Enhancing Interpretability of Deep Learning Techniques for Oncological Diseases: Current Trends and Future Horizons*

Abstract

The healthcare industry is currently collecting a vast amount of patient data, which due to its volume and diversity of modalities can be called "big data". Different machine learning methods have already been applied in oncology successfully. Deep learning, a sub-field of machine learning, have reached human-level performance in some tasks such as melanoma classification from dermoscopic images and lymph node metastases in breast cancers from pathology images. Moreover many methods have already been approved by the FDA. Although advances are being made for the introduction of artificial intelligence in the workflow of healthcare practitioners, the lack of interpretability of these methods is still a barrier for their adoption in clinical practice. The research community has recognized that the lack of interpretability is a problem and is focusing on developing methods to solve this problem. The aim of this article is to review the methods for interpreting deep learning models in the specific case of the oncology field. Furthermore, a literature review is presented to identify current problems and future directions of work on this field.

Keywords: Big Data, Intepretability, Deep Learning, Oncology, Machine Learning, Decision-support System

1. Introduction

Today, in healthcare scenarios, we are living a digital era where physical patient records are mapped to digital formats. This has opened the possibility to improve the efficiency and quality of treatment provided to patients by building decision-support systems.

Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) which studies algorithms that are capable to construct data driven models. The construction of such models follows two distinct phases - training and inference. During training, the algorithm builds a model which fits the data received as input, while in inference, the now trained model will produced results based on a new set of information that it receives exclusively in this phase and can be used to test its performance.

Between 2014 and 2019 the Food Drugs Administration approved 46 ML algorithms [1] encompassing different areas like mammogram screening and ultrasound image diagnosis, turning the application of ML in healthcare context a reality.

The majority of these algorithms are supervised which means that in these scenarios, they need a help of a physician to label the data before the mining process starts. As an example, in overall survival prediction of breast cancer patients it is necessary that a physician labels the set of patient data that will be used in the training process with the target variable (overall survival – typically measured in months). When this target variable is discrete we are present to a classification problem, or a regression problem in case the variable is continuous.

The Artificial Neural Network (ANN) is a popular supervised algorithm inspired by biological neuron, and began to be used in healthcare in the early 90s [2]. The ANN is an analogy used by computer scientists to emulate the behaviour of the human brain and are composed by an input, an output and intermediate layers, which are also called hidden layers. Similarly to biological neurons, each artificial neuron, or perceptron [3], receives a set of inputs, either from the input layer or other neurons, performs a linear combination based on its weights

*This article is a result of the project NORTE-01-0145-FEDER-000027, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

and make a non-linear decision whether to activate the neuron and fire.

Due to the increasing computational power, the complexity of these networks has substantially grown, materialized in the use of dozens of layers and millions of neurons. In this context, Deep Learning (DL) techniques - a subset of ANN techniques - emerged as the state of the art for many real world problems, surpassing other ML techniques, and reaching human-level performance in several task such as in the classification of melanoma from dermoscopic images [4], or the detection of lymph node metastases in breast cancers from pathology images [5].

Despite its vast potential DL suffers from several disadvantages. First is the dependency on large amounts of data and computational power. Also the black-box nature of DL make it difficult to interpret it.

The objective of this study goes towards knowing what strategies can we use in the oncological field to interpret the results produced by DL techniques. Other reviews already covered specific medical areas such as radiology [6] or more broadly at the medical field and machine learning models in general [7], but this is the first study to review in detail work of interpretability of DL techniques in the oncological field.

Results from this study were compiled by searching the PubMed database for articles published between January 2014 and March 2020, searching individually and in combination search terms such as "interpretability", "deep learning", "oncology", "cancer" and "decision support systems".

In this study we have found that the majority of works focus on medical imaging (e.g. mammogram, histological images and dermoscopic images) related to breast and skin cancer. Possible explanations are related to the highest prevalence of such diseases and also the dissemination of well curated datasets and challenges target at those diseases. Overall, the most frequent strategy to interpret DL decisions is to highlight the most relevant regions of the image.

Future work includes the evaluation of interpretability methods so that they can be compared and validated. Also, the development of workflows where the clinician can interact with the model and the interpretation to make decisions.

Throughout the next two overview sections, we will talk about various ANN techniques illustrating their internal architectures and learning processes using a self-explanatory oncological example, that consists of the classification of a breast tumor based on handcrafted features such as mass density (fat-containing - 0, low - 1, equal - 2, high - 3), shape (round - 0, oval - 1, irregular - 2) and the breast side that it was found (left - 0 or right - 1) as well as the raw mammogram. Using such features as an input, the goal of the different types of ANN's will be predict an output related to the malignancy of the tumor (0 means benign and 1 malignant).

2. ANN Techniques Overview

The Perceptron [8] is the the precursor to the ANN techniques.

Training process - As seen in Figure 1a, after receiving a set of variables as input (x_1, x_2, ..., x_n), the perceptron will attribute weights for each variable (w_1, w_2, ..., w_n) and afterwards will use a mathematical function also known as activation function (green) that will use the input variables combination (yellow) to produce a desired output (y). For each set of input variables, the output (y) is compared to the label corresponding to expected output, also known as target. During training, the weights are continuously changed to move the output of the perceptron and the target closer together.

In the example provided in Figure 1b, the perceptron is given the breast cancer tumor variables density, shape, and side (blue) and given the weights obtained during training (0.8, 0.7 and 0 respectively), predicts the tumor to be malignant.

The Multilayer perceptron (MLP) [8] is the natural extension of the perceptron to solve more complex problems. Rather than having a single unit, or neuron, the MLP has multiple layers with multiple neurons each, as can be seen in yellow in Figure 2 a). Due to its multiple layered structure, the MLP can be seen as a deep neural network.

Training process - After receiving a set of variables as input (x_0, x_1 ..., x_n), each intermediate neurons (yellow) acts like a perceptron, performing the weighted combination of its inputs and apply a nonlinear activation function which help to solve nonlinear problems. The output of activation function of each neuron, also known as activation, acts as input for the neurons of the next layer. The combination of activation of the last intermediate layer produces a desired output (y).

In the example (Figure 2b), given the breast cancer tumor variables density, shape, and side (blue),

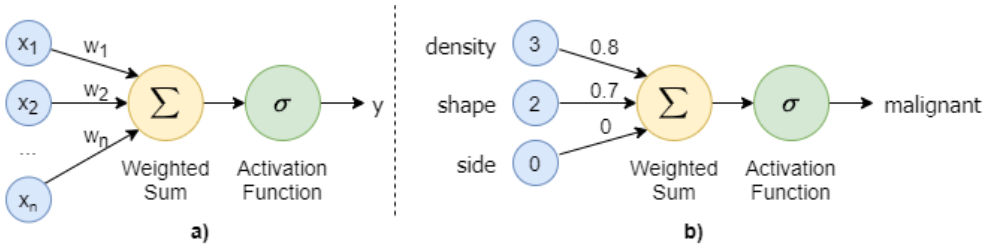

Figure 1: An illustration of the Perceptron. a) Theoretical scheme - input variables (blue), weighted sum which combines the inputs (yellow), activation function (green) which turns the output into a binary prediction y. b) Practical example - breast cancer tumor variables (blue) are combined using the weights (3, 2 and 0) into making the prediction of malignant.

the MLP with its weights obtained during training, predicts the tumor to be malignant.

