network via $y_2 = f_2(f_1(x))$ where x is the input, $f_1(.)$ consists of 7x7 convolutions of stride 2 followed by max-pooling and likewise $f_2(.)$. Then [71] reconstructs an image using a deconvolution network by approximately inversing the trained convolutional network $\tilde{x} = deconv(y) = \hat{f}_2^{-1}\hat{f}_1^{-1}(y)$ which is an approximation, because layers such as max-pooling have no unique inverse. It is shown that $\tilde{x}$ does appear like slightly blurred version of the original image, which is distinct to human eye. Inversion of image representations within the layers has also been used to demonstrate that CNN layers do store important information of an input image accurately [72], [73]. Guided backpropagation [74] modifies the way backpropagation is performed to achieve inversion by zeroing negative signals from both the output or input signals backwards through a layer. Indeed, inversion-based methods do use saliency maps for visualization of the activated signals.

Activation Optimization. Besides transforming the activation of neurons, signal method also includes finding input images that optimize the activation of a neuron or a collection of neurons. This is called the *activation maximization.* Starting with a noise as an input x, the noise is slowly adjusted to increase the activation of a select (collection of) neuron(s) $\{a_k\}$. In simple mathematical terms, the task is to find $x_0 = argmax\ ||\{a_k\}||$ where optimization is performed over input x and $||.||$ is a suitable metric to measure the combined strength of activations. Finally the optimized input that maximizes the activation of the neuron(s) can emerge as something visually recognizable. For example, the image could be a surreal fuzzy combination of swirling patterns and parts of dog faces, as shown in fig. 2(B).

Research works on activation maximization include [75] on MNIST dataset, [76] and [77] that uses a regularization function. In particular, [37] provides an excellent interactive interface (feature visualization) demonstrating activation-maximized images for GoogLeNet [78]. GoogLeNet has a deep architecture, from which we can see how neurons in deeper layer stores complex features while shallower layer stores simple patterns; see fig. 2(B). To bring this one step further, the semantic dictionary is used [38] to provide a visualization of activations within a higher-level organization and semantically more meaningful arrangements.

Other Observations of Signal Activations. Ablation studies [79], [80] also study the roles of neurons in shallower and deeper layers. In essence, some neurons are corrupted and the output of the corrupted neural network is compared to the original network.

Challenges and Future Prospects. Signal methods might have revealed some parts of the black-box mechanisms. Many questions still remain.

- What do we do with the (partially) reconstructed images and images that optimize activation?
- We might have learned how to approximately inverse signals to recover images, can this help improve interpretability further?
- The components and parts in the intermediate process that reconstruct the approximate images might contain important information; will we be able to utilize them in the future?
- How is explaining the components in this "inverse space" more useful than explaining signals that are forward propagated?
- Similarly, how does looking at intermediate signals that lead to activation optimization help us pinpoint the role of a collection of neurons?
- Optimization of highly parameterized functions notoriously gives non-unique solutions. Can we be sure that optimization that yields combination of surreal dog faces will not yield other strange images with minor alteration?

In the process of answering these questions, we may find hidden clues required to get closer to interpretable AI.

A.3) Verbal Interpretability

This form of interpretability takes the form of verbal chunks that human can grasp naturally. Examples include sentences that indicate causality, as shown in the examples below.

Logical statements can be formed from proper concatenation of predicates, connectives etc. An example of logical statement is the conditional statement. Conditional statements are statements of the form $A \to B$, in another words *if A then B*. An ML model from which logical statements can be extracted directly has been considered obviously interpretable. The survey [30] shows how interpretability methods in general can be viewed under such symbolic and relational system. In the medical field, see for example [81], [82].

Similarly, *decision sets* or *rule sets* have been studied for interpretability [83]. The following is a single line in a rule set "rainy and grumpy or calm $\to$ dairy or vegetables", directly quoted from the paper. Each line in a rule set contains a clause with an input in *disjunctive normal form* (DNF) mapped to an output in DNF as well. The example above is formally written (rainy$\wedge$grumpy)$\vee$calm$\to$dairy$\vee$vegetables. Comparing three different variables, it is suggested that interpretability of explanations in the form of rule sets is most affected by cognitive chunks, explanation size and little effected by variable repetition. Here, a cognitive chunk is defined as a clause of inputs in DNF and the number of (repeated) cognitive chunks in a rule set is varied. The explanation size is self-explanatory (a longer/shorter line in a rule set, or more/less lines in a rule set). MUSE [84] also produces explanation in the form of decision sets, where interpretable model is chosen to approximate the black-box function and optimized against a number of metrics, including direct optimization of interpretability metrics.

It is not surprising that verbal segments are provided as the explanation in NLP problems. An encoder-generator framework [85] extracts segment like "a very pleasant ruby red-amber color" to justify 5 out of 5-star rating for a product review. Given a sequence of words $x = (x_1, ..., x_l)$ with $x_k \in \mathbb{R}^d$, explanation is given as the subset of the sentence that gives a summary of why the rating is justified. The subset can be expressed as the binary sequence $(z_1, ..., z_l)$ where $z_k = 1(0)$ indicates x_k is (not) in the subset. Then z follows a probability distribution with $p(z|x)$ decomposed by assuming independence to $\Pi_k p(z_k|x)$ where $p(z_k|x) = \sigma_z(W^z[\overrightarrow{h_k}, \overleftarrow{h_k}] + b^z)$, with $\overrightarrow{h_t}, \overleftarrow{h_t}$ being the usual hidden units in the recurrent cell (forward and backward respectively). Similar segments are generated using filter-attribute probability density function to improve the relation between the activation of certain filters and specific attributes [86]. Earlier works on Visual Question Answering (VQA) [87]–[89] are concerned with the generation of texts discussing objects appearing in images.

Challenges and Future Prospects. While texts appear to provide explanations, the underlying mechanisms used to generate the texts are not necessarily explained. For example, NNs and the common variants/components used in text-related tasks such as RNN (recurrent NN), LSTM (long short term memory) are still black-boxes that are hard to troubleshoot in the case of wrong predictions. There have been less works that probe into the inner signals of LSTM and RNN neural networks. This is a possible research direction, although similar problem as mentioned in the previous sub-subsection may arise (what to do with the intermediate signals?). Furthermore, while word embedding is often optimized with the usual loss minimization, there does not seem to be a coherent explanation to the process and shape of the optimized embedding. There may be some clues regarding optimization residing within the embedding, and thus successfully interpreting the shape of embedding may help shed light into the mechanism of the algorithm.

B. Interpretability via Mathematical Structure

Mathematical structures have been used to reveal the mechanisms of ML and NN algorithms. In the previous section, deeper layer of NN is shown to store complex information while shallower layer stores simpler information [71]. TCAV [95] has been used to show similar trend, as suggested by fig. 5(A2). Other methods include clustering such as t-SNE (t-Distributed Stochastic Neighbor Embedding) shown in fig. 5(B) and subspace-related methods, for example correlation-based Singular Vector Canonical Correlation Analysis (SVCCA) [96] is used to find the significant directions in the subspace of input for accurate prediction, as shown in figure 5(C). Information theory has been used to study interpretability by considering Information Bottleneck principle [97], [98]. The rich ways in which mathematical structures add to the interpretability pave ways to a comprehensive view of the interpretability of algorithms, hopefully providing a ground for unifying the different views under a coherent framework in the future. Fig. 4 provides an overview of ideas under this category.

B.1) Pre-defined Model

To study a system of interest, especially complex systems with not well-understood behaviour, mathematical formula such as parametric models can help simplify the tasks. With a proper hypothesis, relevant terms and parameters can be designed into the model. Interpretation of the terms come naturally if the hypothesis is either consistent with available knowledge or at least developed with good reasons. When the systems are better understood, these formula can be improved by the inclusion of more complex components. In the medical field (see later section), an example is *kinetic modelling*. Machine learning can be used to compute the parameters defined in the models. Other methods exist, such as integrating commonly available methodologies with subject specific contents etc. For example, Generative Discriminative Models [99], combining ridge regression and least square method to handle variables for analyzing Alzheimer's disease and schizophrenia.

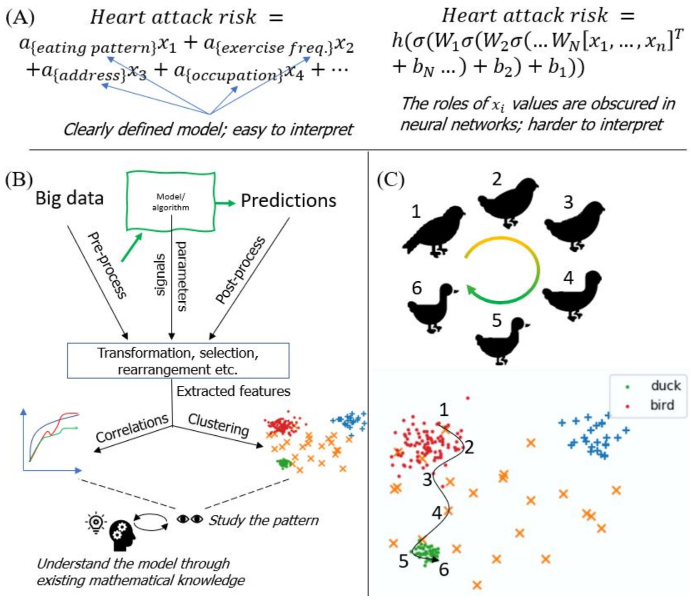

(A)
$Heart\ attack\ risk\ = a_{\{eating\ pattern\}}x_1 + a_{\{exercise\ freq.\}}x_2 + a_{\{address\}}x_3 + a_{\{occupation\}}x_4 + \cdots$
Clearly defined model; easy to interpret
$Heart\ attack\ risk\ = h(\sigma(W_1\sigma(W_2\sigma(\dots W_N[x_1,\dots,x_n]^T + b_N \dots) + b_2) + b_1))$
The roles of x_i values are obscured in neural networks; harder to interpret
(B)
Big data
Model/ algorithm
Predictions
Pre-process
signals
parameters
Post-process
Transformation, selection, rearrangement etc.
Extracted features
Correlations
Clustering
Study the pattern
Understand the model through existing mathematical knowledge
(C)
1
2
3
4
5
6
duck
bird
1
2
3
4
5
6

Fig. 4. Overview of methods whose interpretability depend on interpreting underlying mathematical structure. (A) *Pre-defined models*. Modeling with clear, easily understandable model, such as linear model can help improve readability, and hence interpretability. On the other hand, using neural network could obscure the meaning of input variables. (B) *Feature extraction*. Data, predicted values, signals and parameters from a model are processed, transformed and selectively picked to provide useful information. Mathematical knowledge is usually required to understand the resulting pattern. (C) *Sensitivity*. Models that rely on sensitivity, gradients, perturbations and related concepts will try to account for how different data are differently represented. In the figure, the small changes transforming the bird to the duck can be traced along a map obtained using clustering.

Linearity. The simplest interpretable pre-defined model is the linear combination of variables $y = \Sigma_i a_i x_i$ where a_i is the degree of how much x_i contributes to the prediction y. A linear combination model with $x_i \in \{0, 1\}$ has been referred to as the *additive feature attribution method* [39]. If the model performs well, this can be considered highly interpretable. However, many models are highly non-linear. In such cases, studying interpretability via linear properties (for example, using linear probe; see below) are useful in several ways, including the ease of implementation. When linear property appears to be insufficient, non-linearity can be introduced; it is typically not difficult to replace the linear component $\overrightarrow{w} \cdot \overrightarrow{a}$ within the system with a non-linear version $f(\overrightarrow{w}, \overrightarrow{a})$.

