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

Erico Tjoa and Cuntai Guan Fellow, IEEE

Abstract—**Recently, artificial intelligence, especially machine learning has demonstrated remarkable performances in many tasks, from image processing to natural language processing, especially with the advent of deep learning. Along with research progress, machine learning has encroached into many different fields and disciplines. Some of them, such as the medical field, require high level of accountability, and thus transparency, which means we need to be able to explain machine decisions, predictions and justify their reliability. This requires greater interpretability, which often means we need to understand the mechanism underlying the algorithms. Unfortunately, the black-box 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, with the intention of providing alternative perspective that is hopefully more tractable for future adoption of interpretability standard. We explore further into interpretability in the medical field, illustrating the complexity of interpretability issue.**

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

1. 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] [2] [3] [4] [5] [6] [7] [8].

Interpretability and explainability of a ML algorithm 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 on 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. 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 [9] [10] [11].

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 the interpretability in the medical field. We will not attempt to cover all the related works many of which are already presented in the research and survey works we cite [1] [2] [9] [12] [13] [14] [15] [16] [17] [18] [19]. This paper is arranged as the following. Section 2 explores the types of interpretability. Section 3 lists some concepts that frequently occur in interpretability studies. Section 4 examines a proposed category for interpretabilities. Section 5 extends the previous sections into the medical field. Lastly, it is also imperative to point out that the issue of accountability and interpretability has even entered the sphere of ethics and law enforcements [20], engendering movements to protect the society from possible misuses and harms in the wake of the increasing use of AI.

2. 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 [9] [21] [22]. In fact, different works use different criteria, and they are justifiable in one way or another. [23] [24] have suggested a unified framework to study interpretabilities that have thus-far been studied separately. [25] defines a unified measure of feature importance. Here, we categorize existing interpretabilities and present non-exhaustive list of works in each category.

2.1. Visual and Textual Explanations as Interpretability

The paper [26] aptly summarizes interpretability issue within ML research with the phrase "Why Should I Trust You?". It introduces LIME algorithm (Local Interpretable Model-agnostic Explanations) that explains the output of other models visually or textually. In this vein, it suggests that an interpretable model is one that provides a humanly easy-to-understand explanations. For example, an interpretable classifier will distinguish an elephant from a yellow-striped cat by explaining its choice with words such as "grey" and "trunk". In text classification problems, a model will distinguish a food-related article from a sports news by providing an importance chart highlighting words such as "delicious" and "tasty".

In image processing, saliency maps, heat maps and super-pixels may be the natural way towards interpretability. For example, in an object detection problem, the detection of a dog in an image can be explained by providing an image obtained from element-wise multiplication of the image with its super-pixel map. The output, i.e. the explanation, is the same image with the dog left intact and greyed out region elsewhere. More generally, if $x \in \mathbb{R}^{m\times n}$ is the original representation having $m \times n$ pixels, then the super-pixel map $x' \in \mathbb{R}^{m\times n} \in \{0,1\}^{m\times n}$ indicates the presence of some human-interpretable feature around some patches in the image, labeling the pixels in the patches with 1 (and labelling other patches with 0). The explanation is then the element-wise product $x * x'$. [22] [27] both use Layer-wise Relevance Propagation (LRP) to construct heat-maps for interpretability. Recently developed Automatic Concept-based Explanations (ACE) algorithm [28] also uses super-pixels as explanations.

Class Activation Map (CAM) has been introduced as a kind of saliency-map that corresponds to discriminative features for classifications [29] [30] [31]. In an image of a person with a dog, to explain the detection the dog, localization is performed by computing a heatmap showing high intensity around the image patches where the dog is. This has been shown to correspond to a reasonably high accuracy of object classification. Note that, so far, there is neither explanation nor interpretation of the inner working of an ML algorithm.

A feature map is the output of the convolution of a filter with the output from a previous layer in a convolutional neural network (CNN). [32] demonstrates that feature maps in the deeper layer activate more strongly to complex features, such as human face, keyboard etc. Conversely, it is observed that earlier layers activate more strongly to simple features such as edges, vertical lines etc. Methods of interpretability that observe the stimulation of neuron layers like this are also called *signal methods* [33] (although this paper specifically refers to the observation of neuron layers at higher layers).

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 [32]. However, they might be too distorted to be acceptable. Then, how else can a feature map be related to a humanly visible 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 [32] reconstructs an image using a deconvolution network by approximately inversing the trained convolutional network $\tilde{x} = deconv(y) = \hat{f}_2^{-1}\tilde{f}_1^{-1}(y)$ 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.

Feature visualization by optimization is described by [23] with an excellent interactive interface. In essence, choose a neuron (or a set of them) from a neural network (NN), then an input image is optimized so as to maximize or minimize the activation of the neuron. Starting with a noise as an input, the optimized input that maximizes the activation of a neuron can emerge as something visually interpretable. For example, the image could be a surreal fuzzy combination of swirling patterns and parts of dog faces.

In [23] , GoogLeNet [34] is used, which has the advantage of having deep layers. Images that optimize neuron activations at deeper layer turn out to be interesting as well, some being combinations of visibly distinct parts of cat faces with cars, suggesting that each neuron might not always correspond to concepts that are similar with respect to human understanding. Unfortunately, many high frequencies or intricate patterns are still hard to explain.

To bring this one step further, [24] introduces the "semantic dictionary". Given an image of a cat, pick a small patch, for example, corresponding to the paw of the cat. Observe the activation intensities $a_1, a_2, ...$ of several corresponding neurons $n_1, n_2, ...$ generally from different layers in decreasing intensity of activation. Using the feature visualization by optimization explained before, each neuron corresponds to an optimized input $x_1, x_2, ...$ and $\{x_i, i = 1, ..., n\}$ forms the semantic dictionary. Then, we can say that the paw corresponds to $a_1 \times x_1 + a_2 \times x_2 + \cdots$. The good news is, if x_1 is a somewhat clear image, for example, a surreal combinations of paws and claws, then we can at least say the small patch we chose from the cat image has a_1 degree of "paw-ness". The bad news is, say, if x_2 is some undecipherable swirls of patterns, we have nothing much to interpret.