Due to its nature, MLPs do not scale well to images. As an example, for an image with a width and height of 100 pixels, the MLP would require 10,000 neurons just in the first layer and this number would grow exponentially with each layer. To address this issue, Convolutional Neural Networks (CNN) [9, 10] techniques emerged as a possible solution.

Training process - CNNs treat the image as a matrix (Figure 3), extracting features using a mathematical operation called convolution which helps preserve the spatial relationship between neighboring pixels. The convolution slides a small matrix, called filter, over the original image, and for every position, it computes the element-wise multiplication between the two matrices, and the resulting value forms a single element of the output matrix, called feature map. The filter is composed of weights (w) that are learned during training.

During feature extraction, each convolutional layer is composed by n filters resulting in n feature maps (Figure 4a purple). The values of the feature maps of the last convolutional layer are concatenated into a single vector (blue) and used as an input for a MLP (yellow) which makes the prediction y. During training, the values of the filter matrices and of the MLP are continuously changed to move the output closer to the expected targets.

In the example provided in Figure 4b), given a mammogram, the CNN has already learn the weights of the filters, and during the feature extraction (purple) is able to extract features (blue) which may include the density and the shape of the tumor. The features are used to make the classification, which predicts the tumor to be malignant.

Although CNNs are able to take advantage of the spatial relationships between pixels, they struggle with large sequence data such as text. Recurrent Neural Networks (RNN) techniques solve this issue by having a small network looped for each element of the sequence, allowing information to persist.

Training process - RNNs are usually composed by only a layer of neurons (yellow), which taking an input (blue) predicts the output (green) in a recurrent way (Figure 5a). This refers to the fact that its processing unit (yellow) is looped n times, where n represents the number of elements of the sequence. During training, the weights of the RNN are continuously changed to minimize the difference between the target sequenced, and the predicted one.

In the example provided in Figure 5b, the RNN has already learned to generate text based on the a set of features extracted from a mammogram. At each set, and based on the information that is passed from the previous step, it generates the word that is most likely. So for the first word, it predicts 'Found' as it learned that the sentences given during training usually start this way. Next, based on the word 'Found' it decides that word 'mass' is the most likely word to follow. This process repeats until the RNN generates the end of sequence. For example the sequence "Found mass in right breast with irregular shape and high density.".

Unlike the other revised techniques, the autoencoder [11] is a unsupervised algorithm - in the training process, this technique does not require labelled data. The goal of autoencoders is to learn a compressed representation (code) of the input data by reconstructing it as the output of the network. By restricting the size of the code, the technique can discover the interesting structures of the data, and in the case of denoising autoencoder, even reconstruct noisy images.

Training process - The autoencoder (Figure 6a) contains an encoder (purple) which re-

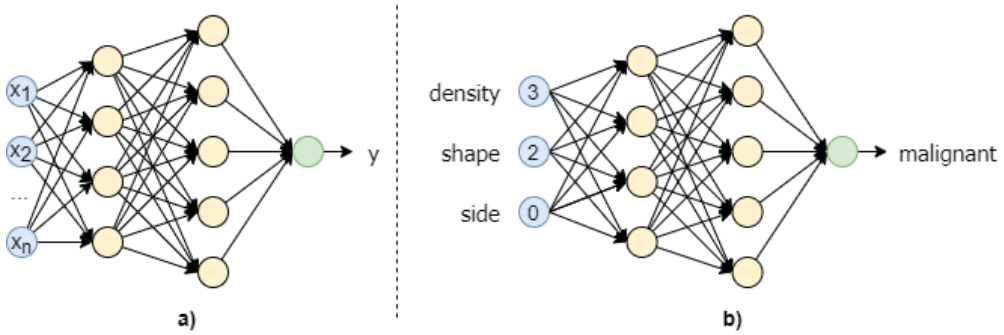

Figure 2: An illustration of the Multilayer Perceptron (MLP). a) Theoretical scheme - input variables (blue), intermediate layer with neurons similar to perceptrons (yellow), activation function (green) which transforms the output into a binary classification y. b) Practical example - breast cancer tumor variables (blue) are combined during multiple layers into making the prediction of malignant.

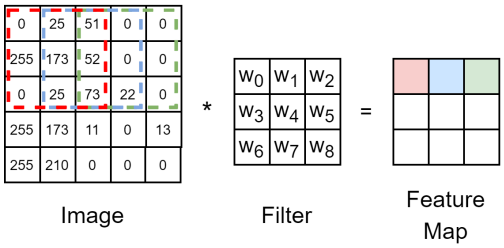

Figure 3: An illustration of the convolution operation. Each element of the feature map is the result of the element-wise multiplication between the region of the image and the filter.

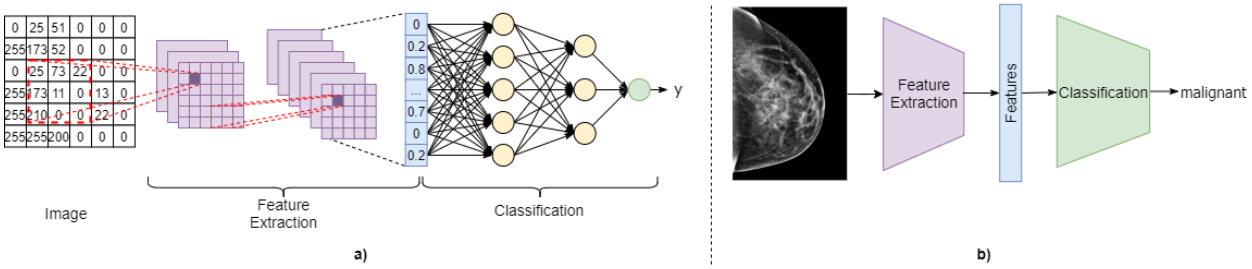

Figure 4: An illustration of the Convolutional Neural Network (CNN). a) Theoretical scheme - image is represented as matrix. Feature extraction extracts the feature maps (purple) using the convolution operator (red). The output of the last convolution layer is concatenated into a feature vector (blue) which serves as input for the classification MLP (yellow). The activation function (green) transforms the output into a binary prediction y. b) Practical example - A mammogram showing a tumor is provided, the feature extraction (purple) extracts features (blue) which may include the density and shape of the tumor. The features are used to classify the tumor as malignant.

ceives the noisy input (blue), compresses into a small representation, called code (yellow), and is reconstructed by a decoder (green) into the original noiseless input. During training, the weights of the neurons present in the encoder and decoder are continuously changed to reduce the difference between the original input and the output, called reconstruction error, to find useful patterns in the data.