A linear probe is used in [100] to extract information from each layer in a neural network. More technically, assume we have deep learning classifier $F(x) \in [0, 1]^D$ where $F_i(x) \in [0, 1]$ is the probability that input x is classified into class i out of D classes. Given a set of features H_k at layer k of a neural network, then the linear probe f_k at layer k is defined as a linear classifier $f_k : H_k \rightarrow [0, 1]^D$ i.e. $f(h_k) = softmax(Wh_k + b)$. In another words, the probe tells us how well the information from only layer k can predict the output, and each of this predictive probe is a linear classifier by design. The paper then shows plots of the error rate of the prediction made by each f_k against k and demonstrates that these linear

TABLE I

LIST OF JOURNAL PAPERS ARRANGED ACCORDING TO THE INTERPRETABILITY METHODS USED, HOW INTERPRETABILITY IS PRESENTED OR THE SUGGESTED MEANS OF INTERPRETABILITY. THE TABULATION PROVIDES A NON-EXHAUSTIVE OVERVIEW OF INTERPRETABILITY METHODS, PLACING SOME DERIVATIVE METHODS UNDER THE UMBRELLA OF THE MAIN METHODS THEY DERIVE FROM. HSI: HUMAN STUDY ON INTERPRETABILITY ✓MEANS THERE IS HUMAN STUDY DESIGNED TO VERIFY IF THE SUGGESTED METHODS ARE INTERPRETABLE BY THE HUMAN SUBJECT. ANN: ✓MEANS EXPLICITLY INTRODUCES NEW ARTIFICIAL NEURAL NETWORK ARCHITECTURE, MODIFIES EXISTING NETWORKS OR PERFORMS TESTS ON NEURAL NETWORKS.

Methods	HSI	ANN	Mechansim		
CAM with global average pooling [41], [90]	✗	✓	Decomposition	Saliency	Perceptive Interpretability
+ Grad-CAM [42] generalizes CAM, utilizing gradient	✓	✓			
+ Guided Grad-CAM and Feature Occlusion [67]	✗	✓			
+ Respond CAM [43]	✗	✓			
+ Multi-layer CAM [91]	✗	✓			
LRP (Layer-wise Relevance Propagation) [13], [52]	✗	✓			
+ Image classifications. PASCAL VOC 2009 etc [44]	✗	✓			
+ Audio classification. AudioMNIST [46]	✗	✓			
+ LRP on DeepLight. fMRI data from Human Connectome Project [47]	✗	✓			
+ LRP on CNN and on BoW(bag of words)/SVM [48]	✗	✓			
+ LRP on compressed domain action recognition algorithm [49]	✗	✗			
+ LRP on video deep learning, *selective relevance method* [51]	✗	✓			
+ BiLRP [50]	✗	✓			
DeepLIFT [56]	✗	✓			
Prediction Difference Analysis [57]	✗	✓			
Slot Activation Vectors [40]	✗	✓			
PRM (Peak Response Mapping) [58]	✗	✓			
LIME (Local Interpretable Model-agnostic Explanations) [14]	✓	✓	Sensitivity		
+ MUSE with LIME [84]	✓	✓			
+ Guidelinebased Additive eXplanation optimizes complexity, similar to LIME [92]	✓	✓			
# Also listed elsewhere: [55], [68], [70], [93]	N.A.	N.A.			
Others. Also listed elsewhere: [94]	N.A.	N.A.	Others		
+ Direct output labels. Training NN via multiple instance learning [64]	✗	✓			
+ Image corruption and testing Region of Interest statistically [65]	✗	✓			
+ Attention map with autofocus convolutional layer [66]	✗	✓			
DeconvNet [71]	✗	✓	Inversion	Signal	
Inverting representation with natural image prior [72]	✗	✓			
Inversion using CNN [73]	✗	✓			
Guided backpropagation [74], [90]	✗	✓			
Activation maximization/optimization [37]	✗	✓	Optimization		
+ Activation maximization on DBN (Deep Belief Network) [75]	✗	✓			
+ Activation maximization, multifaceted feature visualization [76]	✗	✓			
Visualization via regularized optimization [77]	✗	✓			
Semantic dictionary [38]	✗	✓			
Decision trees	N.A.	N.A.	Verbal		
Propositional logic, rule-based [81]	✗	✗			
Sparse decision list [82]	✗	✗			
Decision sets, rule sets [83], [84]	✓	✗			
Encoder-generator framework [85]	✗	✓			
Filter Attribute Probability Density Function [86]	✗	✗			
MUSE (Model Understanding through Subspace Explanations) [84]	✓	✓			

classifiers generally perform better at deeper layer, that is, at larger k.

General Additive Models. Linear model is generalized by the Generalized Additive Model (GAM) [101], [102] with standard form $g(E[y]) = \beta_0 + \Sigma f_j(x_j)$ where g is the *link function*. The equation is general, and specific implementations of f_j and link function depend on the task. The familiar General Linear Model (GLM) is GAM with the specific implementation of linear f_j and g is the identity. Modifications can be duly implemented. As a natural extension to the model, interaction terms between variables $f_{ij}(x_i, x_j)$ are used [103]; we can certainly extend this indefinitely. ProtoAttend [104] uses probabilities as weights in the linear component of the NN. Such model is considered inherently interpretable by the authors. In the medical field, see [81], [99], [105], [106].

Content-subject-specific model. Some algorithms are considered obviously interpretable within its field. Models are designed based on existing knowledge or empirical evidence, and thus interpretation of the models is innately embedded into the system. ML algorithms can then be incorporated in rich and diverse ways, for example, through parameter fitting. The following lists just a few works to illustrate the usage diversity of ML algorithms. Deep Tensor Neural Network is used for quantum many-body systems [107]. Atomistic neural network architecture for quantum chemistry is used in [108], where each atom is like a node in a graph with a set of feature vectors. The specifics depend on the neural network used, but this model is considered inherently interpretable. Neural network has been used for programmable wireless environments (PWE) [109]. TS approximation [110] is a fuzzy network approximation of other neural networks. The approximate fuzzy system is constructed with choices of components that can be adapted to the context of interpretation. The paper itself uses sigmoid-based membership function, which it considers interpretable.

A so-called model-based reinforcement learning is suggested to be interpretable after the addition of high level knowledge about the system that is realized as Bayesian structure [111].

Challenges and Future Prospects. The challenge of formulating the "correct" model exists regardless of machine learning trend. It might be interesting if a system is found that is fundamentally operating on a specific machine learning model. Backpropagation-based deep NN (DNN) itself is inspired by the brain, but they are not operating at fundamental level of similarity (nor is there any guarantee that such model exists). When interpretability is concerned, having fundamental similarity to real, existing systems may push forward our understanding of machine learning model in unprecedented ways. Otherwise, in the standard uses of machine learning algorithm, different optimization paradigms are still being discovered. Having optimization paradigm that is specialized for specific models may be contribute to a new aspect of interpretable machine learning.

B.2) Feature Extraction

We give an intuitive explanation via a hypothetical example of a classifier for heart-attack prediction. Given, say, 100-dimensional features including eating pattern, job and residential area of a subject. A kernel function can be used to find out that the strong predictor for heart attack is a 100-dimensional vector which is significant in the following axes: eating pattern, exercise frequency and sleeping pattern. Then, this model is considered interpretable because we can link heart-attack risk with healthy habits rather than, say socio-geographical factors. More information can be drawn from the next most significant predictor and so on.

Correlation. The methods discussed in this section include the use of correlation in a general sense. This will naturally include covariance matrix and correlation coefficients after transformation by kernel functions. A kernel function transforms high-dimensional vectors such that the transformed vectors better distinguish different features in the data. For example, the Principal Component Analysis transforms vectors into the principal components (PC) that can be ordered by the eigenvalues of singular-value-decomposed (SVD) covariance matrix. The PC with the highest eigenvalue is roughly the most informative feature. Many kernel functions have been introduced, including the Canonical Correlation Analysis (CCA) [112]. CCA provides the set of features that transforms the original variables to the pairs of canonical variables, where each pair is a pair of variables that are "best correlated" but not correlated to other pairs. Quoted from [113], "such features can inherently characterize the object and thus it can better explore the insights and finer details of the problems at hand". In the previous sections, interpretability research using correlation includes [59].

SVCCA combines CCA and SVD to analyze interpretability [96]. Given an input dataset $X = \{x_1, ..., x_m\}$ where each input x_i is possibly multi-dimensional. Denote the activation of neuron i at layer l as $z_i^l = (z_i^l(x_1), ..., z_i^l(x_m))$. Note that one such output is defined for the entire input dataset. SVCCA finds out the relation between 2 layers of a network $l_k = \{z_i^{l_k} | i = 1, ..., m_k\}$ for $k = 1, 2$ by taking l_1 and l_2 as the input (generally, l_k does not have to be the entire layer). SVCCA uses SVD to extract the most informative components l'_k and uses CCA to transform l'_1 and l'_2 such that $\bar{l}'_1 = W_X l'_1$ and $\bar{l}'_2 = W_X l'_2$ have the maximum correlation $\rho = \{\rho_1, ..., \rho_{min(m_1,m_2)}\}$. One of the SVCCA experiments on CIFAR-10 demonstrates that only 25 most-significant axes in l'_k are needed to obtain nearly the full accuracy of a full-network with 512 dimensions. Besides, the similarity between 2 compared layers is defined to be $\bar{\rho} = \frac{1}{min(m_1,m_2)} \Sigma_i \rho_i$.

The successful development of generative adversarial networks (GAN) [114]–[116] for generative tasks have spawned many derivative works. GAN-based models have been able to generate new images not distinguishable from synthetic images and perform many other tasks, including transferring style from one set of images to another or even producing new designs for products and arts. Studies related to interpretabilities exist. For example [117] uses encoder-decoder system to perform multi-stage PCA. Generative model is used to show that natural image distribution modelled using probability density is fundamentally difficult to interpret [118]. This is demonstrated through the use of GAN for the estimation of image distribution density. The resulting density shows preferential accumulation of density of images with certain features (for examples, images featuring small object with few foreground distractions) in the pixel space. The paper then suggests that interpretability is improved once it is embedded in the deep feature space, for example, from GAN. In this sense, the interpretability is offered by better correlation between the density of images with the correct identification of the objects. Consider also the GAN-based works they cite.

Clustering. Algorithm such as t-SNE has been used to cluster input images based on their activation of neurons in a network [76], [119]. The core idea relies on the distance between objects being considered. If the distance between two objects are short in some measurement space, then they are similar. This possibly appeals to the notion of human learning by the *Law of Association*. It differs from correlation-based method which provides some metrics that relate the change of one variable with another, where the two related objects can originate from completely different domains; clustering simply presents their similarity, more sensibly in similar domain or in the subsets thereof. In [119], the activations $\{f_{fc7}(x)\}$ of 4096-dimensional layer fc7 in the CNN are collected over all input $\{x\}$. Then $\{f_{fc7}(x)\}$ is fed into t-SNE to be arranged and embedded into two-dimension for visualization (each point then is visually represented by the input image x). Activation atlases are introduced in [120], which similarly uses t-SNE to arrange some activations $\{f_{act}(x)\}$, except that each point is represented by the average activations of feature visualization. In meta-material design [121], design pattern and optical responses are encoded into latent variables to be characterized by Variational Auto Encoder (VAE). Then, t-SNE is used to visualize the latent space.

In the medical field (also see later section), we have [122], [123] (uses Laplacian Eigenmap for interpretability) [124] (introduces a low-rank representation method for Autistic Spectrum Diagnosis).

Challenges and Future Prospects. This section exemplifies the difficulty in integrating mathematics and human intuition.