In the medical field (see later section), [31] [35] [36] [37] [38] have studied methods employing visual explanations.

2.2. Logical Statements as Interpretability

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 $A \rightarrow B$, in another words "*if A then B*". An ML model from which logical statements can be extracted directly has been considered obviously interpretable. In [39], a rule-based system could provide the statement "*has asthma→lower risk*", where risk here refers to death risk due to pneumonia. Likewise, [40] creates a model that provides such statements for stroke prediction.

One can indeed question the interpretability there. 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 "*asthma→ ICU→ lower risk*". In the

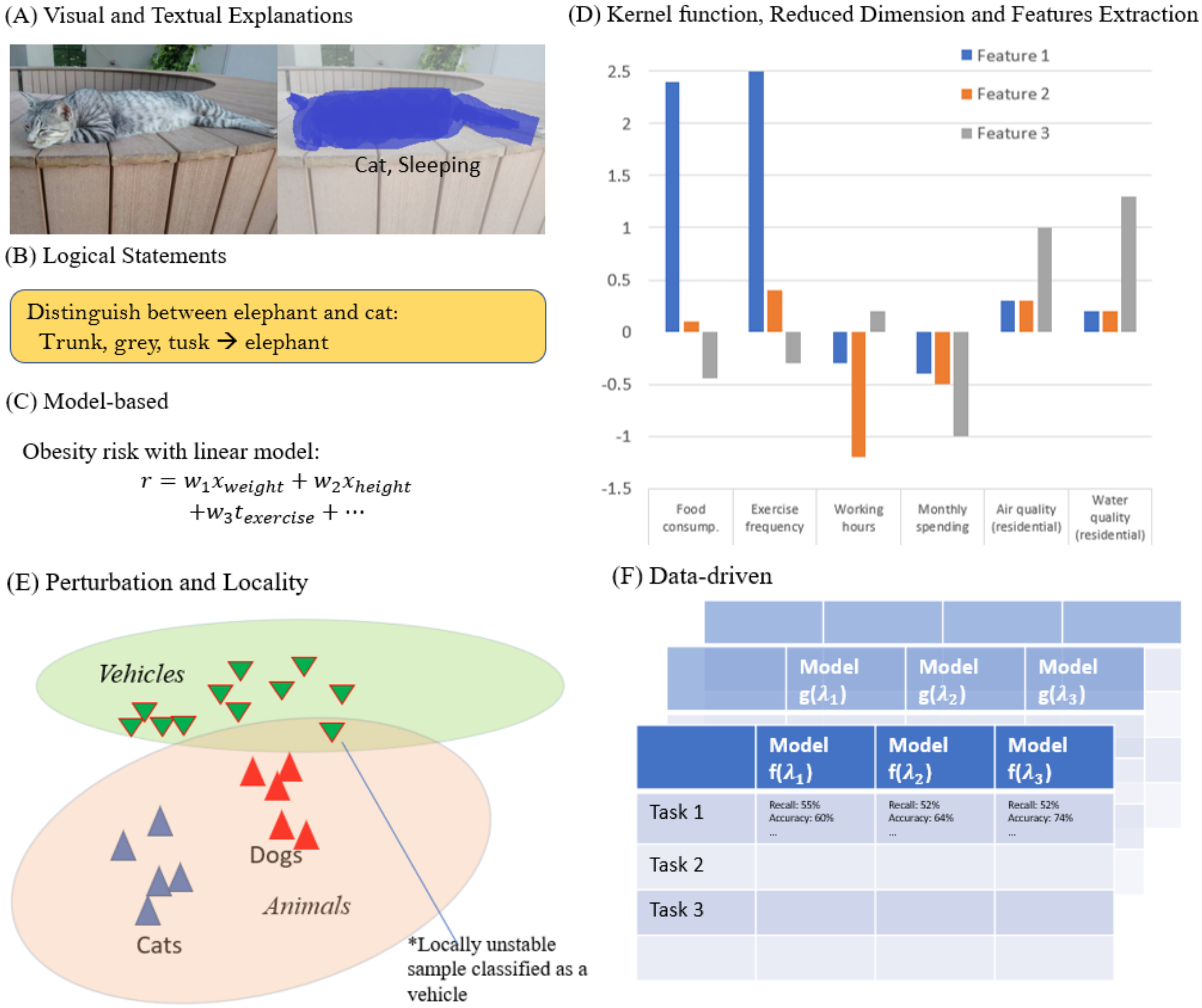
(A) Visual and Textual Explanations
Cat, Sleeping
(B) Logical Statements
Distinguish between elephant and cat:
Trunk, grey, tusk → elephant
(C) Model-based
Obesity risk with linear model:
$r = w_1 x_{weight} + w_2 x_{height} + w_3 t_{exercise} + \cdots$
(D) Kernel function, Reduced Dimension and Features Extraction
Feature 1
Feature 2
Feature 3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
Food consump.
Exercise frequency
Working hours
Monthly spending
Air quality (residential)
Water quality (residential)
(E) Perturbation and Locality
Vehicles
Dogs
Animals
Cats
*Locally unstable sample classified as a vehicle
(F) Data-driven
Model g(λ_1)
Model g(λ_2)
Model g(λ_3)
Model f(λ_1)
Model f(λ_2)
Model f(λ_3)
Task 1
Recall: 55% Accuracy: 60% ...
Recall: 52% Accuracy: 64% ...
Recall: 52% Accuracy: 74% ...
Task 2
Task 3