In the example provided in Figure 6 b), the autoencoder is given a noisy mammogram and its task is to denoise it. First the encoder (purple) compresses the mammogram into the code maintaining useful information for the reconstruction, then the decoder (green) reconstructs code into the mammogram without noise.

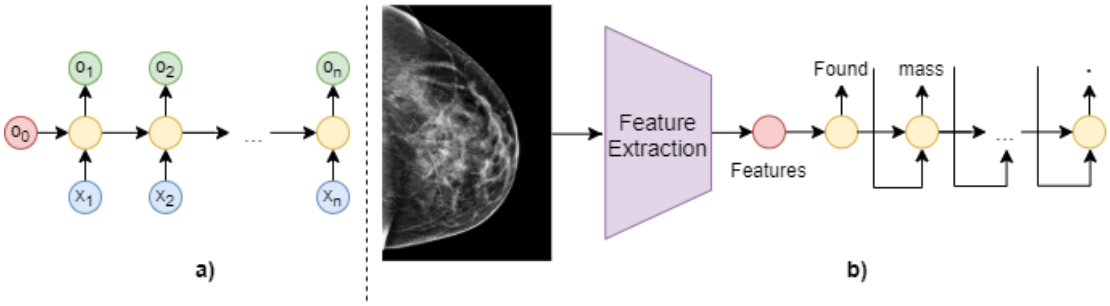

Figure 5: An illustration of the Recurrent Neural Network (RNN). a) Theoretical scheme - an initial information is provided to the RNN (red), and for each element of the input sequence (blue), the layer of neurons (yellow) predicts the next element of the sequence (green). Practical example - report of a mammogram is generated by extracting visual features and providing them to the RNN. The RNN then generates at each step the word most likely to come next, given the previous generated word.

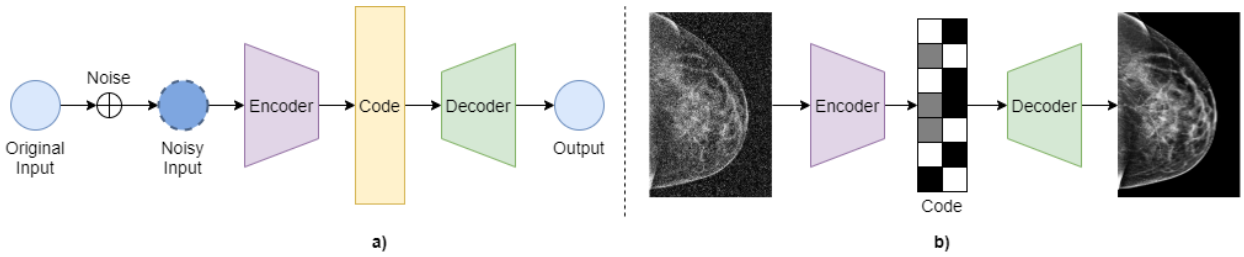

Figure 6: An illustration of the Denoising Autoencoder. a) Theoretical scheme - noise is added to the original input (blue), the encoder (purple) learn a compressed representation from the input (blue), the decoder (green) then reconstructs the code into the original input (blue). b) Practical example - a noisy mammogram is encoded into a compressed representation (code) and then the code is decoded into the denoised version of original mammogram.

3. Interpretability Concepts Overview

There is no consensus upon the definition of interpretability [12]. However one of the most used was presented by Doshi-Velez and Kim [13] which defined interpretability as the "ability to explain or to present in understandable terms to a human", and will be used in this work. Since deep learning techniques reach human performance in melanoma diagnosis from dermoscopic images [4], or the detection of lymph node metastases in breast cancers from pathology images [5], the need of interpret them emerge specially in healthcare contexts.

3.1. Dimensions of Interpretability

Interpretability methods can be characterized by a set of dimensions [14]: global and local interpretability, intrinsic and post-hoc interpretability and model-specific and model-agnostic interpretability.

Global and Local Interpretability. to perform a classification task a machine learning algorithm first creates a data-driven model based on a set of input features (e.g. age and sex) during the training phase. The objective of this phase is allowing neurons to select important features and learn relationships between them and the target output. In global interpretability we are interested in analyzing this model, to understand the common patterns in the overall data that help make decisions, by studying the model's parameters (i.e. weights), and the learned relationships. In local interpretability we are interested in understanding the relationship between a specific set on input features and the model decision.

In our example, based on the examples provided, the network learned relationships that help predict the tumor maligancy, based on its density, shape and breast side. As the breast side (left or right) where the tumor appears is not indicative of the level of malignancy, the network should have learned to discard this input feature. Global interpretability could be help understand which relationships did the network learn, and for the example of breast side confirm that it was not used. Local

interpretability could help understand the importance of the input features, when given a specific set of values.

Intrinsic and Post-hoc Interpretability. While the increase of complexity of ANNs (i.e. number of neurons), help solve complex problems, it increase the difficulty to interpret them. Restricting the network's complexity and adding self-explanatory components during the training phase helps the model to obtain intrinsic interpretability, while post-hoc interpretability is obtained using interpretability methods after training.

In our mammogram example, we could add a self-explanatory component which learns to segment the most important region (e.g. breast quadrant with the tumor) and use it to make the prediction. The same can type of explanation can also be achieved after training by using a interpretability method without the need for self-explanatory components.

Model-specific and Model-agnostic. Another way to classify interpretability methods is based on the dependence the method has on the type of model which it tries to explain. Model-agnostic methods can be applied to different types of models, while model-specific methods are only applicable to a specific type of model [14].

In our example, while a model-agnostic method could extract the importance of the density and shape from a model trained from any ML algorithm, a model-specific method would only be able to do the same for similar models.

3.2. Interpretability Strategies

During the training phase, DL algorithms create data-driven models that can be interpreted using different strategies producing different types of explanations.

3.2.1. Feature Importance

One of the more explored explanations is feature importance, which gives the importance or contribution of an input feature on the prediction of an example. Two main approaches are used for computing feature importance: sensitivity analysis [15] and decomposition [16, 17].

Sensitivity analysis computes the effects of the variation in the input variables in the model's output and help us answer the question "What change would make the instance more or less like a specific category?".

Decomposition approaches successively decomposes the importance of the output of a layer into previous layers, until the contribution that the input features have on the output is found. It help us answer the question "What was the feature's influence on the model's output?".