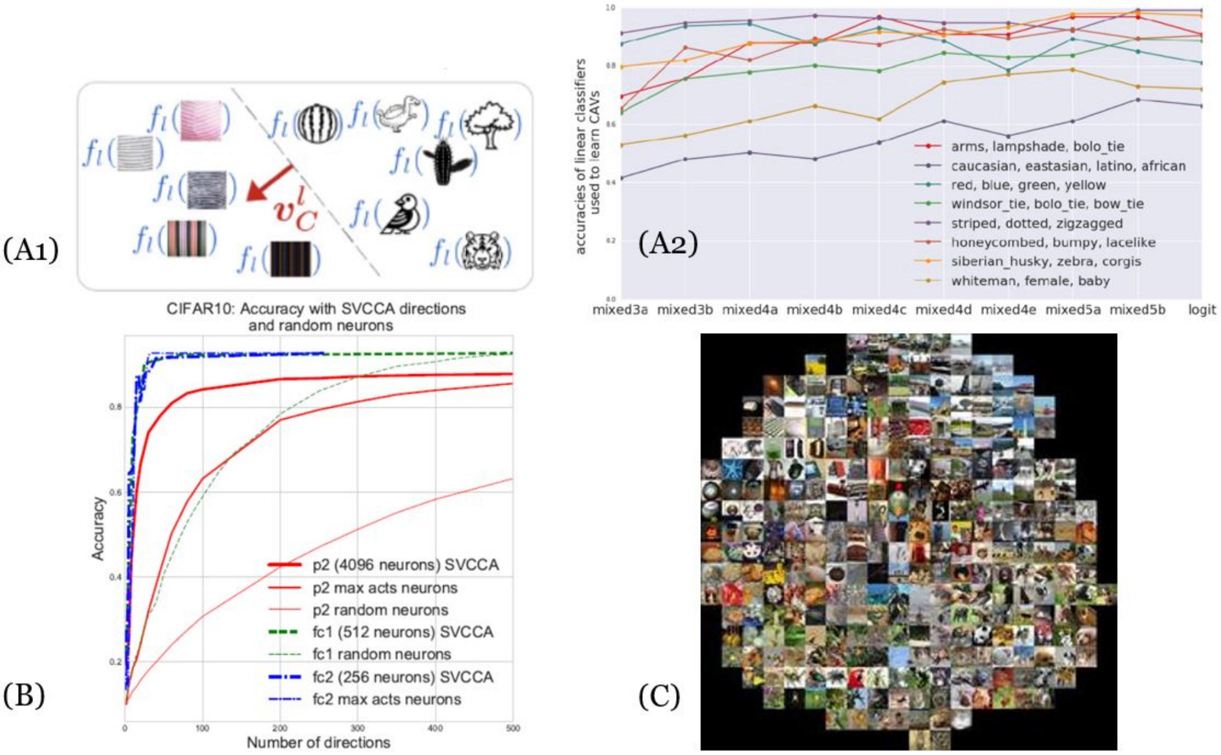


Fig. 5. (A1) TCAV [95] method finds the hyperplane CAV that separates concepts of interest. (A2) Accuracies of CAV applied to different layers supports the idea that deeper NN layers contain more complex concepts, and shallower layers contain simpler concepts. (B) SVCCA [96] finds the most significant subspace (direction) that contains the most information. The graph shows that as few as 25 directions out of 500 are enough to produce the accuracies of the full network. (C) t-SNE clusters images in meaningful arrangement, for example dog images are close together. Figures (A1,A2) are used with permission from the authors Been Kim; figures (B,C) from Maithra Raghu and Jascha Sohl-dickstein.

Having extracted ”relevant” or ”significant” features, sometimes we are left with still a combination of high dimensional vectors. Further analysis comes in the form of correlations or other metrics that attempt to show similarities or proximity. The interpretation may stay as mathematical artifact, but there is a potential that separation of concepts attained by these methods can be used to reorganize a black-box model from within. It might be an interesting research direction that lacks justification in terms of real-life application: however, progress in unraveling black-boxes may be a high-risk high-return investment.

B.3) Sensitivity

We group together methods that rely on localization, gradients and perturbations under the category of sensitivity. These methods rely on the notion of small changes dx in calculus and the neighborhood of a point in metric spaces.

Sensitivity to input noises or neighborhood of data points. Some methods rely on the locality of some input x. Let a model $f(.)$ predicts $f(x)$ accurately for some x. Denote $x+\delta$ as a slightly noisy version of x. The model is locally faithful if $f(x+\delta)$ produces correct prediction, otherwise, the model is unfaithful and clearly such instability reduces its reliability. Reference [125] introduces *meta-predictors* as interpretability methods and emphasizes the importance of the variation of input x to neural network in explaining a network. They define *explanation* and *local explanation* in terms of the response of blackbox f to some input. Amongst many of the studies conducted, they provide experimental results on the effect of varying input such as via deletion of some regions in the input. Likewise, when random pixels of an image are deleted (hence the data point is shifted to its neighborhood in the feature space) and the resulting change in the output is tested [56], pixels that are important to the prediction can be determined. In text classification, [126] provides explanations in the form of partitioned graphs. The explanation is produced in three main steps, where the first step involves sampling perturbed versions of the data using VAE.

Testing with Concept Activation Vectors (TCAV) has also been introduced as a technique to interpret the low-level representation of neural network layer [95]. First, the concept activation vector (CAV) is defined. Given input $x \in \mathbb{R}^n$ and a feedforward layer l having m neurons, the activation at that layer is given by $f_l : \mathbb{R}^n \to \mathbb{R}^m$. If we are interested in the concept C, for example striped pattern, then, using TCAV, we supply a set P_C of examples corresponding to striped pattern (zebra, clothing pattern etc) and the negative examples N. This collection is used to train a binary classifier $v_C^l \in \mathbb{R}^m$ for layer l that partitions $\{f_l(x) : x \in P_C\}$ and $\{f_l(x) : x \in N\}$. In another words, a kernel function extracts features by mapping out a set of activations that has relevant information about the stripe-ness. CAV is thus defined as the normal vector to the hyperplane that separates the positive examples from the negative ones, as shown in fig. 5(A1). It then computes directional derivative $S_{v,k,l}(x) = \nabla h_{l,k}(f_l(x)) \cdot v_C^l$ to obtain the sensitivity of the model w.r.t. the concept C, where $h_{l,k}$ is the logit function for class k of C for layer l.

LIME [14] optimizes over models $g \in G$ where G is a set of interpretable models G by minimizing locality-aware loss and complexity. In another words, it seeks to obtain the optimal model $\xi(x) = argmin_{g\in G}\ L(f, g, \pi_x) + \Omega(g)$ where Ω is the complexity and f is the true function we want to model. An example of the loss function is $L(f, g, \pi_x) = \Sigma_{z,z'\in Z}\pi_x(z)[f(x) - g(z')]^2$ with $\pi_x(z)$ being, for example, Euclidean distance and Z is the vicinity of x. From the equation, it can be seen that the desired g will be close to f in the vicinity Z of x, because $f(z) \approx g(z')$ for $z, z' \in Z$. In another words, noisy inputs z, z' do not add too much losses.

Gradient-based explanation vector $\xi(x_0) = \frac{\partial}{\partial x}P(Y \neq g(x_0)|X = x)$ is introduced by [127] for Bayesian classifier $g(x) = argmin_{c\in\{i,...,C\}}\ P(Y \neq c|X = x)$, where x, ξ are d-dimensional. For any $i = 1, ..., d$, high absolute value of $[\xi(x_0)]_i$ means that component i contributes significantly to the decision of the classifier. If it is positive, the higher the value is, the less likely x_0 contributes to decision $g(x_0)$.

ACE algorithm [55] uses TCAV to compute saliency score and generate super-pixels as explanations. Grad-CAM [42] is a saliency method that uses gradient for its sensitivity measure. In [128], *influence function* is used. While theoretical, the paper also practically demonstrates how understanding the underlying mathematics will help develop perturbative training point for adversarial attack.

Sensitivity to dataset. A model is possibly sensitive to the training dataset $\{x_i\}$ as well. Influence function is also used to understand the effect of removing x_i for some i and shows the consequent possibility of adversarial attack [128]. Studies on adversarial training examples can be found in the paper and its citations, where seemingly random, insignificant noises can degrade machine decision considerably. The *representer theorem* is introduced for studying the extent of effect x_i has on a decision made by a deep NN [129].

Challenges and Future Prospects. There seems to be a concern with locality and globality of the concepts. As mentioned in [95], to achieve *global quantification* for interpretability, explanation must be given for a set of examples or the entire class rather than "just explain individual data inputs". As a specific example, there may be a concern with the globality of TCAV. From our understanding, TCAV is a perturbation method by the virtue of stable continuity in the usual derivative and it is global because the whole subset of dataset with label k of concept C has been shown to be well-distinguished by TCAV. However, we may want to point out that despite their claim to globality, it is possible to view the success of TCAV as local, since it is only "global within each label k rather than within all dataset considered at once.

From the point of view of image processing, the neighborhood of a data point (an image) in the feature space poses a rather subtle question; also refer to fig. 4(C) for related illustration. For example, after rotating and stretching the image or deleting some pixels, how does the position of the image in the feature space change? Is there any way to control the effect of random noises and improve robustness of machine prediction in a way that is sensible to human's perception? The transition in the feature space from one point to another point that belongs to different classes is also unexplored.

On a related note, gradients have played important roles in formulating interpretability methods, be it in image processing or other fields. Current trend recognizes that regions in the input space with significant gradients provide interpretability. Deforming these regions quickly degrades the prediction; conversely, the particular values at these regions are important to the reach a certain prediction. This is helpful, since calculus exists to help analyse gradients. However, this has shown to be disruptive as well. For example, imperceptible noises can degrade prediction drastically (see *manipulation of explanations* under the section *Risk of Machine Interpretation in Medical Field*). Since gradient is also in the core of loss optimization, it is a natural target for further studies.

B.4) Optimization

We have described several researches that seek to attain interpretability via optimization methods. Some have optimization at the core of their algorithm, but the interpretability is left to visual observation, while others optimize interpretability mathematically.

Quantitatively maximizing interpretability. To approximate a function f, as previously mentioned, LIME [14] performs optimization by finding optimal model $\xi \in G$ so that $f(z) \approx \xi(z')$ for $z, z' \in Z$ where Z is the vicinity of x, so that local fidelity is said to be achieved. Concurrently the complexity $\Omega(\xi)$ is minimized. Minimized Ω means the models interpretability is maximized. MUSE [84] takes in blackbox model, prediction and user-input features to output decision sets based on optimization w.r.t fidelity, interpretability and unambiguity. The available measures of interpretability that can be optimized include *size*,*featureoverlap* etc (refer to table 2 of its appendix).

Activation Optimization. Activation optimizations are used in research works such as [37], [75]–[77] as explained in a previous section. The interpretability relies on direct observation of the neuron-activation-optimized images. While the quality of the optimized images are not evaluated, the fact that parts of coherent images emerge with respect to a (collection of) neuron(s) does demonstrate some organization of information in the neural networks.

C. Other Perspectives to Interpretability

There are many other concepts that can be related to interpretability. Reference [42] conducted experiments to test the improvements of human performance on a task after being given explanations (in the form of visualization) produced by ML algorithms. We believe this might be an exemplary form of interpretability evaluation. For example, we want to compare machine learning algorithms ML_A with ML_B. Say, human subjects are given difficult classification tasks and attain a baseline 40% accuracy. Repeat the task with different set of human subjects, but they are given explanations churned out by ML_A and ML_B. If the accuracies attained are now 50% and 80% respectively, then ML_B is more interpretable.