Figure 1. Types of interpretability and an example in each category. (A) A heatmap provides the intermediate "thought processes" of an algorithm that can be used to justify image classification. (B) A model that explicitly provides the logical statement leading to some conclusion is sometimes considered obviously interpretable. (C) A parametric model can be pre-specified. Variables are then interpreted with respect to the equation and how the output y might vary with them. (D) Features can be extracted in the form of multi-dimensional vectors with meaningful weightage of variables. For example, blue feature in the figure corresponds to the effect of healthy-lifestyle on the output variable. (E) Decisions can be interpreted as the amount of confidence a model groups an object into different categories. (F) Typically, large collection of data is fed into algorithm to minimize loss function. This example, however, shows tabulation of vast data collection across different dimensions and different models which has been suggested as a means to achieve interpretability.

paper, the model appears not to do so. We can see that interpretability issues are not always clear-cut.

Similarly, decision sets or rule sets have been studied for interpretability [41]. A single line in a rule set is for example " rainy and gumpy or calm → dairy or vegetables ". 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 $(\text{rainy} \wedge \text{gumpy}) \vee \text{calm} \rightarrow \text{dairy} \wedge \text{veget}ables$. Comparing three different variables, [41] finds out 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).

2.3. Model-based Interpretability

A problem can be put into the framework of a model, and the explanation can then be extracted from the typically parametric model. In the medical field (see later section), kinetic modelling is a popular choice, and machine learning can be used to compute the parameters. Other methods exist, for example, [42] utilizes regression to create interpretable solution indirectly.

A good regression model will provide a relation that reveals a trend. For example, $y = f(\vec{x}) = f(x_1, \dots, x_i, \dots)$ can provide a trend like "*y increases linearly with x_i but does not depend on x_j*". In other cases, the explanation can be in the form of more general logical statement "if $P[\vec{x}]$ then $Q[y(\vec{x})]$", where P, Q are the appropriate predicates. As in the previous section, such statements have sometimes been taken as obviously interpretable. Note that the dependence on the variables $\{x_i\}$ might be hard to interpret if $y(.)$ turns out to be a very complex function.

In [39], 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. This idea is represented by the Generalized Additive Model (GAM) [43] [44] with standard form

$$g(E[y]) = \beta_0 + \Sigma f_j(x_j)$$

where g is the link function. The familiar General Linear Model (GLM) is GAM with linear f_j. Besides, as a natural extension to the model, interaction terms like $f_{ij}(x_{ij})$ can be used as well [45].

2.4. Kernel function, Reduced Dimension and Features Extraction for Interpretability

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) [46].

How does a kernel function help with interpretability? 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. The kernel function of a model finds 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 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.

Recently, Singular Vector Canonical Correlation Analysis (SVCCA) is suggested as a tool to analyze interpretability [47]. Given an input dataset $X = \{x_1, \dots, x_m\}$ where each input x_i is possibly multi-dimensional. Denote the activation of neuron i at layer l as $z_i^l = \left(z_i^l(x_1), \dots, z_i^l(x_m)\right)$. Note that one such output is defined for the entire input dataset. SVCCA finds out the relation between 2 layers of a network $l_k = \{z_i^{l_k} | i = 1, \dots, m_k\}$ 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 $\bar{l}'_2 = W_Y l'_2$ have the maximum correlation $\rho = \{\rho_1, \dots, \rho_{\min(m_1,m_2)}\}$. 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 $\bar{\rho} = \frac{1}{\min(m_1,m_2)} \Sigma_i \rho_i$.

Testing with Concept Activation Vectors (TCAV) has also been introduced as a technique to interpret the low-level representation of neural network layer [48]. Given input $x \in \mathbb{R}^n$ and a feedforward layer l having m neurons, then the activation at that layer can be 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 maps out a set of activations that has relevant information about the "stripe"-ness. And then CAV is defined as the normal vector to the hyperplane that separates the positive examples from the negative ones. ACE algorithm [28] uses TCAV to compute saliency score and generate super-pixels as explanations.

Algorithm such as t-SNE has been used to cluster input images based on their activation of neurons in a network [49] [50]. In [49], 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). [51] introduces activation atlases, which is similarly using t-SNE to arrange some activations $\{f_{act}(x)\}$, except that each point is represented by the average activations of feature visualization.

In the medical field (see later section), [52] uses Laplacian Eigenmap for interpretability; [53] introduces a low-rank representation method for Autistic Spectrum Diagnosis.

2.5. Locality and Perturbation-based interpretability

To interpret a model, we can perform post-processing. For example, a customized probe can be designed specific to a model to observe activation of neurons in a trained layer. Perturbative methods include observation of output z' for an input x', where x' is in the locality of some input x with output z. This abstraction may make sense in terms of interpretability if we imagine a trained model as a highly multi-dimensional space. The points in the space form a whole space (possibly a continuum) where each patch or locality corresponds to a notion, concept, or object as shown in figure 1(E). In this sense, an interpretable model groups together similar objects, or objects that are "near" with respect to some metric.

The difficulty yet to be resolved is how this space can be represented concisely and accurately. The interpretability of a model might suffer if x' is similar to x, but somehow z' produces a radically different output. Such cases have been used by [54] to generate adversarial training data to further train the model.

In the section on *locality* later, we see that LIME from [26] is an optimization problem using $argmin[L + \Omega]$. The sampling method used to find an interpretable model, for example a linear $g \in G$, is perturbative because given in input x' to the model, its perturbed counterparts z' which contains some of the non-zero elements of x' are sampled (uniform randomly). See a section later on *Locality, Sensitivity, Gradients* for more elaboration of the model.