If we extract the feature importance of a decision of our example, it can have different meanings depending on the type of method used (Figure 7). High sensitivity values for density and shape means that their growth would also increase the prediction of malignancy. While high contribution values of density and shape means that the prediction of malignancy was highly influenced by the value of these features.

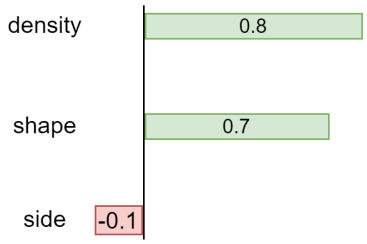

Figure 7: Illustration of feature importance of density, shape and side on the classification of breast tumor malignancy. Green - positive importance in the classification, red -negative importance in the classification.

3.2.2. Saliency Map

When dealing with images, feature importance can be visualized using an heatmap or a saliency map [13]. Using an image as output, this will indicate the region that serves as the base for a particular decision.

An example of a saliency map, extracted from a CNN trained to predict the malignancy based on mammogram patches is seen in Figure 8. The red and yellow regions correspond to the most important regions of the image. The method correctly focus on the mass, supporting our confidence in the model's decisions.

3.2.3. Model Visualization

The ML algorithm receives an example with a set of input features, and in their internal process creates a combination of its features also called internal features. Some strategies help visualize patterns detected in an image [19], while others help visualize the feature distribution in the dataset [20, 21].

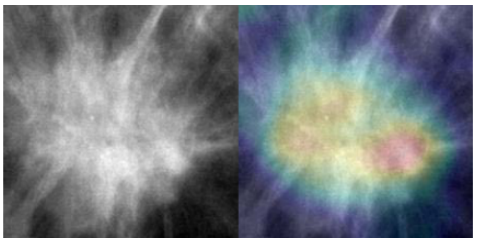

Figure 8: Example of a saliency map depicting the important pixels for the malignancy prediction based on mammogram patches [18]. Left: mammogram patch used during classification. Right: saliency map, where the pixel color indicates the importance of the pixel in the classification (red - high importance, blue - low importance).

Also, while some strategies help to find the image which contain a pattern detected by the network [22], others artificially create images which accentuate the same patterns [23, 24].

In Figure 9 we can see regions of mammograms which contain patterns detected by individual filters of the CNN trained to diagnose the tumor malignancy.

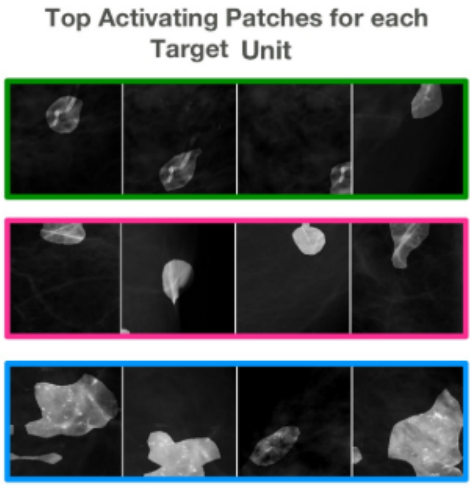

Figure 9: Illustration of the internal behaviour of a network unit by visualizing regions of mammograms with patterns detected by individual units of the network [22].

3.2.4. Surrogate model

A surrogate model is an interpretable model which was trained with the objective of extracting a list of rules allowing the clinician to understand the knowledge produced by the algorithm. One way of doing this is by creating a new dataset where each example of the dataset used to train the DL model is combined with its prediction and the task of the surrogate model is to predict this values.

To better understand what is a surrogate model, let's consider the example in Figure 10. On the left we can see a MLP trained to classify the malignancy of a tumor based on its density and shape. On the right we can see a list of rules extracted from the MLP that demonstrate its decisions.

3.2.5. Domain Knowledge

Although DL algortihms extract internal features (combination of input features) automatically during the training phase, the domain knowledge of the medical field which physicians have can be used to validate the decision of the network.

In the case of tumor malignancy prediction, physicians take into account the density and shape of the tumor to make their decision. In order for physicians to trust the prediction of malignancy of a network trained on mammograms, intermediate predictions of density and shape can be provided (Figure 11).

3.2.6. Example-based explanation

Example-based explanation methods select examples of the dataset that explain the behavior of the network [14]. This behavior is usually explained using the internal features (combination of input features) extracted from the examples by the network.

Similar examples are examples of the dataset that have similar values on the internal features and produce the same prediction as the example whose prediction we are explaining [25].

Counterfactual explanations can be used to explain predictions of examples by finding small changes in the example that cause the network to change its prediction. Some of these changes, although imperceptible to the human eye can fool the network into misclassifying instances with high confidence [26], these are called adversarial examples.

Usually examples of a dataset can be grouped together based on existing patterns. A prototype is a particular example of the dataset representative of its group.

Figure 12 shows examples of similar and counterfactual examples based on our example of breast cancer diagnosis. The color represents the predicted class, green is benign and red is malignant. The similar example are two instances close together with same predicted class, while the counterfactual

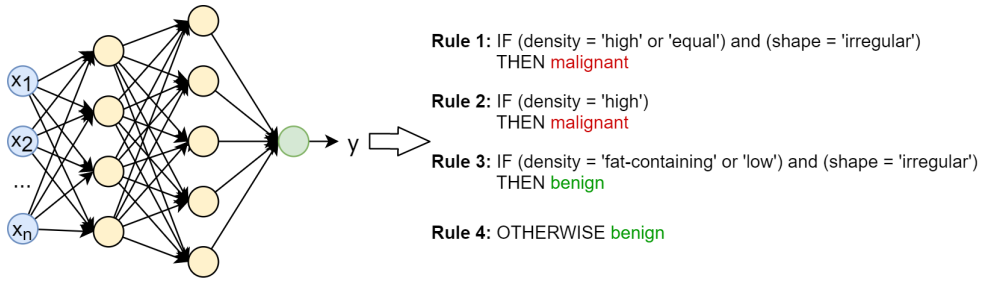

Figure 10: Illustration of a surrogate model extracting a rule list (right) from a MLP (left) trained to predict the malignancy of a breast tumor.

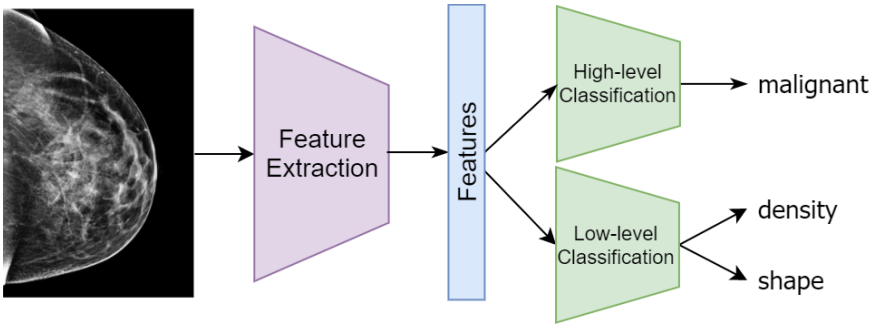

Figure 11: Illustration of how domain knowledge can be incorporated into the network to make more trustworthy predictions.

example are two instances close together but with different predicted class.