Even then, if human subjects cannot really explain why they can perform better with the given explanations, then the interpretability may be questionable. This brings us to the question of what kind of interpretability is necessary in different tasks and certainly points to the possibility that there is no need for a unified version of interpretability.

C.1) Data-driven Interpretability

Data in catalogue. A large amount of data has been crucial to the functioning of many ML algorithms, mainly as the input data. In this section, we mention works that put a different emphasize on the treatment of these data arranged in catalogue. In essence, [10] suggests that we create a matrix whose rows are different real-world tasks (e.g. pneumonia detection), columns are different methods (e.g. decision tree with different depths) and the entries are the performance of the methods on some *end-task*. How can we gather a large collection of entries into such a large matrix? Apart from competitions and challenges, crowd-sourcing efforts will aid the formation of such database [147], [148]. A clear problem is how multi-dimensional and gigantic such tabulation will become, not to mention that the collection of entries is very likely uncountably many. Formalizing interpretability here means we pick latent dimensions (common criteria) that human can evaluate e.g. time constraint or time-spent, cognitive chunks (defined as the basic unit of explanation, also see the definition in [83]) etc. These dimensions are to be refined along iterative processes as more user-inputs enter the repository.

Incompleteness. In [10], the problem of *incompleteness* of problem formulation is first posed as the issue in interpretability. Incompleteness is present in many forms, from the impracticality to produce all test-cases to the difficulty in justifying why a choice of proxy is the best for some scenarios. At the end, it suggests that interpretability criteria are to be born out of collective agreements of the majority, through a cyclical process of discoveries, justifications and rebuttals. In our opinion, a disadvantage is that there is a possibility that no unique convergence will be born, and the situation may aggravate if, say, two different conflicting factions are born, each with enough advocate. The advantage lies in the existence of strong roots for the advocacy of certain choice of interpretability. This prevents malicious intent from tweaking interpretability criteria to suit ad hoc purposes.

C.2) Invariances

Implementation invariance. Reference [93] suggests implementation invariance as an axiomatic requirement to interpretability. In the paper, it is stated as the following. Define two *functionally equivalent* functions as f_1, f_2 so that $f_1(x) = f_x(x)$ for any x regardless of their implementation details. Given any two such networks using attribution method, then the attribution functional A will map the importance of each component of an input to f_1 the same way it does to f_2. In another words, $(A[f_1](x))_j = (A[f_2](x))_j$ for any $j = 1, , d$ where d is the dimension of the input. The statement can be easily extended to methods that do not use attribution as well.

Input invariance. To illustrate using image classification problem, translating an image will also translate super-pixels demarcating the area that provides an explanation to the choice of classification correspondingly. Clearly, this property is desirable and has been proposed as an axiomatic invariance of a reliable saliency method. There has also been a study on the input invariance of some saliency methods with respect to translation of input $x \to x+c$ for some c [70]. Of the methods studied, gradients/sensitivity-based methods [127] and signal methods [71], [74] are input invariant while some attribution methods, such as integrated gradient [93], are not.

C.3) Interpretabilities by Utilities

The following utilities-based categorization of interpretability is proposed by [10].

Application-based. First, an evaluation is application-grounded if human A gives explanation X_A on a specific application, so-called the end-task (e.g. a doctor performs diagnosis) to human B, and B performs the same task. Then A has given B a useful explanation if B performs better in the task. Suppose A is now a machine learning model, then the model is highly interpretable if human B performs the same task with improved performance after given X_A. Some medical segmentation works will fall into this category as well, since the segmentation will constitute a visual explanation for further diagnosis/prognosis [143], [144] (also see other categories of the grand challenge). Such evaluation is performed, for example, by [42]. They proposed Grad-CAM applied on guided backpropagation (proposed by [74]) of AlexNet CNN and VGG. The produced visualizations are used to help human subjects in Amazon Mechanical Turks identify objects with higher accuracy in predicting VOC 2007 images. The human subjects achieved 61.23% accuracy, which is 16.79% higher than visualization provided by guided backpropagation.

Human-based. This evaluation involves real humans and simplified tasks. It can be used when, for some reasons or another, having human A give a good explanation X_A is challenging, possibly because the performance on the task cannot be evaluated easily or the explanation itself requires specialized knowledge. In this case, a simplified or partial problem may be posed and X_A is still demanded. Unlike the application-based approach, it is now necessary to look at X_A specifically for interpretability evaluation. Bigger pool of human subjects can then be hired to give a generic valuation to X_A or create a model answer $\hat{X}_A$ to compare X_A with, and then a generic valuation is computed.

Now, suppose A is a machine learning model, A is more interpretable compared to another ML model if it scores better in this generic valuation. In [145], a ML model is given a document containing the conversation of humans making a plan. The ML model produces a ”report” containing relevant predicates (words) for the task of inferring what the final plan is. The metric used for interpretability evaluation is, for example, the percentage of the predicates that appear, compared to human-made report. We believe the format of human-based evaluation needs not be strictly like the above. For example, hybrid human and interactive ML classifiers require human users to nominate features for training [146]. Two different standard MLs can be compared to the hybrid, and one can be said to be more interpretable than another if it picks up features similar to the hybrid, assuming they perform at similarly acceptable level.

Functions-based. Third, an evaluation is functionally-grounded if there exist proxies (which can be defined a priori) for evaluation, for example sparsity [10]. Some papers [2],

TABLE II
(CONTINUED FROM TABLE I) LIST OF JOURNAL PAPERS ARRANGED ACCORDING TO THE INTERPRETABILITY METHODS USED, HOW INTERPRETABILITY IS PRESENTED OR THE SUGGESTED MEANS OF INTERPRETABILITY.

Methods	HSI	ANN	Mechanism		
Linear probe [100]	✗	✓	Pre-defined models		Interpretability via Mathematical Structure
Regression based on CNN [105]	✗	✓			
Backwards model for interpretability of linear models [106]	✗	✗			
GDM (Generative Discriminative Models): ridge regression + least square [99]	✗	✗			
GAM, GA^{2}M (Generative Additive Model) [81], [101], [102]	✗	✗			
ProtoAttend [104]	✗	✓			
Other content-subject-specific models:	N.A.	N.A.			
+ Kinetic model for CBF (cerebral blood flow) [130]	N.A.	✓			
+ CNN for PK (Pharmacokinetic) modelling [131]	N.A.	✓			
+ CNN for brain midline shift detection [132]	N.A.	✓			
+ Group-driven RL (reinforcement learning) on personalized healthcare [133]	N.A.	✓			
+ Also see [107]–[111]	N.A.	✓			
PCA (Principal Components Analysis), SVD (Singular Value Decomposition)	N.A.	N.A.	Correlation	Feature Extraction	
CCA (Canonical Correlation Analysis) [112]	✗	✗			
SVCCA (Singular Vector Canonical Correlation Analysis) [96] = CCA+SVD	✗	✓			
F-SVD (Frame Singular Value Decomposition) [113] on electromyography data	✗	✗			
DWT (Discrete Wavelet Transform) + Neural Network [134]	✗	✓			
MODWPT (Maximal Overlap Discrete Wavelet Package Transform) [135]	✗	✗			
GAN-based Multi-stage PCA [117]	✓	✗			
Estimating probability density with deep feature embedding [118]	✗	✓			
t-SNE (t-Distributed Stochastic Neighbour Embedding) [76]	✗	✓	Clustering		
+ t-SNE on CNN [119]	✗	✓			
+ t-SNE, activation atlas on GoogleNet [120]	✗	✓			
+ t-SNE on latent space in meta-material design [121]	✗	✓			
+ t-SNE on genetic data [136]	✗	✓			
+ mm-t-SNE on phenotype grouping [137]	✗	✓			
Laplacian Eigenmaps visualization for Deep Generative Model [123]	✗	✓			
KNN (k-nearest neighbour) on multi-center low-rank rep. learning (MCLRR) [124]	✗	✓			
KNN with triplet loss and *query-result activation map pair* [138]	✗	✓			
Group-based Interpretable NN with RW-based Graph Convolutional Layer [122]	✗	✓			
TCAV (Testing with Concept Activation Vectors) [95]	✓	✓	Sensitivity		
+ RCV (Regression Concept Vectors) uses TCAV with Br score [139]	✗	✓			
+ Concept Vectors with UBS [140]	✗	✓			
+ ACE (Automatic Concept-based Explanations) [55] uses TCAV	✓	✓			
Influence function [128] helps understand adversarial training points	✗	✓			
Representer theorem [129]	✗	✓			
SocRat (Structured-output Causual Rationalizer) [126]	✗	✓			
Meta-predictors [125]	✗	✓			
Explanation vector [127]	✗	✗			
# Also listed elsewhere: [14], [42], [84], [93]	N.A.	N.A.			
# Also listed elsewhere: [14], [59], [84] etc	N.A.	N.A.	Optimization		
CNN with separable model [141]	✗	✓	Others		
Information theoretic: Information Bottleneck [97], [98]	✗	✓			
Database of methods v.s. interpretability [10]	N.A.	N.A.	Data Driven		Other Persp.
Case-Based Reasoning [142]	✓	✗			
Integrated Gradients [68], [93]	✗	✓	Invariance		
Input invariance [70]	✗	✓			
Application-based [143], [144]	N.A.	N.A.	Utilities		
Human-based [145], [146]					
Function-based [2], [5], [41]–[43], [95], [96], [143], [144]					

[5], [41]–[43], [95], [96], [143], [144] use metrics that rely on this evaluation include many supervised learning models with clearly defined metrics such as (1) Dice coefficients (related to visual interpretability), (2) attribution values, components of canonically transformed variables (see for example CCA) or values obtained from dimensionality reduction methods (such as components of principal components from PCA and their corresponding eigenvalues), where interpretability is related to the degree an object relates to a feature, for example, classification of a dog has high values in the feature space related to four limbs, shape of snout and paws etc. Which suitable metrics to use are highly dependent on the tasks at hand.

III. XAI IN MEDICAL FIELD

ML has also gained traction recently in the medical field, with large volume of works on automated diagnosis, prognosis [149]. From the grand-challenge.org, we can see many different challenges in the medical field have emerged and galvanized researches that use ML and AI methods. Amongst successful deep learning models are [2], [5], using U-Net for medical segmentation. However, being a deep learning neural network, U-Net is still a blackbox; it is not very interpretable. Other domain specific methods and special transformations (denoising etc) have been published as well; consider for example [130] and many other works in MICCAI publications.

In the medical field the question of interpretability is far from just intellectual curiosity. More specifically, it is pointed out that interpretabilities in the medical fields include factors other fields do not consider, including risk and responsibilities [21], [150], [151]. When medical responses are made, lives may be at stake. To leave such important decisions to machines that could not provide accountabilities would be akin to shirking the responsibilities altogether. Apart from ethical issues, this is a serious loophole that could turn catastrophic when exploited with malicious intent.

TABLE III
CATEGORIZATION BY THE ORGANS AFFECTED BY THE DISEASES. NEURO* REFERS TO ANY NEUROLOGICAL, NEURODEVELOPMENTAL, NEURODEGENERATIVE ETC DISEASES. THE ROWS ARE ARRANGED ACCORDING TO THE FOCUS OF THE INTERPRETABILITY AS THE FOLLOWING: APPL.=APPLICATION, METHOD.=METHODOLOGY, COMP.=COMPARISON

Appl.	brain, neuro* [47], [67], [130], [152] [131], [132], [135], [155]	breast [68], lung [6], [81], sleep [153], skin [154] others [105]
Method.	brain, neuro* [65], [66], [82], [90], [99] [113], [122], [134], [156]	breast [64], [69], [139], [140] skin [138], heart [123] others [43], [66], [137], [141]
Comp.	brain, neuro* [106], [157]	lung [92], sleep [158] skin [159], other [136]

Many more works have thus been dedicated to exploring explainability in the medical fields [11], [20], [43]. They provide summaries of previous works [21] including subfield-specific reviews such as [25] for chest radiograph and sentiment analysis in medicine [160], or at least set aside a section to promote awareness for the importance of interpretability in the medical field [161]. In [162], it is stated directly that being a black-box is a "strong limitation" for AI in dermatology, as it is not capable of performing customized assessment by certified dermatologist that can be used to explain clinical evidence. On the other hand, the exposition [163] argues that a certain degree of opaqueness is acceptable, i.e. it might be more important that we produce empirically verified accurate results than focusing too much on how to the unravel the black-box. We recommend readers to consider them first, at least for an overview of interpretability in the medical field.