TCAV is claimed to be a global-perturbation-based tool for interpretability [48]. It uses directional derivative $S_{v,k,l}(x) = \nabla h_{l,k}\big(f_l(x)\big) \cdot v_C^l$ where $h_{l,k}$ is the logit function for class k of C for layer l, TCAV computes the score $TCAV_{Q_{C,k,l}} = \frac{|\{x \in X_k : S_{C,k,l}(x) > 0\}|}{|X_k|} \in [0,1]$. A two-sided t-test can be used where null hypothesis, TCAV=0.5, means there is no prediction related to the classification w.r.t concept C, i.e. not interpretable where concept C is concerned. From our understanding, it 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. Furthermore, perturbation is, by definition, local. More related works are explained in the section *Locality, Sensitivity, Gradients*.

2.6. Data-driven interpretability

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. In essence, [9] 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*. A row in the matrix can be, for example, identifying pneumonia patients and the columns can correspond to decision trees of increasing depths. How can we gather a large collection of entries in such a large matrix? Apart from competitions and challenges, crowd-sourcing efforts will aid the formation of such database [55] [56]. 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 task- and method-related common criteria, the so-called latent dimensions, 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 [41]) etc. These dimensions are to be refined along iterative processes as more user-inputs enter the repository.

The paper begins by posing the problem of *incompleteness* of problem formulation as an 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.

3. CONCEPTS ASSOCIATED WITH INTERPRETABILITY

[9] points out that a major way to understand interpretability is through a proxy. Depending on the system, a metric measures some property such as sparsity, and an interpretable model will be one that attains some value assigned by the metric. We explore properties that might be quantified as proxies in this section.

3.1. Locality, Sensitivity, Gradients

There is a concern of locality vs globality in a model. Let a model $f(.)$ predicts $f(x)$ accurately for some x. Denote $\tilde{x}$ as a slightly noisy version of x. The model is a locally faithful $f(\tilde{x})$ produces correct prediction, otherwise, the model is unfaithful and clearly such instability reduces its reliability. In another words, the function is able to identify a locally important feature. There is no guarantee there exists globally important feature that is invariant in a machine learning task.

[57] defines gradient-based explanation vector $\zeta(x_0) = \frac{\partial}{\partial x} P(Y \neq g(x_0)|X = x)_{|x=x_0}$ for Bayesian classifier $g(x) = \underbrace{argmin}_{c\in\{1,\ldots C\}} P(Y \neq c|X = x)$ where x, ζ are d-dimensional. For any $i = 1, \ldots, d$, high absolute value of $[\zeta(x_0)]_i$ 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 $g(x_0)$. The paper also suggests a different definition.

[26] shows that at least locally, fidelity can be achieved. Suppose a binary problem is perfectly captured by $f: \mathbb{R}^d \to \mathbb{R}$ with $f(x) \in [-1,1]$ and suppose there is a model $g \in G$ that is trained to estimate f and g contains explanations either textual or visual forms can be presented directly to user that. Some examples of possible g include linear models and decision trees. Let $\pi_x(z)$ be the "distance" between x and z, and $L(f, g, \pi_x)$ be the measure of unfaithfulness. As an example, $L(f, g, \pi_x) = \Sigma_{z,z' \in Z} \pi_x(z)[f(z) - g(z')]^2$ and $\pi_x(z) = e^{-D(x,z)^2/\sigma^2}$ where $D(x, z)$ can be Euclidean distance or similar variants, and σ some constant. It is clear that the closer f is to g at some locality of x, the smaller is the contribution of x to unfaithfulness. Furthermore, define $\Omega(x)$ as the complexity of the classifier. As a simple example, Ω can quantify the depth of decision tree classifier; if x can be computed at a shallow level of the tree, then there will be a few features that a human can easily extract. With fewer features, it is more humanly readable, hence more explainable. Then, an optimally interpretable model is

$$g_0 = \underbrace{argmin[L(f, g, \pi_x) + \Omega(g)]}_{g\in G}$$

If we denote LIME as the mapping ξ from a locality around x to a model $g \in G$, then $\xi(x) = g_0$, i.e. LIME computes a model that tries to output value $g_0(z)$, where z is near x, that estimates the best possible $f(z)$ while keeping the interpretability high by keeping the complexity low.

[48] categorizes its TCAV (see the section on perturbation) as A global method, though it should be taken with a pinch of salt due to the computation of their metrics that strictly partition the labels beforehand. [25] identifies local accuracy as an important property for models that use additive feature attributions (see next section on linearity).

3.2. Linearity

The simplest interpretable model is the linear combination of variables $y = \Sigma_i a_i x_i$ where a_i is the degree of x_i-ness of the prediction y. If the model performs well, this can be considered highly interpretable. However, in other cases, while linearity might not be directly associated with interpretability, studying interpretability via linear properties are useful in several ways, including the ease of implementation. When non-linearity is required, it is typically not difficult to replace the linear function $\vec{w}\cdot\vec{a}$ within the system with a non-linear version $f(w_1a_1,\dots,w_na_n)$. [25] refers to a linear combination method with $x_i\in\{0,1\}$ as the additive feature attribution method.

A linear probe is used in [58] to extract information from each layer in a neural network. More technically, assume we have deep learning classifier $F(x)\in\mathbb{R}^D$ where $\mathrm{F_i(x)}\in[0,1]$ 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 classifiers generally perform better at deeper layer, that is, at larger k.