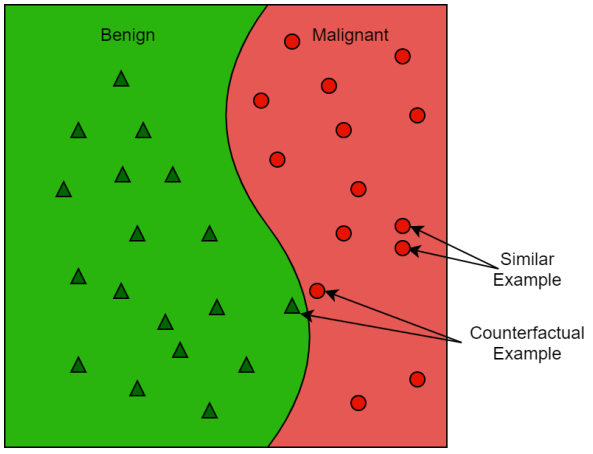

Figure 12: Illustration of similar example and counterfactual example. Color represents the predicted class (green - benign, red - malignant).

4. Interpreting Deep Learning in Oncology

The use of DL techniques has become widespread in the oncology area, covering different pathologies, but their interpretation remains an unexplored field [27, 28]. In this section, an overview about interpretably strategies applied to oncological disease will be performed.

As previously mentioned, we conducted a search of papers combining deep learning and the oncological field, and compiled the results in Table 1. In total, 44 works were found, where the majority target in breast cancer (30%), skin cancer (23%), lung cancer (9%) and brain cancer (11%). The most commons interpretability strategies were saliency maps (43%) and feature importance (27%) and among the prediction tasks, most works focused on diagnosis of malignancy (45%) and of different pathologies (27%),.

4.1. Breast Cancer

Prediction of breast cancer malignancy has been one the most successful applications of deep learning in oncology, achieving 86.7% sensitivity and 96.1% specificity when diagnosing mammograms [71]. It also is the main task on interpretability work (69%). Due to the availability of well-curated public datasets on breast cancer, mainly mammograms and hematoxylin and eosin (H&E)

Table 1: Summary of papers reviewed.

Ref	Disease	Task	Modality	Explanation	Dataset
[29]	Breast Cancer	Metastases Detection	WSI H&E	Model Visualization, Saliency Map	Public
[22]	Breast Cancer	Malignancy Diagnosis	Mammogram	Model Visualization	Public
[30, 31]	Breast Cancer	Malignancy Diagnosis	WSI H&E	Feature Importance, Domain Knowledge	Public
[32]	Breast Cancer	Malignancy Diagnosis	Mammogram	Domain Knowledge, Saliency Map	Public
[33]	Breast Cancer	Malignancy Diagnosis	Mammogram, Ultrasound, DCE-MRI, Hand-crafted	Domain Knowledge	Public
[18]	Breast Cancer	Malignancy Diagnosis	Mammogram	Saliency Map, Text	Public
[34]	Breast Cancer	Malignancy Diagnosis	Hand-crafted	Feature Importance	Public
[35]	Breast Cancer	Malignancy Diagnosis	Hand-crafted from H&E	Surrogate Model	Private
[36]	Breast Cancer	Malignancy Diagnosis	Hand-crafted from H&E	Surrogate Model	Public
[37]	Breast Cancer	Survival Prediction	Gene expression, Cancer biomarkers	Feature Importance	Public
[38]	Breast Cancer	Predict Estrogen Receptor Status	Metabolomics Data	Feature Importance	Public
[39]	Breast Cancer	Clustering	Gene expression, CNA data	Model Visualization	Public
[40]	Skin Cancer	Malignancy Diagnosis	Dermoscopic images	Model Visualization	Public
[41]	Skin Cancer	Malignancy Diagnosis	WSI H&E	Saliency Map	Private
[42]	Skin Cancer	Malignancy Diagnosis	Dermoscopic images	Saliency Map	Public
[43]	Skin Cancer	Diagnosis of skin lesion	WSI H&E	Saliency Map	Public
[44]	Skin Cancer	Diagnosis of skin lesion	Dermoscopic images	Saliency Map	Public
[45]	Skin Cancer	Malignancy Diagnosis	WSI H&E	Saliency Map	Public
[46]	Skin Cancer	Diagnosis of skin lesion	Dermoscopic images	Example	Public
[47, 48]	Skin Cancer	Malignancy Diagnosis	Dermoscopic images	Feature importance, Example, Surrogate Model	Public
[49]	Skin Cancer	Diagnosis of skin lesion	Dermoscopic images	Example, Saliency Map	Public
[50]	Lung Cancer	Disease Diagnosis	Chest Radiograph	Saliency Map	Public
[51]	Lung Cancer	Malignancy Diagnosis	CT	Domain knowledge	Public
[52]	Lung Cancer	Malignancy Diagnosis	CT	Domain knowledge	Public
[53]	Lung Cancer	Prediction radiation reaction	Biomarker, clinical data	Domain knowledge	Private
[54]	Brain Cancer	Tumor Grading	MRI	Saliency Map	Public
[55]	Brain Cancer	Tumor Grading	MRI	Feature Importance, Saliency Map	Public
[56]	Brain Cancer	Predict Methylation State	MRI	Model Visualization	Public
[57]	Brain Cancer	Survival Prediction	MRI	Feature Importance	Public
[58]	Brain Cancer	Survival Prediction	WSI H&E and Genomic Biomarkers	Saliency Map	Public
[59]	Other	Malignancy Diagnosis	Gene expression	Feature Importance	Public
[60]	Other	Survival Prediction	Gene and protein expression	Feature Importance	Public
[61]	Other	Disease Diagnosis	RNA-seq expression, SVN data	Surrogate Model, Feature Importance	Private
[62]	Other	Disease Diagnosis	Volumetric Laser Endomicroscopy	Saliency Map	Private
[63]	Other	Disease Diagnosis	Endoscopic images	Saliency Map	Public
[64]	Other	Disease Diagnosis	WSI H&E	Saliency Map	Private
[65]	Other	Disease Diagnosis	DESI	Cluster	Private
[66]	Other	Disease Diagnosis	Ophtalmic images	Domain Knowledge	Private
[67]	Other	Malignancy Diagnosis	Ultrasound	Domain knowledge	Private
[68]	Other	Malignancy Diagnosis	WSI H&E	Text, Saliency Map	Public
[69]	Other	Disease Diagnosis	Chest Radiograph	Text, Saliency Map, Text	Public
[70]	Other	Tumor Grading	WSI H&E	Text, Saliency Map	Private

stained histological images, research in this area has as taken a step forward.