We apply categorization from the previous section to the ML and AI in the medical field. Table III shows categorization obtained by tagging (1) how interpretability method is incorporated: either through direct application of existing methods, methodology improvements or comparison between interpretability methods and (2) the organs targeted by the diseases e.g. brain, skin etc. As there is not yet a substantial number of significant medical researches that address interpretability, we will refrain from presenting any conclusive trend. However, from a quick overview, we see that the XAI research community might benefit from more studies comparing different existing methods, especially those with more informative conclusion on how they contribute to interpretability.

A. Perceptive Interpretability

Medical data could come in the form of traditional 2D images or more complex formats such as NIFTI or DCOM which contain 3D images with multiple modalities and even 4D images which are time-evolving 3D volumes. The difficulties in using ML for these data include the following. Medical images are sometimes far less available in quantity than common images. Obtaining these data requires consent from the patients and other administrative barriers. High dimensional data also add complexity to data processing and the large memory space requirement might prevent data to be input without modification, random sampling or down-sizing, which may compromise analysis. Other possible difficulties with data collection and management include as left/right-censoring, patients' death due to unrelated causes or other complications etc.

When medical data is available, ground-truth images may not be correct. Not only do these data require some specialized knowledge to understand, the lack of comprehensive understanding of biological components complicates the analysis. For example, ADC modality of MR images and the isotropic version of DWI are in some sense derivative, since both are computed from raw images collected by the scanner. Furthermore, many CT or MRI scans are presented with skull-stripping or other pre-processing. However, without a more complete knowledge of what fine details might have been accidentally removed, we cannot guarantee that an algorithm can capture the correct features.

A.1) Saliency

The following articles consist of direct applications of existing saliency methods. Chexpert [6] uses GradCAM for visualization of pleural effusion in a radiograph. CAM is also used for interpretability in brain tumour grading [152]. Reference [67] uses Guided Grad-CAM and feature occlusion, providing complementary heatmaps for the classification of Alzheimer's disease pathologies. Integrated gradient method and SmoothGrad are applied for the visualization of CNN ensemble that classifies estrogen receptor status using breast MRI [68]. LRP on DeepLight [47] was applied on fMRI data from Human Connectome Project to generate heatmap visualization. Saliency map has also been computed using primitive gradient of loss, providing interpretability to the neural network used for EEG (Electroencephalogram) sleep stage scoring [153]. There has even been a direct comparison between the feature maps within CNN and skin lesion images [154], overlaying the scaled feature maps on top of the images as a means to interpretability. Some images correspond to relevant features in the lesion, while others appear to explicitly capture artifacts that might lead to prediction bias.

The following articles are focused more on comparison between popular saliency methods, including their derivative/improved versions. Reference [158] trains an artificial neural network for the classification of insomnia using physiological network (PN). The feature relevance scores are computed from several methods, including DeepLIFT [56]. Comparison between 4 different visualizations is performed in [157]. It shows different attributions between different methods and concluded that LRP and guided backpropagation provide the most coherent attribution maps in their Alzheimer's disease study. Basic tests on GradCAM and SHAP on dermoscopy images for melanoma classification are conducted, concluding with the need for significant improvements to heatmaps before practical deployment [159].

The following includes slightly different focus on methodological improvements on top of the visualization. Respond-CAM [43] is derived from [41], [42], and provides a saliency-map in the form of heat-map on 3D images obtained from Cellular Electron Cryo-Tomography. High intensity in the heatmap marks the region where macromolecular complexes are present. Multi-layer class activation map (MLCAM) is introduced in [90] for glioma (a type of brain tumor) localization. Multi-instance (MI) aggregation method is used with CNN to classify breast tumour tissue microarray (TMA) image's for 5 different tasks [64], for example the classification of the histologic subtype. Super-pixel maps indicate the region in each TMA image where the tumour cells are; each label corresponds to a class of tumour. These maps are proposed as the means for visual interpretability. Also, see the activation maps in [65] where interpretability is studied by corrupting image and inspecting region of interest (ROI). The autofocus module from [66] promises improvements in visual interpretability for segmentation on pelvic CT scans and segmentation of tumor in brain MRI using CNN. It uses attention mechanism (proposed by [91]) and improves it with adaptive selection of scale with which the network "sees" an object within an image. With the correct scale adopted by the network while performing a single task, human observer analysing the network can understand that a neural network is properly identifying the object, rather than mistaking the combination of the object plus the surrounding as the object itself.

There is also a different formulation for the generation of saliency maps [69]. It defines a different softmax-like formula to extract signals from DNN for visual justification in classification of breast mass (malignant/benign). Textual justification is generated as well.

A.2) Verbal

In [81], a rule-based system could provide the statement *has asthma → lower risk*, where risk here refers to death risk due to pneumonia. Likewise, [82] creates a model called *Bayesian Rule Lists* that provides such statements for stroke prediction. Textual justification is also provided in the LSTM-based breast mass classifier system [69]. The *argumentation theory* is implemented in the machine learning training process [155], extracting arguments or decision rules as the explanations for the prediction of stroke based on the Asymptomatic Carotid Stenosis and Risk of Stroke (ACSRS) dataset.

One should indeed look closer at the interpretability in [81]. Just as many MLs are able to extract some humanly non-intuitive pattern, the rule-based system seems to have captured the strange link between asthma and pneumonia. The link becomes clear once the actual explanation based on real situation is provided: a pneumonia patient which also suffers from asthma is often sent directly to the Intensive Care Unit (ICU) rather than a standard ward. Obviously, if there is a variable ICU=0 or 1 that indicates admission to ICU, then a better model can provide more coherent explanation "*asthma→ICU→lower risk*". In the paper, the model appears not to identify such variable. We can see that interpretability issues are not always clear-cut.

Several researches on Visual Question Answering in the medical field have also been developed. The initiative by ImageCLEF [164], [165] appears to be at its center, though VQA itself has yet to gain more traction and successful practical demonstration in the medical sector before widespread adoption.

Challenges and Future Prospects for perceptive interpretability in medical sector. In many cases, where saliency maps are provided, they are provided with insufficient evaluation with respect to their utilities within the medical practices. For example, when providing importance attribution to a CT scan used for lesion detection, are radiologists interested in heatmaps highlighting just the lesion? Are they more interested in looking for reasons why a haemorrhage is epidural or subdural when the lesion is not very clear to the naked eyes? There may be many such medically-related subtleties that interpretable AI researchers may need to know about.

B. Interpretability via Mathematical Structure

B.1) Pre-defined Model

Models help with interpretability by providing a generic sense of what a variable does to the output variable in question, whether in medical fields or not. A parametric model is usually designed with at least an estimate of the working mechanism of the system, with simplification and based on empirically observed patterns. For example, [130] uses kinetic model for the cerebral blood flow in $ml/100g/min$ with

$$CBF = f(\Delta M) \frac{6000\beta \Delta M exp(\frac{PLD}{T_{1b}})}{2\alpha T_{1b}(SI_{PD})(1-exp(-\frac{\tau}{T_{1b}}))} \tag{1}$$

which depends on perfusion-weighted image ΔM obtained from the signal difference between labelled image of arterial blood water treated with RF pulses and the control image. This function is incorporated in the loss function in the training pipeline of a fully convolutional neural network. At least, an interpretation can be made partially: the neural network model is designed to denoise a perfusion-weighted image (and thus improve its quality) by considering CBF. How the network understands the CBF is again an interpretability problem of a neural network which has yet to be resolved.

There is an inherent simplicity in the interpretability of models based on linearity, and thus they have been considered obviously interpretable as well; some examples include linear combination of clinical variables [99], metabolites signals for MRS [105] etc. Linearity in different models used in the estimation of brain states is discussed in [106], including how it is misinterpreted. It compares what it refers to as forward and backward models and then suggested improvement on linear models. In [81], a logistic regression model picked up a relation between asthma and lower risk of pneumonia death, i.e. asthma has a negative weight as a risk predictor in the regression model. Generative Discriminative Machine (GDM) combines ordinary least square regression and ridge regression to handle confounding variables in Alzheimers disease and schizophrenia dataset [99]. GDM parameters are said to be interpretable, since they are linear combinations of the clinical variables. Deep learning has been used for PET pharmacokinetic (PK) modelling to quantify tracer target density [131]. CNN has helped PK modelling as a part of a sequence of processes to reduce PET acquisition time, and the output is interpreted with respect to the golden standard PK model, which is the linearized version of Simplified Reference Tissue Model (SRTM). Deep learning method is also used to perform parameters fitting for Magnetic Resonance Spectroscopy (MRS) [105]. The parametric part of the MRS signal model specified, $x(t) = \Sigma a_m x_m(t) e^{\Delta\alpha_m t + 2\pi i \Delta f_m t}$, consists of linear combination of metabolite signals $x_m(t)$. The paper shows that the error measured in SMAPE (symmetric mean absolute percentage error) is smallest for most metabolites when their CNN model is used. In cases like this, clinicians may find the model interpretable as long as the parameters are well-fit, although the neural network itself may still not be interpretable.

The models above use linearity for studies related to brain or neuro-related diseases. Beyond linear models, other brain and neuro-systems can be modelled with relevant subject-content knowledge for better interpretability as well. Segmentation task for the detection of brain midline shift is performed using using CNN with standard structural knowledge incorporated [132]. A template called *model-derived age norm* is derived from mean values of sleep EEG features of healthy subjects [156]. Interpretability is given as the deviation of the features of unhealthy subject from the age norm.

On a different note, reinforcement learning (RL) has been applied to personalized healthcare. In particular, [133] introduces group-driven RL in personalized healthcare, taking into considerations different groups, each having similar agents. As usual, Q-value is optimized w.r.t policy π_θ, which can be qualitatively interpreted as the maximization of rewards over time over the choices of action selected by many participating agents in the system.

Challenges and Future Prospects. Models may be simplifying intractable system. As such, the full potential of machine learning, especially DNN with huge number of parameters, may be under-used. A possible research direction that taps onto the hype of predictive science is as the following: given a model, is it possible to augment the model with new, sophisticated components, such that parts of these components can be identified with (and thus interpreted as) new insights? Naturally, the augmented model needs to be comparable to previous models and shown with clear interpretation why the new components correspond to insights previously missed. Do note that there are critiques against the hype around the potential of AI which we will leave to the readers.

B.2) Feature extraction

Vanilla CNN is used in [141] but it is suggested that interpretability can be attained by using a separable model. The separability is achieved by polynomial-transforming scalar variables and further processing, giving rise to weights useful for interpretation. In [122], fMRI is analyzed using correlation-based functional graphs. They are then clustered into super-graph, consisting of subnetworks that are defined to be interpretable. A convolutional layer is then used on the super-graph. For more references about neural networks designed for graph-based problems, see the papers citations. The following are further sub-categorization for methods that revolve around feature extraction and the evaluations or measurements (such as correlations) used to obtain the features, similar to the previous section.