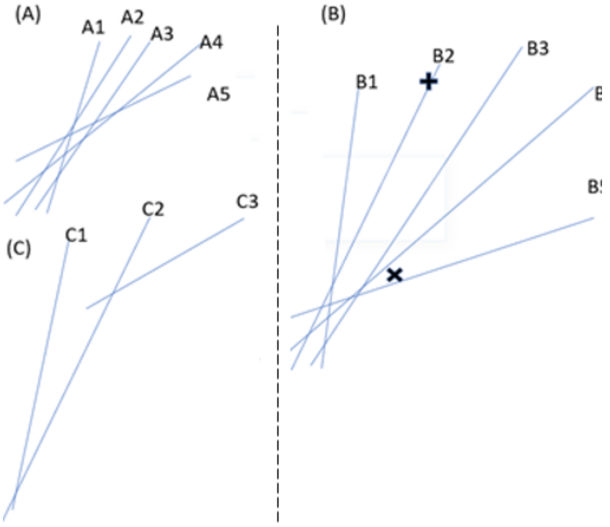

Figure 2. Illustration of linear separability. (A) Not highly linearly separable (B) Good linear separability. + mark shows an input classified as B2 with high confidence, while x mark corresponds to an input classified to either B4 or B5 since it might share both properties (C) Potential problematic situation.

In their words, linear separability is low at earlier layers (figure 2A) and improves deeper into layer (figure 2B). We can say that an ML model is interpretable because it learns to distinguish classes of objects more linearly deeper into the layers. Since any 2 straight-lines intersect at one point at most, the improvement in linear separability means that different classes are separated further in some D-dimensional space as shown by the + mark in figure 2B, while an object sharing the properties of some classes will turn out nearer the intersection points as shown by the x mark in figure 2B. This is visual enough to be easily understandable by a human (at least one with basic notion of linear algebra). The biggest problem we can procure on the fly is the possibility that the classes might not intersect around the same locality (figure 2C), which renders the interpretation of the probe possibly noisy.

In the medical field (see later section), [42] [59] suggest linear combination of variables as the means to interpretability (for example clinical variables, metabolites signals for MRS etc). [60] discusses the linearity in models used in the estimation of brain states, including how it is misinterpreted.

3.3. Invariances

Implementation invariance. [61] suggests implementation invariance as an axiomatic requirement. In the paper, it is stated as the following. Define two *functionally equivalent* functions as f_1, f_2 so that $f_1(x)=f_2(x)$ 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, $\left(A[f_1](x)\right)_j=\left(A[f_2](x)\right)_j$ for any $j=1,\dots,d$ 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. We use figure 1(A) as an illustration of input invariance: if we move the sleeping cat to the right, then the explanation provided by a model (in this case the blue super-pixels) will shift right correspondingly. Clearly, this property is desirable and has been proposed as an axiomatic invariance of a reliable saliency method. [33] studies the input invariance of some saliency methods with respect to translation of input $x\rightarrow x+c$ for some c. Of the methods studied, gradients/sensitivity-based methods [57] and signal methods [32] [62] are input invariant while some attribution methods, such as integrated gradient, do not [61].

3.4. Others

There are many other concepts that can be related to interpretability. [30] conducted tests on the improvements of human performance on a task after being given explanations (in the form of visualization) produced by machine learning 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 classfication tasks and attain a baseline 40% accuracies. 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.

4. Bases of interpretability evaluation

This section explores how researchers classify different interpretability classes. This is relevant because not all scenarios require the same explanations. For example, it might not be useful to provide words searched from a common dictionary if a medical imaging diagnosis depends on the location and size of a lesion. [10] specifically studies "What Clinicians Want", providing what qualifies as explainable for clinicians. We group some works based on the recommendation by [9] which suggests three different bases for evaluation (though we do not include most works it already cited). Fitting them into one of the following categories may or may not enhance their usefulness and interpretability, but we attempt to do so as part of an overview.

4.1. Application-grounded

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 [63] [64] (also see other categories of the grand challenge).

Such evaluation is performed, for example, by [30]. They proposed Grad-CAM applied on Guided Back-propagation (proposed by [62]) of AlexNet CNN and VGG, whose visualizations 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 Back-propagation.

4.2. Human-grounded

Second evaluation suggested by [9] is **human-grounded**. This evaluation involves real humans and simplified tasks. It can be used when, for some reasons or another, having human A gives 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 $\widehat{\mathrm{X}}_{\mathrm{A}}$ 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 [65], 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. In [66], hybrid human and interactive ML classifiers require human users to nominate features for training. 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.

4.3. Functionally-grounded

Third, an evaluation is **functionally-grounded** evaluation if there exist proxies (which can be defined a priori) for evaluation, for example sparsity [9]. This evaluation is appropriate when specific requirements are met (e.g. there is already a strong justification to use a specific ML model, and the remaining concern is which relevant proxy is best) or impossible to meet (e.g. unethical to use human). Some papers [2] [5] [29] [67] [31] [47] [68] [63] [64] 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. values from dimensionality reduction methods: interpretability is related to the degree an object relates to a feature, for example, classification of a dog has high values related to four limbs, snout and paws

and others.

5. XAI IN MEDICAL FIELD

ML has also gained traction recently in the medical field, with large volume of works on automated diagnosis, prognosis [69]. 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 black-box; it is not very interpretable. Other domain specific methods and special functions (denoising etc) have been published as well ([70] and many other works, for example 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 [14] [71] [72]. 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.

Many more works have thus been dedicated to exploring explainability in the medical fields [10] [13] [31], providing summaries of previous works [14] including subfield specific review [19] (for chest radiograph), or at least set aside a section to promote awareness for the importance of interpretability in the medical field [73].

5.1. Categorizing interpretabilities in Medical Field

Here, we consider the interpretabilities that have been discussed in the previous section but are also prevalent in the medical imaging field.