When dealing with imaging data, researchers found it important to visualize the patterns detected by the networks either through model visualization techniques or with saliency maps. These patterns were then either validated by experts or correlated with medical concepts. For other types of data (e.g. gene expression, hand-crafted features), researchers mainly focused on computing feature importance or extracting surrogate models (i.e. rule lists).

Graziani *et al.* [29] visualized the patterns of a metastases detection CNN for WSI H&E images by synthesizing images that increase the network's confidence on the prediction (Activation Maximization [23, 24]) and by extracting saliency maps [72]. They found that the network detected nuclei-resembling shapes and regions of nuclei with marked variations in size and irregular shapes. Hsieh *et al.* [22] used Network Dissection method [73] to visualize the patterns of individual filters of a malignancy classifier based on mammograms and developed a web-based tool which let experts label the patterns. Figure 13 shows an example of a pattern which was labeled as 'Calcified Vessels'. Also, other BI-RADS [74] medical concepts (e.g. mass margin) were found to overlap with patterns detected by the network.

Rather than being validated by experts, Graziani *et al.* [30, 31] introduced Regression Concept Vectors (an extension of Concept Activation Vectors [75]) which let them detected the importance of medical concepts (i.e. area, perimeter and contrast) on the decisions of a breast cancer malignancy classifier based on WSI H&E network, even though they were not present in the training dataset. Constrast was found to be positively correlated with malignancy, while correlation was negatively correlated. Kim *et al.* [32] used medical concepts during training, computing their importance alongside saliency maps to help explain the malignancy diagnosis of mammograms.

Antropova *et al.* [33] visualized the values of both

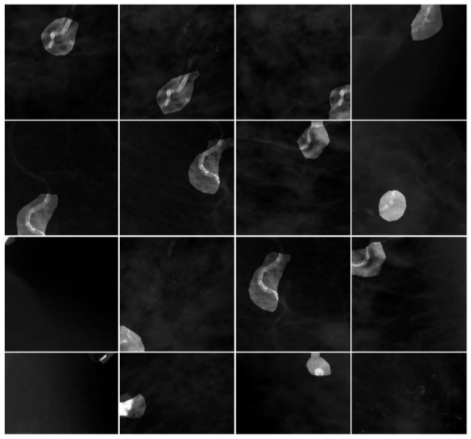

Figure 13: Example of pattern detected by the network and labeled by an expert as 'Calcified Vessels' in the web-based labeling tool [22].

deep features and hand-crafted features from different image modalities (i.e. Mammogram, Ultrasound, DCE-MRI) and found that their fusion improved malignancy diagnosis performance, most likely due to the low agreement between deep and handcrafted features.

Lee *et al.* [18] trained a malignancy diagnosis network able to justify its decisions both visually and textually. It trained a a language model that composes text description [32, 70, 69, 76] from mammograms. Although the descriptions are still not sufficiently good (i.e. "There are sharp lines on some part of complexly formed mass."), they show that this interpretability strategy has great potential.

When dealing with hand-crafted features relating with tumor size and shape, researchers found it important to simplify the network to behave linearly [34] making it easier to compute the feature importance, or extract simpler classifiers that could present physicians with simple rules (i.e. decision rules [35] and symbolic rules [36]) increasing interpretability.

Feature importance was the focus of most works dealing with gene expression data. For example, SALMON [37] predicted survival risk of patients with breast cancer, and feature importance of eigengene's modules and other clinical information, they confirmed that age, progesterone receptor status and other five mRNA sequence data coexpression modules play pivotal roles in patient prognosis. Similar methods, using the H2O [77] library, were used to detect the important features in the classification of the Estrogen Receptor Status of patients with breast cancer based on metabolomics data [38]. They found eight commonly enriched significant metabolomics pathways: isoleucine, putrescine, glycerol, 5'-deoxy-5'-methylthioadenosine, ornithine, tocopherol beta, phenylalanine, and arachidonic acid. Finally, Liu *et al.* [39] used an autoencoder to find clusters of breast cancer patients based on their gene expression and copy number alteration data, and visualized them using heatmaps. They found that the cluster of patients with ER-negative breast cancer patients usually have a poor prognosis.

4.2. Skin Cancer

Works in skin cancer almost evenly divided on the malignancy diagnosis and diagnosis of multiple skin diseases. The modality used was also divided between two types, dermoscopic images (70%) and hematoxylin and eosin (H&E) stained histopathological images (30%). Similarly to breast cancer detection, DL has also achieved great results in skin cancer detection based on medical imaging [78]. Interpretability methods for these pathologies ranged from saliency maps, model visualization, rule extraction, text explanations and example-based explanations.

A simple visualization method was used to visualize the activation of neurons of a CNN trained to predict the malignancy of dermoscopic images [40]. Inspection of activations let to finding neurons related to medical concepts such as borders, lesions, and skin type, as well as different image artifacts such as hairs.

Cruz-Roa *et al.* [41] proposed a DL technique for the malignancy diagnosis using histological images and visualized the most salient patterns in that task which when validated by pathologists were found to be related large-dark nuclei. Researchers also tried to improve the quality of saliency maps by making changes on the architecture of the network when diagnosis malignancy based dermoscopic images [42] and diagnosis of skin diseases based on WSI H&E images [43]. PatchNet [42] found a trade-off between interpretability and performance, as smaller patch sizes provided saliency maps with better visual interpretability at the expense of worse generalization capabilities. Paschali *et al.* [43] also found that smaller convolutional filters resulted in more fine-grained saliency maps. González-Díaz [44] incorporated segmentation of lesion areas based on

high-level dermoscopic features, and used these segmentations to diagnose of skin lesions and show relevant regions.

Example-based explanation are also useful interpretability strategies in skin cancer, as shown by Sadeghi *et al.* [46] which conducted a study which revealed that similar examples provided by DL techniques help users in classifying skin lesions from dermoscopic images. In the study, accuracy increased from 51.56% to 60.94% when the 15 most similar cases were provided to the users. Silva *et al.* [47, 48] unified complementary explanations to explain skin lesion predictions from dermoscopic images. The method extracted rules and presented them as text sentences alongside positive and a counter-factual examples for every decision. Also on the same task, Codella *et al.* [49] explained the decision with similar examples using k-nearest neighbors on the deep features and highlighted the most salient regions of the image.