Correlation. DWT-based method (discrete wavelet transform) is used to perform feature extraction before eventually feeding the EEG data (after a series of processings) into a neural network for epilepsy classification [134]. A *fuzzy relation* analogous to correlation coefficient is then defined. Furthermore, as with other transform methods, the components (the wavelets) can be interpreted component-wise. As a simple illustration, the components for Fourier transform could be taken as how much certain frequency is contained in a time series. Reference [135] mentioned a host of wavelet-based feature extraction methods and introduced maximal overlap discrete wavelet package transform (MODWPT) also applied on EEG data for epilepsy classification.

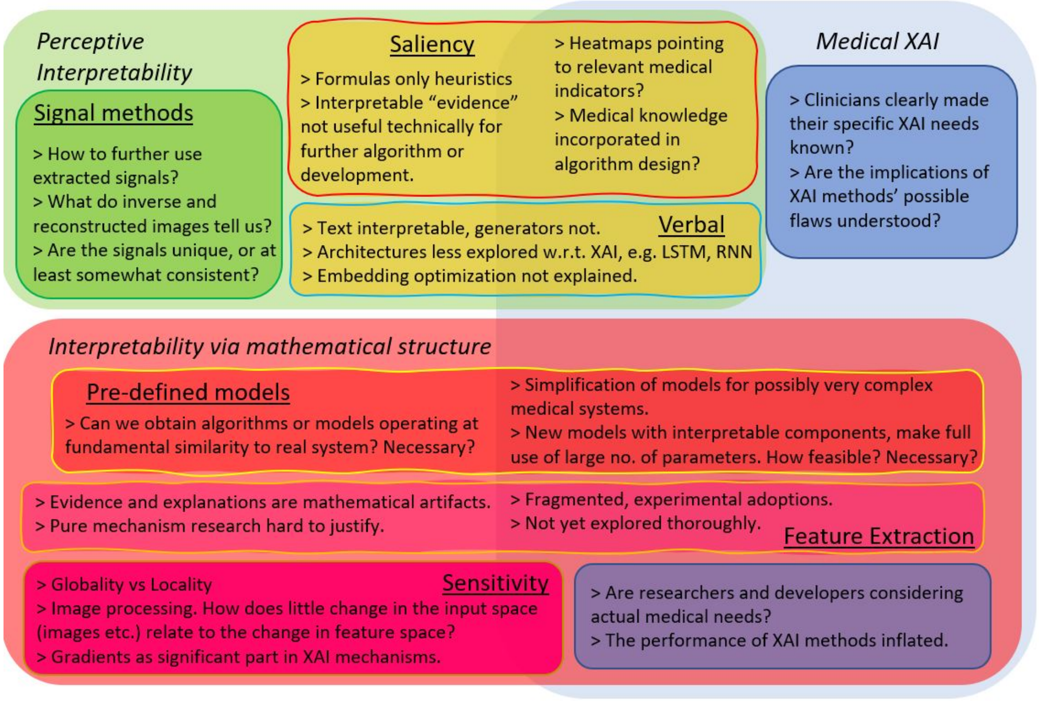
Frame singular value decomposition (F-SVD) is introduced for classifications of electromyography (EMG) data [113]. It is a pipeline involving a number of processing that includes DWT, CCA and SVD, achieving around 98$\%$ accuracies on classifications between amyotrophic lateral sclerosis, myopathy and healthy subjects. Consider also CCA-based papers that are cited in the paper, in particular citations 18 to 21 for EMG and EEG signals.

Clustering. VAE is used to obtain vectors in 64-dimensional latent dimension in order to predict whether the subjects suffer from hypertrophic cardiomyopathy (HCM) [123]. A non-linear transformation is used to create Laplacian Eigenmap (LE) with two dimensions, which is suggested as the means for interpretability. Skin images are clustered [138] for melanoma classification using k-nearest-neighbour that is customized to include CNN and triplet loss. A queried image is then compared with training images ranked according to similarity measure visually displayed as *query-result activation map pair*.

t-SNE has been applied on human genetic data and shown to provide more robust dimensionality reduction compared to PCA and other methods [136]. Multiple maps t-SNE (mm-t-SNE) is introduced by [137], performing clustering on phenotype similarity data.

Sensitivity. Regression Concept Vectors (RCV) is proposed along with a metric *Br* score as improvements to TCAV's concept separation [139]. The method is applied on breast cancer histopathology classification problem. Furthermore, Unit Ball Surface Sampling metric (UBS) is introduced [140] to address the shortcoming of *Br* score. It uses neural networks for classification of nodules for mammographic images. Guidelinebased Additive eXplanation (GAX) is introduced in [92] for diagnosis using CT lung images. Its pipeline includes LIME-like perturbation analysis and SHAP. Comparisons are then made with LIME, Grad-CAM and feature importance generated by SHAP.

Challenges and Future Prospects. We observe popular uses of certain methods ingrained in specific sectors on the one hand and, on the other hand, emerging applications of sophisticated ML algorithms. As medical ML (in particular



Perceptive Interpretability

Signal methods

> How to further use extracted signals?
> What do inverse and reconstructed images tell us?
> Are the signals unique, or at least somewhat consistent?

Saliency

> Formulas only heuristics
> Interpretable "evidence" not useful technically for further algorithm or development.

> Heatmaps pointing to relevant medical indicators?
> Medical knowledge incorporated in algorithm design?

Verbal

> Text interpretable, generators not.
> Architectures less explored w.r.t. XAI, e.g. LSTM, RNN
> Embedding optimization not explained.

Medical XAI

> Clinicians clearly made their specific XAI needs known?
> Are the implications of XAI methods' possible flaws understood?

Interpretability via mathematical structure

Pre-defined models

> Can we obtain algorithms or models operating at fundamental similarity to real system? Necessary?

> Simplification of models for possibly very complex medical systems.
> New models with interpretable components, make full use of large no. of parameters. How feasible? Necessary?

> Evidence and explanations are mathematical artifacts.
> Pure mechanism research hard to justify.

> Fragmented, experimental adoptions.
> Not yet explored thoroughly.

Feature Extraction

Sensitivity

> Globality vs Locality
> Image processing. How does little change in the input space (images etc.) relate to the change in feature space?
> Gradients as significant part in XAI mechanisms.

> Are researchers and developers considering actual medical needs?
> The performance of XAI methods inflated.

Fig. 6. Overview of challenges and future prospects arranged in a Venn diagram.

the application of recently successful DNN) is still a young field, we see fragmented and experimental uses of existing or customized interpretable methods. As medical ML research progresses, the trade-off between many practical factors of ML methods (such as ease of use, ease of interpretation of mathematical structure possibly regarded as complex) and its contribution to the subject matter will become clearer. Future research and application may benefit from a practice of consciously and consistently extracting interpretable information for further processing, and the process should be systematically documented for good dissemination. Currently, with feature selections and extractions focused on improving accuracy and performance, we may still have vast unexplored opportunities in interpretability research.

C. Other Perspectives

Data-driven. Case-Based Reasoning (CBR) performs medical evaluation (classifications etc) by comparing a query case (new data) with similar existing data from a database. [142] combines CBR with an algorithm that presents the similarity between these cases by visually providing proxies and measures for users to interpret. By observing these proxies, the user can decide to take the decision suggested by the algorithm or not. The paper also asserts that medical experts appreciate such visual information with clear decision-support system.

D. Risk of Machine Interpretation in Medical Field

Jumping conclusion. According to [81], logical statements such as *has asthma*$\rightarrow$*lower risk* are considered interpretable. However, in the example, the statement indicates that a patient with asthma has lower risk of death from pneumonia, which might be strange without any clarification from the intermediate thought process. While human can infer that the lowered risk is due to the fact that pneumonia patients with asthma history tend to be given more aggressive treatment, we cannot always assume there is a similar humanly inferable reason behind each decision. Furthermore, interpretability method such as LRP, deconvolution and guided backpropagation introduced earlier are shown to not work for simple model, such as linear model, bringing into question their reliability [59].

IV. Conclusion

We present a survey on interpretability and explainability of ML algorithms in general, and place different interpretations suggested by different research works into distinct categories. From general interpretabilities, we apply the categorization into the medical field. Some attempts are made to formalize interpretabilities mathematically, some provide visual explanations, while others might focus on the improvement in task performance after being given explanations produced by algorithms. At each section, we also discuss related challenges and future prospects. Fig. 6 provides a diagram that summarizes all the challenges and prospects.

Manipulation of explanations. Given an image, a similar image can be generated that is perceptibly indistinguishable from the original, yet produces radically different output [94]. Naturally, its significance attribution and interpretable information become unreliable. Furthermore, explanation can even be manipulated arbitrarily [166]. For example, an explanation for the classification of a cat image (i.e. particular significant values that contribute to the prediction of cat) can be implanted into the image of a dog, and the algorithm could be fooled into classifying the dog image as a cat image. The risk in medical field is clear: even without malicious, intentional manipulation, noises can render explanations wrong. Manipulation of algorithm that is designed to provide explanation is also explored in [167].

Incomplete constraints. In [130], the loss function for the training of a fully convolutional network includes CBF as a constraint. However, many other constraints may play important roles in the mechanism of a living organ or tissue, not to mention applying kinetic model is itself a simplification. Giving an interpretation within limited constraints may place undue emphasis on the constraint itself. Other works that use predefined models might suffer similar problems [99], [105], [131].

Noisy training data. The so-called ground truths for medical tasks, provided by professionals, are not always absolutely correct. In fact, news regarding how AI beats human performance in medical imaging diagnosis [168] indicates that human judgment could be brittle. This is true even of trained medical personnel. This might give rise to the classic garbage-in-garbage-out situation.

The above risks are presented in large part as a reminder of the nature of automation. It is true that algorithms have been used to extract invisible patterns with some successes. However, one ought to view scientific problems with the correct order of priority. The society should not risk over-allocating resources into building machine and deep learning models, especially since due improvements to understanding the underlying science might be the key to solving the root problem. For example, higher quality MRI scans might reveal key information not visible with current technology, and many models built nowadays might not be very successful because there is simply not enough detailed information contained in currently available MRI scans.

Future directions for clinicians and practitioners. Visual and textual explanation supplied by an algorithm might seem like the obvious choice; unfortunately, the details of decision-making by algorithms such as deep neural networks are still not clearly exposed. When an otherwise reliable deep learning model provides a strangely wrong visual or textual explanation, systematic methods to probe into the wrong explanations do not seem to exist, let alone methods to correct them. A specialized education combining medical expertise, applied mathematics, data science etc might be necessary to overcome this. For now, if ”interpretable” algorithms are deployed in medical practices, human supervision is still necessary. Interpretability information should be considered nothing more than complementary support for the medical practices before there is a robust way to handle interpretability.

Future directions for algorithm developers and researchers. Before the blackbox is un-blackboxed, machine decision always carries some exploitable risks. It is also clear that a unified notion of interpretability is elusive. For medical ML interpretability, more comparative studies between the performance of methods will be useful. The interpretability output such as heatmaps should be displayed and compared clearly, including poor results. In the best case scenario, clinicians and practitioners recognize the shortcomings of interpretable methods but have a general idea on how to handle them in ways that are suitable to medical practices. In the worst case scenario, the inconsistencies between these methods can be exposed. The very troubling trend of journal publications emphasizing good results is precarious, and we should thus continue interpretability research with a mindset open to evaluation from all related parties. Clinicians and practitioners need to be given the opportunity for fair judgment of utilities of the proposed interpretability methods, not just flooded with performance metrics possibly irrelevant to the adoption of medical technology.