Visual interpretability - Visual interpretability applies to the medical field as well. However, the images in medical field comes in many different formats such as NIFTI or DCOM. They could come in traditional 2D images, 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, such as photographs of animals. Obtaining these data certainly requires some level of consents from the patients and other administrative barriers. High dimensional data also add complexity to data processing and the large RAM space requirement might prevent data to be input without modification, random sampling or down-sizing, which may compromise analysis.

When data is available, ground-truth images may not be "correct", in the sense that human can correctly identify different common animals relatively easily (see the risk of machine interpretation in a later section: *noisy training data*). Not only do these data require some specialized knowledge to understand, the lack of comprehensive understanding of how biological components such as the brain complicates the analysis. For example, 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 ML algorithm can capture the correct features, even if an ML can capture features in principle. All these considered, ML still holds great potentials for both reliability and interpretability.

[31] develops Grad-CAM which is derived from [29] [30], 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. [36] uses multi-instance (MI) aggregation method during pre-processing for the training of CNNs to classify breast tumour tissue microarray (TMA) images for 5 different tasks, 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. Likewise, see [37] [74].

The autofocus module from [38] promises improvements for CNN in terms of visual interpretability. It uses attention mechanism (proposed by [75]) 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.

Case-Based Reasoning (CBR) performs medical evaluation (classifications etc) by comparing a query case (new data) with similar existing data from a database. [76] combines CBR with an algorithm that presents the similarity between these cases by visually providing proxies and measures. 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.

Logical statements as interpretability – as previously mentioned, such works include [39] [40] [41].

Kernel function, dimensionality reduction for interpretability - Kernel function and dimensionality reduction are also used in the medical field to provide interpretability. Dimensionality reduction methods in medical field might be important since they afford us analysis methods independent of pre-defined models since the complexity of biological models often render pre-defined models insufficient. Medical data can thus be treated similarly as common images, except that medical data are often higher dimensional in their raw form. However, interpretability problems for unclear features extracted might have just been delayed and remain unresolved. Eventually, human interpretation will be required in such situations. The following are some examples.

[52] uses Variational Autoencoder (VAE) to obtain vectors in 64-dimensional latent dimension in order to predict whether samples suffer from hypertrophic cardiomyopathy (HCM). A non-linear transformation is used to create Laplacian Eigenmap (LE) with two dimensions, which is suggested as the means for interpretability. [42] proposes Generative Discriminative Machine (GDM) that combines ordinary least square regression and ridge regression to handle confounding variables in Alzheimer's disease dataset. GDM parameters are said to be interpretable, since they are linear combinations of the clinical variables. [77] introduces frame singular value decomposition (F-SVD) for classifications for electromyography (EMG) data. [78] uses DWT-based method (discrete wavelet transform) to perform feature extraction before eventually feeding the EEG data into a neural network for epilepsy classification. [79] developed a host of wavelet-based feature extraction methods as well applied on EEG data for epilepsy classification.

Model-based interpretability - As previously mentioned, 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, [70] uses kinetic model for the cerebral blood flow in $ml/100g/min$ with

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

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 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.

Deep learning method is also used to perform parameters fitting for Magnetic Resonance Spectroscopy [59]. The parametric model specified, $x(t) = \Sigma \mathrm{a_m} x_m(t) e^{\Delta\alpha_m t+2\pi i\Delta f_m t}$, consists of linear combination of metabolite signals $x_m(t)$. 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. Likewise, deep learning has been used for PET pharmacokinetic (PK) modelling to quantify tracer target density [80]. 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).

On a different note, reinforcement learning (RL) has been applied to personalized healthcare. In particular, [81] 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.

5.2. Risk of Machine Interpretation in Medical Field

- **Jumping conclusion**. According to [39], logical statements such as "*has asthma → 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 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.
- **Manipulation of explanations.** [82] shows that an image can be generated that is perceptibly indistinguishable from the original but produces radically different interpretation. Furthermore, explanation can even be manipulated arbitrarily [83]. For example, an explanation for the classification of a cat image can be implanted into the prediction of the image of a dog. The risk in medical field is clear: even without malicious, intentional manipulation, noises can render "explanations" wrong.
- **Incomplete constraints**. In [70], the loss function 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 [42] [59] [80].
- **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 [84] 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 for lesion segmentation nowadays might not be very successful because there is simply not enough useful information contained in the MRI scans fed into the models.

6. 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. Visual and textual explanation supplied by an algorithm might seem like the obvious choice; unfortunately, imagine an otherwise reliable deep learning model providing a strangely wrong visual or textual explanation. Before the black-box is unblack-boxed, machine decision always carries some exploitable risks. It is also clear that a unified notion of interpretability is elusive. 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 side by side 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,* 2017.

[2] Olaf Ronneberger, Philipp Fischer and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Springer Link, MICCAI,* 2015.

[3] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce and H. P. Beck, "The role of trust in automation," *International Journal of Human–Computer Interaction,* 2003.

[4] Liang Chen, Paul Bentley and Daniel Rueckert, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *Neuroimage: Clinical,* 2017.

[5] Ozgun Cicek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox and Olaf Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," *Springer Link, MICCAI,* 2016.

[6] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Silviana Ciurea-Ilcus, Chris Chute, Henrik Mark,

Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Yifan Yu, 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," 2019. [Online]. Available: https://arxiv.org/pdf/1901.07031.pdf.

[7] Fausto Milletari, Nassir Navab and Seyed-Ahmad Ahmadi, "V-Net: fully convolutional neural network for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision (3DV)*, 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," 2016. [Online]. Available: https://arxiv.org/abs/1606.00915.

[9] Finale Doshi-Velez and Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning," 2017. [Online]. Available: https://arxiv.org/pdf/1702.08608.pdf.

[10] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden and Anna Goldenberg, "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use," 2019. [Online]. Available: https://arxiv.org/abs/1905.05134.