4.3. Lung Cancer

Interpretability research on the diagnosis of lung cancer focused mainly on two modalities, Chest Radiography (X-Ray) or Computed Tomography (CT). Similarly, to breast and skin cancer, DT techniques have been shown to be able to reach human-level performance. In the diagnosis of 14 different pathologies from chest radiographs, a CNN achieved radiologist-level performance [50]. Radiologists confirmed, by inspecting saliency maps [79], that the network localizes accurately the lung masses.

Other works focused on the integration between hand-crafted features related to medical concepts and deep features. Paul *et al.* [51] developed a model for the malignancy diagnosis of lung cancer using CT images, and interpreted their correlation with medical features used by physicians by iteratively replacing deep features and evaluating the drop in confidence. Although deep features were not found to be perfectly correlated with medical features, they could represent 9 of the medical features with the deep features without losing performance. In the same task, Shen *et al.* [52] proposed to model that made high-level predictions for the tumor malignancy, and low-level predictions of medical features - calcification, subtlety, lobulation, sphericity, internal structure, margin, texture and spiculation. The approach achieved comparable or better results with state-of-the-art methods in the public Lung Image Database Consortium (LIDC).

Finally, Cui *et al.* [53] used a combination of hand-crafted features composed of clinical features and cancer biomarkers in a nonsmall cell lung cancer who received radiotherapy to predict the damage caused by the treatment. The results found that better performance was achieved by integrating the hand-crafted features with the deep features extracted from a autoencoder [80].

4.4. Brain Cancer

Unlike previous pathologies, brain cancer research deviates from diagnosis of diseases and focus on survival prediction (40%) and tumor grading (40%), almost entirely based on Magnetic Resonance Imaging (MRI) (83%).

When performing tumor grading - distinguishing from lower grade gliomas from high grade gliomas from MRI - researched have focused on producing saliency maps from the 3D MRI scans or Region of Interest (ROI) annotated by experts. Pereira *et al.* [54] extended existing saliency map methods for three dimensional inputs [81, 72]. The ROI classifier achieved better performance than the 3D scan (92.98% and 89.50% accuracy), but they were both able to locate the tumor. Pereira *et al.* [55] also used a feature importance method [82] to identified MRI sequences which were relevant for features extracted from the network, and then produce saliency maps. The sequences chosen were consistent with domain knowledge.

Han and Kamdar [56] train a model to predict the methylation state of the MGMT regulatory regions using MRI of Glioblastoma Multiforme (GBM) patients, resulting in 62% accuracy. The MRI scans were extracted from the Cancer Imaging Archive (TCIA) [83] and the methylation data from the Cancer Genome Atlas (TCGA) [84]. The authors developed a online visualization tool which allows the user to load an MRI scan and visualize the activation of different filters. Through this the model was found to classify lesions with ring enhancement with negative methylation status and tumors with less clearly defined borders and heterogeneous texture with positive methylation status.

Lao *et al.* [57] constructed a model for survival prediction of patients with GBM based on deep features and hand-crafted features extracted from MRI. to reduce the number of features used, feature selection was done using feature importance methods to find features that were robust to tumor segmentation uncertainty, highly predictive and non-redundant. Survival prediction was also performed

using histological samples and genomic data [58] with validation of produced saliency maps by expert pathologists.

4.5. Other Pathologies

Other oncological pathologies have been showed interested in interpretability using different modalities of data (not exclusively image). Researchers that applied DL techniques on data of multiple pathologies have seeked to interpret them using feature importance. For example, Ahn *et al.* [59] trained a network for malignancy diagnosis based on gene-expression data from multiple tissues and by computing the feature importance of individual genes on the diagnosis found a sub-group suspected to be oncogene-addicted as an individual gene contribute extensively in the classification. Similarly, Yousefi *et al.* [60] proposed a model for the survival prediction based on clinical, gene-expression and protein-expression data of multiple tissues and computed the sensitivity of each feature on the survival risk, identifying that TGF-Beta 1 signaling and epithelialmesenchymal transition (EMT) gene sets are associated with poor prognosis. Oni *et al.* [61] diagnosed eight different cancer types from RNA-seq expression and single nucleotide variation (SNV) data. To explain its decisions, a linear surrogate model [82] was extracted, where its coefficient's magnitude corresponded to importance of the genes in the prediction. The location and variability of explanations were visualized using 2D embeddings of the RNA-seq input data. They found genes related to cell proliferation and tumor growth were important for the diagnosis.

In the diagnosis of early Barrett's Neoplasia using Volumetric Laser Endomicroscopy [62], saliency maps [79] focused on the glands located around the first layers of the esophagus in high-grade dysplasia cases, and on homogeneous esophagus layers in non-dysplastic Barrett's esophagus cases. Garcia-Peraza-Herrera *et al.* [63] extended the same saliency map method to interpret the diagnose esophageal cancer based on endoscopic images. By computing saliency maps of different resolutions they were able to detect unhealthy patterns and diseased tissue.

Korbar *et al.* [64] interpreted the diagnosis of colorectal polyps based on histological images using saliency maps[79, 72] and found that by adding a boundary box around them increased their similarity with pathologists' segmentations.

Inglese *et al.*[65] used DL techniques to find a high-level representation of mass spectrometry imaging data from colorectal adenocarcinoma biopsies. The features extracted from the network was visualized in two dimensions using t-SNE [85] unveiling clusters with different chemical and biological interactions occurring.

Zhang *et al.* [66] developed a diagnostic system of ophtalmic images that explained the diagnosis with sub-tasks. In addition to the diagnosis disease, the network segmented important anatomical regions, and detected other illnesses. The results show an accuracy of 93% on the diagnosis, localization accuracy of the foci of 82% in normal lighted images and 90% in fluorescein sodium eye drops.

Zhang *et al.* [67] proposed a system for diagnosing the malignancy of thyroid nodules on ultrasound with performance comparable with radiologists. The network provides prediction on medical concepts based on the TI-RADS lexicon.

The automatic generation of text reports based on medical imaging system is also an active research area. Zhang *et al.* [68] presented network trained on H&E patches for the malignancy diagnosis of bladder cancer, and conditioned a RNN-based language model to generate text descriptions and visual attention (i.e. saliency maps) highlighting regions of the image relevant for specific parts of the text (Figure 14). Similarly, TieNet [69] provided the same explanation for the network which diagnoses diseases based on chest radiographs and generates text descriptions with similar visual attention. MDNet [70] establishes a relationship between histological images of bladder cancer and diagnostic reports to generate text descriptions and provide visual attention for specific parts of the text.

5. Open Issues and Promising Research Directions

As deep learning grow in popularity, so will the need for interpretability in the dichotomy between machine learning and medical practice. From this survey, it becomes clear that are three premising directions to be explored in future research: lack of evaluation metrics and unreliability of some interpretability methods, focusing almost exclusively on cancer diagnosis and confining studies to frequent diagnosed cancer diseases.