Also, there may be a need to shift interpretability study away from algorithm-centric studies. An authoritative body setting up the standard of requirements for the deployment of model building might stifle the progress of the research itself, though it might be the most efficient way to reach an agreement. This might be necessary to prevent damages, seeing that even corporate companies and other bodies non-academic in the traditional sense have joined the fray (consider health-tech start-ups and the implications). Acknowledging that machine and deep learning might not be fully mature for large-scale deployment, it might be wise to deploy the algorithms as a secondary support system for now and leave most decisions to the traditional methods. It might take a long time before humanity graduates from this stage, but it might be timely: we can collect more data to compare machine predictions with traditional predictions and sort out data ownership issues along the way.

ACKNOWLEDGMENT

This research was supported by Alibaba Group Holding Limited, DAMO Academy, Health-AI division under Alibaba-NTU Talent Program. The program is the collaboration between Alibaba and Nanyang Technological university, Singapore.

REFERENCES

[1] Eun-Jae Lee, Yong-Hwan Kim, Namkug Kim, and Dong-Wha Kang. Deep into the brain: Artificial intelligence in stroke imaging. *Journal of Stroke*, 19:277–285, 09 2017.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[3] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.*, 58(6):697718, June 2003.

[4] Liang Chen, Paul Bentley, and Daniel Rueckert. Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks. *NeuroImage: Clinical*, 15:633 – 643, 2017.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016.

[6] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019.

[7] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797, 2016.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[9] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, 2019.

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. cite arxiv:1702.08608.

[11] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. *CoRR*, abs/1905.05134, 2019.

[12] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW 00, page 241250, New York, NY, USA, 2000. Association for Computing Machinery.

[13] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[14] 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*, KDD 16, page 11351144, New York, NY, USA, 2016. Association for Computing Machinery.

[15] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016.

[16] F. K. Doilovi, M. Bri, and N. Hlupi. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, 2018.

[17] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.

[18] Alejandro [Barredo Arrieta], Natalia Daz-Rodrguez, Javier [Del Ser], Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82 – 115, 2020.

[19] Surjo R. Soekadar, Niels Birbaumer, Marc W. Slutzky, and Leonardo G. Cohen. Brainmachine interfaces in neurorehabilitation of stroke. *Neurobiology of Disease*, 83:172 – 179, 2015.

[20] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Mller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[21] Yao Xie, Ge Gao, and Xiang 'Anthony' Chen. Outlining the design space of explainable intelligent systems for medical diagnosis. *CoRR*, abs/1902.06019, 2019.

[22] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 2019.

[23] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[24] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, and F. Marcelloni. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine*, 14(1):69–81, Feb 2019.

[25] K. Kallianos, J. Mongan, S. Antani, T. Henry, A. Taylor, J. Abuya, and M. Kohli. How far have we come?: Artificial intelligence for chest radiograph interpretation. *Clinical Radiology*, 74(5):338–345, May 2019.

[26] Grgoire Montavon, Wojciech Samek, and Klaus-Robert Mller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1 – 15, 2018.

[27] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017.

[28] Laura Rieger, Pattarawat Chormai, Grégoire Montavon, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, chapter Structuring Neural Networks for More Explainable Predictions, pages 115–131. Springer International Publishing, Cham, 2018.

[29] Sofia Meacham, Georgia Isaac, Detlef Nauck, and Botond Virginas. Towards explainable ai: Design and development for explanation of machine learning predictions for a patient readmittance medical application. In Kohei Arai, Rahul Bhatia, and Supriya Kapoor, editors, *Intelligent Computing*, pages 939–955, Cham, 2019. Springer International Publishing.

[30] J. Townsend, T. Chaton, and J. M. Monteiro. Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2019.

[31] Oct 2016. https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[32] Bert Heinrichs and Simon B. Eickhoff. Your evidence? machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*, 41(6):1435–1444, 2020.

[33] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Tho Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Sen higeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy ai development: Mechanisms for supporting verifiable claims, 2020.

[34] Nov 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[35] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI 19, New York, NY, USA, 2019. Association for Computing Machinery.

[36] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, 2017.

[37] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11), November 2017.

[38] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability, Jan 2020.

[39] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[40] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. *CoRR*, abs/1809.08037, 2018.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, June 2016.

[42] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[43] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang, and Min Xu. Respond-cam: Analyzing deep models for 3d imaging data by visualizations. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 485–492, Cham, 2018. Springer International Publishing.

[44] Sebastian Bach, Alexander Binder, Grgoire Montavon, Frederick Klauschen, Klaus-Robert Mller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[45] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Mller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[46] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *CoRR*, abs/1807.03418, 2018.

[47] A. W. Thomas, H. R. Heekeren, K. R. Müller, and W. Samek. Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. *Front Neurosci*, 13:1321, 2019.

[48] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "what is relevant in a text document?": An interpretable machine learning approach. *CoRR*, abs/1612.07843, 2016.

[49] V. Srinivasan, S. Lapuschkin, C. Hellge, K. Mller, and W. Samek. Interpretable human action recognition in compressed domain. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1692–1696, 2017.

[50] Oliver Eberle, Jochen Bttner, Florian Krutli, Klaus-Robert Mller, Matteo Valleriani, and Grgoire Montavon. Building and interpreting deep similarity models, 2020.

[51] Liam Hiley, Alun Preece, Yulia Hicks, Supriyo Chakraborty, Prudhvi Gurram, and Richard Tomsett. Explaining motion relevance for activity recognition in video deep learning models, 2020.

[52] Wojciech Samek, Grgoire Montavon, Alexander Binder, Sebastian Lapuschkin, and Klaus-Robert Mller. Interpreting the predictions of complex ml models by layer-wise relevance propagation, 2016.

[53] Machine learning and ai for the sciences - towards interpretability, 2018. http://www.heatmapping.org/slides/2018_WCCI.pdf.

[54] Deep taylor decomposition of neural networks, 2016. http://iphome.hhi.de/samek/pdf/MonICML16.pdf.

[55] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9277–9286. Curran Associates, Inc., 2019.

[56] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017.

[57] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *CoRR*, abs/1702.04595, 2017.

[58] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. *CoRR*, abs/1804.00880, 2018.

[59] Pieter-Jan Kindermans, Kristof T. Schtt, Maximilian Alber, Klaus-Robert Mller, Dumitru Erhan, Been Kim, and Sven Dhne. Learning how to explain neural networks: Patternnet and patternattribution, 2017.

[60] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Vigas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. https://pair-code.github.io/saliency/.

[61] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[62] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks, 2015.

[63] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 493–501, Cham, 2018. Springer International Publishing.

[64] Heather D. Couture, J. S. Marron, Charles M. Perou, Melissa A. Troester, and Marc Niethammer. Multiple instance learning forheterogeneous images: Training acnn for histopathology. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 254–262, Cham, 2018. Springer International Publishing.

[65] Xiaoxiao Li, Nicha C. Dvornek, Juntang Zhuang, Pamela Ventola, and James S. Duncan. Brain biomarker interpretation in asd using deep learning and fmri. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 206–214, Cham, 2018. Springer International Publishing.

[66] Yao Qin, Konstantinos Kamnitsas, Siddharth Ancha, Jay Nanavati, Garrison W. Cottrell, Antonio Criminisi, and Aditya V. Nori. Autofocus layer for semantic segmentation. *CoRR*, abs/1805.08403, 2018.

[67] Ziqi Tang, Kangway V. Chuang, Charles DeCarli, Lee-Way Jin, Laurel Beckett, Michael J. Keiser, and Brittany N. Dugger. Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature Communications*, 10(1):2173, 2019.

[68] Zachary Papanastasopoulos, Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Chintana Paramagul, Mark A. Helvie M.D., and Colleen H. Neal M.D. Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In Horst K. Hahn and Maciej A. Mazurowski, editors, *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 228 – 235. International Society for Optics and Photonics, SPIE, 2020.

[69] Hyebin Lee, Seong Tae Kim, and Yong Man Ro. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In Kenji Suzuki, Mauricio Reyes, Tanveer Syeda-Mahmood, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 21–29, Cham, 2019. Springer International Publishing.

[70] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham, 2019.

[71] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[72] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *CoRR*, abs/1412.0035, 2014.

[73] Alexey Dosovitskiy and Thomas Brox. Inverting convolutional networks with convolutional networks. *CoRR*, abs/1506.02753, 2015.

[74] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2014.

[75] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009. Also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.

[76] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *CoRR*, abs/1602.03616, 2016.

[77] Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.

[78] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[79] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *CoRR*, abs/1901.08644, 2019.

[80] Richard Meyes, Constantin Waubert de Puiseau, Andres Posada-Moreno, and Tobias Meisen. Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations, 2020.

[81] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 15, page 17211730, New York, NY, USA, 2015. Association for Computing Machinery.

[82] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):13501371, Sep 2015.

[83] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *CoRR*, abs/1902.00006, 2019.

[84] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES 19, page 131138, New York, NY, USA, 2019. Association for Computing Machinery.

[85] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics.

[86] Pei Guo, Connor Anderson, Kolten Pearson, and Ryan Farrell. Neural network interpretation via fine grained textual summarization. *CoRR*, abs/1805.08969, 2018.

[87] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1):431, May 2017.

[88] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS16, page 289297, Red Hook, NY, USA, 2016. Curran Associates Inc.

[89] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[90] Mohammadhassan Izadyyazdanabadi, Evgenii Belykh, Claudio Cavallo, Xiaochun Zhao, Sirin Gandhi, Leandro Borba Moreira, Jennifer Eschbacher, Peter Nakaji, Mark C. Preul, and Yezhou Yang. Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 300–308, Cham, 2018. Springer International Publishing.

[91] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014.

[92] Peifei Zhu and Masahiro Ogino. Guideline-based additive explanation for computer-aided diagnosis of lung nodules. In Kenji Suzuki, Mauricio Reyes, Tanveer Syeda-Mahmood, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 39–47, Cham, 2019. Springer International Publishing.

[93] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML17, page 33193328. JMLR.org, 2017.

[94] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile, 2017.

[95] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda B. Vigas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Jennifer G. Dy and Andreas Krause, editors, *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 2673–2682. JMLR.org, 2018.

[96] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS17, page 60786087, Red Hook, NY, USA, 2017. Curran Associates Inc.

[97] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.

[98] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

[99] Erdem Varol, Aristeidis Sotiras, Ke Zeng, and Christos Davatzikos. Generative discriminative models for multivariate inference and statistical mapping in medical imaging. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 540–548, Cham, 2018. Springer International Publishing.

[100] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016.

[101] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statist. Sci.*, 1(3):297–310, 08 1986.

[102] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 12, page 150158, New York, NY, USA, 2012. Association for Computing Machinery.

[103] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 13, page 623631, New York, NY, USA, 2013. Association for Computing Machinery.

[104] Sercan Ömer Arik and Tomas Pfister. Attention-based prototypical learning towards interpretable, confident and robust deep neural networks. *CoRR*, abs/1902.06292, 2019.

[105] Nima Hatami, Michaël Sdika, and Hélène Ratiney. Magnetic resonance spectroscopy quantification using deep learning. *CoRR*, abs/1806.07237, 2018.

[106] Stefan Haufe, Frank Meinecke, Kai Grgen, Sven Dhne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biemann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96 – 110, 2014.

[107] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1):13890, 2017.

[108] Kristof T. Schütt, Michael Gastegger, Alexandre Tkatchenko, and Klaus-Robert Müller. *Quantum-Chemical Insights from Interpretable Atomistic Neural Networks*, pages 311–330. Springer International Publishing, Cham, 2019.

[109] Christos Liaskos, Ageliki Tsioliaridou, Shuai Nie, Andreas Pitsillides, Sotiris Ioannidis, and Ian F. Akyildiz. An interpretable neural network for configuring programmable wireless environments. *CoRR*, abs/1905.02495, 2019.