[11] J. L. Herlocker, J. A. Konstan and J. Riedl, "Explaining collaborative filtering recommendations," *Conference on Computer Supported Cooperative Work,* 2000.

[12] Surjo R. Soekadar, Niels Birbaumer, MarcW. Slutzky and Leonardo G. Cohen, "Brain–machine interfaces in neurorehabilitation of stroke," *Neurobiology of Disease,* 2015.

[13] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal and Heimo Müller, "Causability and explainabilty of artificial intelligence in medicine," 2019. [Online]. Available: https://onlinelibrary.wiley.com/journal/19424795.

[14] Yao Xie, Xiang ‘Anthony’ Chen and Ge Gao, "Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis," in *IUI Workshops*, 2019.

[15] Zachary C. Lipton, "The Mythos of Model Interpretability," 2017. [Online]. Available: https://arxiv.org/pdf/1606.03490.pdf.

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

[17] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine,* 2019.

[18] Alberto Fernandez, Francisco Herrera, Oscar Cordon, Maria Jose del Jesus and Francesco Marcelloni, "Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?," in *IEEE Computational intelligence magazine*, 2019.

[19] 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,* 2019.

[20] "ec.europa.eu," European Union, April 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. [Accessed June 2019].

[21] Danding Wang, Qian Yang, Ashraf Abdul and Brian Y. Lim, "Designing Theory-Driven User-Centric Explainable AI," in *Conference on Human Factors in Computing Systems*, 2019.

[22] 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,* 2019.

[23] Chris Olah, Alexander Mordvintsev and Ludwig Schubert, "distill.pub," Google, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization/.

[24] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye and Alexander Mordvintsev, "distill.pub," Google, 2018. [Online]. Available: https://distill.pub/2018/building-blocks/.

[25] Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," in *Neural Information Processing Systems*, 2017.

[26] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, "“Why Should I Trust You?” Explaining the Predictions of Any Classifier," 2016. [Online]. Available: https://arxiv.org/pdf/1602.04938.pdf.

[27] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller and Wojciech Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE,* 2015.

[28] Amirata Ghorbani, James Wexler, James Zou and Been Kim, "Towards Automatic Concept-based Explanations," 2019. [Online]. Available: https://arxiv.org/abs/1902.03129.

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba, "Learning Deep Features for Discriminative Localization," 2015. [Online]. Available: https://arxiv.org/abs/1512.04150.

[30] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," [Online]. Available: https://arxiv.org/abs/1610.02391.

[31] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang and Min Xu, "Respond-CAM: Analyzing Deep Models for 3D Imaging Data by Visualizations," in *MICCAI*, 2018.

[32] Matthew D. Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision - ECCV*, 2014.

[33] 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," 2017. [Online]. Available: https://arxiv.org/abs/1711.00867.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich, "Going Deeper with Convolutions," 2014. [Online]. Available: https://arxiv.org/abs/1409.4842.

[35] M. Paschali, Sailesh Conjeti, Fernando Navarro and Nassir Navab, "Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples," in *MICCAI*, 2018.

[36] Heather D. Couture, J. S. Marron, Charles M. Perou, Melissa A. Troester and Marc Niethammer, "Multiple Instance Learning for Heterogeneous Images: Training a CNN for Histopathology," in *MICCAI 2018. Lecture Notes in Computer Science*.

[37] Xiaoxiao Li, Nicha C. Dvornek, Juntang Zhuang, Pamela Ventola and James S. Duncan, "Brain Biomarker Interpretation in ASD Using Deep Learning and fMRI," in *MICCAI 2018. Lecture Notes in Computer Science*, 2018.

[38] Yao Qin, Konstantinos Kamnitsas, Siddharth Ancha, Jay Nanavati, Garrison Cottrell, Antonio Criminisi and Aditya Nori, "Autofocus Layer for Semantic Segmentation," 2018. [Online]. Available: https://arxiv.org/abs/1805.08403.

[39] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm and Noémie 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*, 2015.

[40] B. Letham, C. Rudin, T. H. McCormick and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics,* 2015.

[41] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman and Finale Doshi-Velez, "An Evaluation of the Human-Interpretability of Explanation," Google Brain, 2019. [Online]. Available: https://arxiv.org/abs/1902.00006.

[42] Erdem Varol, Aristeidis Sotiras, Ke Zeng and Christos Davatzikos, "Generative Discriminative Models for Multivariate Inference and Statistical Mapping in Medical Imaging," 2018. [Online]. Available: https://arxiv.org/abs/1807.00445.

[43] T. Hastie and R. Tibshirani, Generalized additive models, Chapman & Hall, CRC Press, 1990.

[44] Y. Lou, R. Caruana and J. Gehrke, "Intelligible models for classification and regression," in *KDD*, 2012.

[45] Y. Lou, R. Caruana, J. Gehrke and G. Hooker, "Accurate intelligible models with pairwise interactions," in *KDD*, 2013.

[46] David R. Hardoon, Sandor Szedmak and John Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation,* 2004.

[47] Maithra Raghu, Justin Gilmer, Justin Gilmer and Jascha Sohl-Dickstein, "SVCCA Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability," in *Neural Information Processing Systems*, 2017.

[48] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas and Rory Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," 2018. [Online]. Available: https://arxiv.org/abs/1711.11279.

[49] A. K. account, 2014. [Online]. Available: https://cs.stanford.edu/people/karpathy/cnnembed/. [Accessed 2019].

[50] Anh Nguyen, Jason Yosinski and Jeff Clune, "Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks," 2016. [Online]. Available: https://arxiv.org/pdf/1602.03616.pdf.

[51] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson and Chris Olah, "Exploring Neural Networks with Activation Atlases," 2019. [Online]. Available: https://distill.pub/2019/activation-atlas/.

[52] Carlo Biffi, Ozan Oktay, Giacomo Tarroni, Wenjia Bai, Antonio De Marvao, Georgia Doumou, Martin Rajchl, Reem Bedair, Sanjay Prasad, Stuart Cook, Declan O'Regan and Daniel Rueckert, "Learning Interpretable Anatomical Features Through Deep Generative Models: Application to Cardiac Remodeling," in *MICCAI*, 2018.

[53] Mingliang Wang, Daoqiang Zhang, Jiashuang Huang, Dinggang Shen and Mingxia Liu, "Low-Rank Representation for Multi-center Autism Spectrum Disorder Identification," in *MICCAI*, 2018.

[54] Pang Wei Koh and Percy Liang, "Understanding Black-box Predictions via Influence Functions," in *Proceedings of the 34-th International Conference on Machine Learning*, 2017.

[55] Levin Kuhlmann, Philippa Karoly, Dean R Freestone, Benjamin H Brinkmann, Andriy Temko, Alexandre Barachant, Feng Li, Gilberto Titericz, Jr. , Brian W Lang, Daniel Lavery, Kelly Roman, Derek Broadhead, Scott Dobson, Gareth Jones, Qingnan Tang, Irina Ivanenko, Oleg Panichev, Timothée Proix, Michal Náhlík, Daniel B Grunberg, Chip

Reuben, Gregory Worrell, David B Grayden and Mark J Cook, "Epilepsyecosystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG," *Brain,* vol. 141, no. 9, 2018.

[56] Martin Wiener, Friedrich T.Sommer, Zachary G.Ives, Russell A.Poldrack and BrianLitt, "Enabling an Open Data Ecosystem for the Neurosciences," *Neuron,* vol. 92, no. 4, 2016.

[57] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen and Klaus-Robert Mueller, "How to Explain Individual Classification Decisions," 2009. [Online]. Available: https://arxiv.org/abs/0912.1128.

[58] Guillaume Alain and Yoshua Bengio, "Understanding intermediate layers using linear classifier probes," *arxiv.org,* 2018.

[59] Nima Hatami, Michael Sdika and Helene Ratiney, "Magnetic Resonance Spectroscopy Quantification using Deep Learning," in *MICCAI*, 2018.

[60] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz and Felix Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage,* 2014.

[61] Mukund Sundararajan, Ankur Taly and Qiqi Yan, "Axiomatic Attribution for Deep Networks," 2017. [Online]. Available: https://arxiv.org/abs/1703.01365.

[62] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox and Martin Riedmiller, "Striving for Simplicity. The All Convolutional Net," 2015. [Online]. Available: https://arxiv.org/pdf/1412.6806.pdf.

[63] 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 *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2016.

[64] Oskar Maier and Heinz Handels, "Predicting Stroke Lesion and Clinical Outcome with Random Forests," in *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2016.

[65] B. Kim, C. M. Chacha and J. Shah, "Inferring Robot Task Plans from Human Team Meetings: A Generative Modeling Approach with Logic-Based," 2013. [Online]. Available: https://arxiv.org/abs/1306.0963.

[66] J. Cheng and Michael S. Bernstein, "Flock: Hybrid Crowd-Machine Learning Classifiers," in *ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015.

[67] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *arxiv.org.*

[68] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas and Rory Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," *arxiv.org,* 2018.

[69] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen and Yongjun Wang, "Artificial intelligence in healthcare past, present and future," *Stroke and Vascular Neurology,* 2017`.

[70] 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 *MICCAI*, 2018.

[71] Cassel CK and Jameton AL, "Dementia in the elderly: an analysis of medical responsibility," *Annals of Internal Medicine,* 1981.

[72] Pat Croskerry, Karen Cosby, Mark L. Graber and Hardeep Singh, Diagnosis: Interpreting the Shadows, CRC Press, 2017.

[73] Curtis P. Langlotz, Bibb Allen, Bradley J. Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S. C, 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," *Radiology,* 2019.

[74] 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 *MICCAI*, 2018.

[75] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2016. [Online]. Available: https://arxiv.org/pdf/1409.0473.pdf.

[76] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud and Brigitte Séroussi, "Explainable artificial intelligence for breast cancer: A visual Case-Based Reasoning Approach," *Artificial Intelligence In Medicine,* 2018.

[77] 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,* 2019.

[78] 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,* 2017.

[79] 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,* 2018.

[80] Catherine J. Scott, Jieqing Jiao, M. Jorge Cardoso, Kerstin Kl¨aser, Andrew Melbourne, Pawel J. Markiewicz, Jonathan M. Schott, Brian F. Hutton and Sebastien Ourselin, "Short Acquisition Time PET/MR Pharmacokinetic Modelling Using CNNs," in *MICCAI*, 2018.

[81] Feiyun Zhu, Jun Guo, Zheng Xu, Peng Liao, Liu Yang and Junzhou Huang, "Group-Driven Reinforcement Learning for Personalized mHealth Intervention," in *MICCAI*, 2018.

[82] Amirata Ghorbani, Abubakar Abid and James Zou, "Interpretation of Neural Networks is Fragile," 2017. [Online]. Available: https://arxiv.org/abs/1710.10547.

[83] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller and Pan Kessel, 2019. [Online]. Available: https://arxiv.org/abs/1906.07983.

[84] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Greg S. Corrado, Jason D. Hipp, Lily Peng and Martin C. Stumpe, "Detecting Cancer Metastases on Gigapixel Pathology Images," 2017. [Online]. Available: https://arxiv.org/pdf/1703.02442.pdf.

[85] Chris Olah, Alexander Mordvintsev and Ludwig Schubert, "distill.pub," Google, 2017. [Online]. Available: https://distill.pub/2018/building-blocks/.