[110] Barnabás Bede. Fuzzy systems with sigmoid-based membership functions as interpretable neural networks. In Ralph Baker Kearfott, Ildar Batyrshin, Marek Reformat, Martine Ceberio, and Vladik Kreinovich, editors, *Fuzzy Techniques: Theory and Applications*, pages 157–166, Cham, 2019. Springer International Publishing.

[111] Markus Kaiser, Clemens Otte, Thomas A. Runkler, and Carl Henrik Ek. Interpretable dynamics models for data-efficient reinforcement learning. *CoRR*, abs/1907.04902, 2019.

[112] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[113] Anil Hazarika, Mausumi Barthakur, Lachit Dutta, and Manabendra Bhuyan. F-svd based algorithm for variability and stability measurement of bio-signals, feature extraction and fusion for pattern recognition. *Biomedical Signal Processing and Control*, 47:26 – 40, 2019.

[114] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[115] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML17, page 214223. JMLR.org, 2017.

[116] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[117] Y. Zhu, S. Suri, P. Kulkarni, Y. Chen, J. Duan, and C. . J. Kuo. An interpretable generative model for handwritten digits synthesis. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1910–1914, 2019.

[118] Ryen Krusinga, Sohil Shah, Matthias Zwicker, Tom Goldstein, and David W. Jacobs. Understanding the (un)interpretability of natural image distributions using generative models. *CoRR*, abs/1901.01499, 2019.

[119] A. Karpathy. t-sne visualization of cnn codes, 2014. https://cs.stanford.edu/people/karpathy/cnnembed.

[120] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Exploring neural networks with activation atlases, 2019. https://distill.pub/2019/activation-atlas.

[121] Wei Ma, Feng Cheng, Yihao Xu, Qinlong Wen, and Yongmin Liu. Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy. *Advanced Materials*, 31(35), 2019.

[122] Yujun Yan, Jiong Zhu, Marlena Duda, Eric Solarz, Chandra Sripada, and Danai Koutra. Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 19, page 772782, New York, NY, USA, 2019. Association for Computing Machinery.

[123] Carlo Biffi, Ozan Oktay, Giacomo Tarroni, Wenjia Bai, Antonio M. Simoes Monteiro de Marvao, Georgia Doumou, Martin Rajchl, Reem Bedair, Sanjay K. Prasad, Stuart A. Cook, Declan P. O'Regan, and

Daniel Rueckert. Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. *CoRR*, abs/1807.06843, 2018.

[124] Mingliang Wang, Daoqiang Zhang, Jiashuang Huang, Dinggang Shen, and Mingxia Liu. Low-rank representation for multi-center autism spectrum disorder identification. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 647–654, Cham, 2018. Springer International Publishing.

[125] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *CoRR*, abs/1704.03296, 2017.

[126] David Alvarez-Melis and Tommi Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[127] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:18031831, August 2010.

[128] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML17, page 18851894. JMLR.org, 2017.

[129] Chih-Kuan Yeh, Joon Sik Kim, Ian E.H. Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS18, page 93119321, Red Hook, NY, USA, 2018. Curran Associates Inc.

[130] Cagdas Ulas, Giles Tetteh, Stephan Kaczmarz, Christine Preibisch, and Bjoern H. Menze. Deepasl: Kinetic model incorporated loss for denoising arterial spin labeled mri via deep residual learning. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 30–38, Cham, 2018. Springer International Publishing.

[131] C. J. Scott, J. Jiao, A. Melbourne, N. Burgos, D. M. Cash, E. De Vita, P. J. Markiewicz, A. O'Connor, D. L. Thomas, P. S. Weston, J. M. Schott, B. F. Hutton, and S. Ourselin. Reduced acquisition time PET pharmacokinetic modelling using simultaneous ASL-MRI: proof of concept. *J. Cereb. Blood Flow Metab.*, 39(12):2419–2432, Dec 2019.

[132] 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. In Kenji Suzuki, Mauricio Reyes, Tanveer Syeda-Mahmood, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 30–38, Cham, 2019. Springer International Publishing.

[133] Feiyun Zhu, Jun Guo, Zheng Xu, Peng Liao, Liu Yang, and Junzhou Huang. Group-driven reinforcement learning for personalized mhealth intervention. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 590–598, Cham, 2018. Springer International Publishing.

[134] Ozan Kocadagli and Reza Langari. Classification of eeg signals for epileptic seizures using hybrid artificial neural networks based wavelet transforms and fuzzy relations. *Expert Systems with Applications*, 88:419 – 434, 2017.

[135] Tao Zhang, Wanzhong Chen, and Mingyang Li. Classification of inter-ictal and ictal eegs using multi-basis modwpt, dimensionality reduction algorithms and ls-svm: A comparative study. *Biomedical Signal Processing and Control*, 47:240 – 251, 2019.

[136] Wentian Li, Jane E. Cerise, Yaning Yang, and Henry Han. Application of t-sne to human genetic data. *Journal of Bioinformatics and Computational Biology*, 15(04):1750017, 2017. PMID: 28718343.

[137] W. Xu, X. Jiang, X. Hu, and G. Li. Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization. *BMC Med Genomics*, 7 Suppl 2:S1, 2014.

[138] Noel C. F. Codella, Chung-Ching Lin, Allan Halpern, Michael Hind, Rogerio Feris, and John R. Smith. Collaborative human-ai (chai): Evidence-based interpretable melanoma classification in dermoscopic images. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva, Sergio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 97–105, Cham, 2018. Springer International Publishing.

[139] Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression concept vectors for bidirectional explanations in histopathology. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva, Sergio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132, Cham, 2018. Springer International Publishing.

[140] Hugo Yeche, Justin Harrison, and Tess Berthier. Ubs: A dimension-agnostic metric for concept vector interpretability applied to radiomics. In Kenji Suzuki, Mauricio Reyes, Tanveer Syeda-Mahmood, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 12–20, Cham, 2019. Springer International Publishing.

[141] Alvaro E. Ulloa Cerna, Marios Pattichis, David P. vanMaanen, Linyuan Jing, Aalpen A. Patel, Joshua V. Stough, Christopher M. Haggerty, and Brandon K. Fornwalt. Interpretable neural networks for predicting mortality risk using multi-modal electronic health records. *CoRR*, abs/1901.08125, 2019.

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

[143] Youngwon Choi, Yongchan Kwon, Hanbyul Lee, Beom Joon Kim, Myunghee Cho Paik, and Joong-Ho Won. Ensemble of deep convolutional neural networks for prognosis of ischemic stroke. In Alessandro Crimi, Bjoern Menze, Oskar Maier, Mauricio Reyes, Stefan Winzeck, and Heinz Handels, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 231–243, Cham, 2016. Springer International Publishing.

[144] Oskar Maier and Heinz Handels. Predicting stroke lesion and clinical outcome with random forests. In Alessandro Crimi, Bjoern Menze, Oskar Maier, Mauricio Reyes, Stefan Winzeck, and Heinz Handels, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 219–230, Cham, 2016. Springer International Publishing.

[145] Been Kim, Caleb M. Chacha, and Julie Shah. Inferring robot task plans from human team meetings: A generative modeling approach with logic-based prior. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI13, page 13941400. AAAI Press, 2013.

[146] Justin Cheng and Michael S. Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '15, page 600611, New York, NY, USA, 2015. Association for Computing Machinery.

[147] L. Kuhlmann, P. Karoly, D. R. Freestone, B. H. Brinkmann, A. Temko, A. Barachant, F. Li, G. Titericz, B. W. Lang, D. Lavery, K. Roman, D. Broadhead, S. Dobson, G. Jones, Q. Tang, I. Ivanenko, O. Panichev, T. Proix, M. N?hl?k, D. B. Grunberg, C. Reuben, G. Worrell, B. Litt, D. T. J. Liley, D. B. Grayden, and M. J. Cook. Epilepsyecosystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG. *Brain*, 141(9):2619–2630, 09 2018.

[148] M. Wiener, F. T. Sommer, Z. G. Ives, R. A. Poldrack, and B. Litt. Enabling an Open Data Ecosystem for the Neurosciences. *Neuron*, 92(4):929, 11 2016.

[149] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*, 2(4):230–243, Dec 2017.

[150] C. K. Cassel and A. L. Jameton. Dementia in the elderly: an analysis of medical responsibility. *Ann. Intern. Med.*, 94(6):802–807, Jun 1981.

[151] Pat Croskerry, Karen Cosby, Mark L. Graber, and Hardeep Singh. Diagnosis : Interpreting the shadows., 2017.

[152] Sérgio Pereira, Raphael Meier, Victor Alves, Mauricio Reyes, and Carlos A. Silva. Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva,

Sergio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 106–114, Cham, 2018. Springer International Publishing.

[153] A. Vilamala, K. H. Madsen, and L. K. Hansen. Deep convolutional neural networks for interpretable analysis of eeg sleep stage scoring. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2017.

[154] Pieter Van Molle, Miguel De Strooper, Tim Verbelen, Bert Vankeirsbilck, Pieter Simoens, and Bart Dhoedt. Visualizing convolutional neural networks to improve decision support for skin lesion classification. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva, Sergio Pereira, and Raphael Meier, editors, *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 115–123, Cham, 2018. Springer International Publishing.

[155] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, and C. Pattichis. Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 817–821, 2019.

[156] H. Sun, L. Paixao, J. T. Oliva, B. Goparaju, D. Z. Carvalho, K. G. van Leeuwen, O. Akeju, R. J. Thomas, S. S. Cash, M. T. Bianchi, and M. B. Westover. Brain age from the electroencephalogram of sleep. *Neurobiol. Aging*, 74:112–120, 02 2019.

[157] Fabian Eitel and Kerstin Ritter. Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification. In Kenji Suzuki, Mauricio Reyes, Tanveer Syeda-Mahmood, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2019. Springer International Publishing.

[158] Christoph Jansen, Thomas Penzel, Stephan Hodel, Stefanie Breuer, Martin Spott, and Dagmar Krefting. Network physiology in insomnia patients: Assessment of relevant changes in network topology with interpretable machine learning models. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):123129, 2019.

[159] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. Deep neural network or dermatologist? In Kenji Suzuki, Mauricio Reyes, Tanveer Syeda-Mahmood, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 48–55, Cham, 2019. Springer International Publishing.

[160] C. Zucco, H. Liang, G. D. Fatta, and M. Cannataro. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1740–1747, 2018.

[161] Curtis P. Langlotz, Bibb Allen, Bradley J. Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S. Cook, Adam E. Flanders, Matthew P. Lungren, David S. Mendelson, Jeffrey D. Rudie, Ge Wang, and Krishna Kandarpa. A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 291(3):781–791, 2019. PMID: 30990384.

[162] Arieh Gomolin, Elena Netchiporouk, Robert Gniadecki, and Ivan V. Litvinov. Artificial intelligence applications in dermatology: Where do we stand? *Frontiers in Medicine*, 7:100, 2020.

[163] A. J. London. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep*, 49(1):15–21, Jan 2019.

[164] Sadid Hasan, Yuan Ling, Dimeji Farri, Joey Liu, Henning Mller, and Matthew Lungren. Overview of imageclef 2018 medical domain visual question answering task. 09 2018.

[165] Asma Ben Abacha, Sadid Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Mller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *Lecture Notes in Computer Science*, 09 2019.

[166] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13589–13600. Curran Associates, Inc., 2019.

[167] Himabindu Lakkaraju and Osbert Bastani. how do i fool you?. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Feb 2020.

[168] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Gregory S. Corrado, Jason D. Hipp, Lily Peng, and Martin C. Stumpe. Detecting cancer metastases on gigapixel pathology images. *CoRR*, abs/1703.02442, 2017.