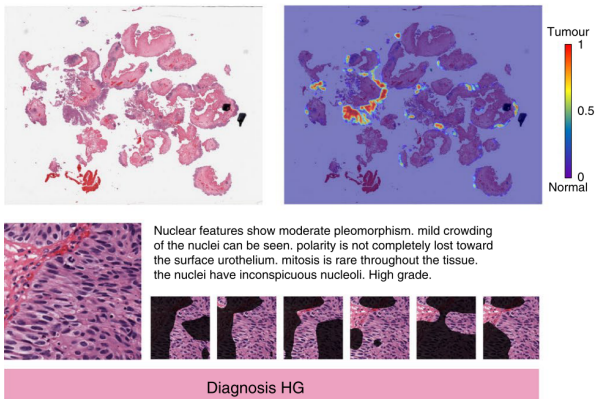

Figure 14: At the left a representative H&E stained whole-slide tissue image and the right the saliency map generated by the method. At the bottom is the generated description for the image and feature-aware attention maps. Adapted from [68].

5.1. Lack of evaluation metrics and unreliability of some interpretability methods

One point of concern is the unreliability of many techniques used to generate saliency maps [86, 87], which might not be representative of the behavior of the model they are trying to explain [88]. Another concern is the susceptibility of neural networks to adversarial attacks, where imperceptible changes to an image can make the network make radical different predictions [89]. More recently, explanations have also been found to be susceptible to adversarial attacks [90], by making imperceptible changes that have no affect on the prediction but alter their explanations.

To quantitatively evaluate an interpretability method without the validation of an expert requires a formal definition of interpretability and the use of a proxy metric describing the quality of the explanation [13]. The lack of ground-truth explanations, such as expert annotated tumor segmentations which indicated what the expected value of a saliency map should be, makes it difficult to make quantitative analysis of the results, hence most studies conduct evaluation by letting experts (e.g. pathologist) compare the explanations of few number of selected examples and their domain knowledge.

Future research should help find interpretability metrics able to assess methods in a number of factors. First, evaluate how faithful are explanations to the actual model's behavior, avoiding adversarial attacks. Also, help understand their uses cases, advantages and weaknesses so to help to choose the appropriate method given a situation, Only by satisfying this requirements can interpretability methods be trusted.

5.2. Focusing almost exclusively on cancer diagnosis

72% of the reviewed studies focus on the diagnosis of cancer diseases, leaving many important tasks which deep learning models have shown to do well still unexplored. The prognosis of a patient with cancer is an estimate of the likely outcome of the disease, whether it will be treated successfully and the patient will recover. Regarding this topic, many DL techniques have been applied for survival prediction and readmission prediction [91] as well as predicting the response to the treatment [92]. Tumor segmentation is a time-consuming and difficult task even for radiotherapy experts, and even though automatic segmentation by DL techniques is possible [93, 94], they interpretability is still low. Recently, DL approaches also has seen success in restoring medical images corrupted with noise or artifacts. The fact that DL hides the reasoning behind this process as also been pointed out as a challenge [95].

Future research should focus on the above tasks as there is many uses cances for interpretability methods, such helping to audit the AI system, finding bias in the data and assisting clinicians reducing their burden when performing the tasks. Furthermore, though studies on interpretability methods have focused mostly on imaging, researchers have begin to combine different modalities of patient data (e.g, imaging, gene information), which result in better performance than single modality data in some cases [96]. Future research should study interpretability methods capable of interpreting DL models training with such heterogeneous data and find relations in it.

5.3. Confining studies to frequent diagnosed cancer diseases

Most studies cover (72%) the application of interpretability methods for DL systems target at the most common cancer diseases, namely breast, skin, lung and brain cancer. In addition to the high number of cases of these diseases, the high number of studies is also due to the proliferation of well-curated public cancer datasets such as TCIA [83], TCGA [84], CBIS-DDSM [97] among others. Also,

due to the creation of special issues, workshops and challenges which focus on interpretability of ML in healthcare, such as the iMIMIC workshop [98] and the ISIC Melanoma Challenge [99] and the BraTS Challenge [100].

In our opinion, there exists a great opportunity to grow research on less common diseases by curating larger datasets and creating challenges directed at increasing the interest.

6. Conclusions

Interpretability of deep learning is a growing field with mostly open problems and many opportunities for the field of medicine and oncology.

The lack of interpretability in deep learning has been pointed out as a major problem by many researchers that have studied the application of deep learning in various areas of medicine and bioinformatics [27, 101, 28].

In this work, we introduced important concepts in this topic and review the related research on the application of interpretability methods for cancer diseases, summarizing their main conclusions.

To the extend of the author's knowledge, such comprehensive review on the interpretability of DL models for cancer diseases has not been previously performed. As discussed in the previous section, as only a small number of interpretability methods have been extended for cancer diseases, future research should extend this methods. Furthermore, we identified a focus on medical imaging and common cancer diseases, namely breast, skin, lung and brain cancer. In the future, research should expand to other modalities and cancer diseases. Also, as AI systems are beginning to take advantage of data from multiple sources (e.g. imaging, gene information, etc.), new interpretability methods must be developed to interpret them. Lastly, future research in the design of evaluation metrics and frameworks is mandatory to assess the reliability of AI systems and for increasing the trust to be used on clinical practice.

7. Conflict of interest

The authors declare no competing interests.

References

[1] FDA Approvals For Smart Algorithms In Medicine In One Giant Infographic, https://medicalfuturist.com/fda-approvals-for-algorithms-in-medicine/, accessed: 2019-12-01.

[2] W. Penny, D. Frost, Neural networks in clinical medicine, Medical decision making : an international journal of the Society for Medical Decision Making 16 (1996) 386–98. doi:10.1177/0272989X9601600409.

[3] F. Rosenblatt, The Perceptron, a Perceiving and Recognizing Automaton Project Para, Report: Cornell Aeronautical Laboratory, Cornell Aeronautical Laboratory, 1957.

[4] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, Schilling, et al, A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task, European Journal of Cancer 111 (2019) 148–154. doi:10.1016/j.ejca.2019.02.005.

[5] B. E. Bejnordi, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, Journal of the American Medical Association (JAMA) 318 (22) (2017) 2199–2210. doi:10.1001/jama.2017.14585.

[6] M. Reyes, R. Meier, S. Pereira, C. Silva, P. M. Dahlweid, MD, H. Tengg-Kobligk, R. Summers, R. Wiest, On the interpretability of artificial intelligence in radiology: Challenges and opportunities, Radiology: Artificial Intelligence 2 (2020) e190043. doi:10.1148/ryai.2020190043.

[7] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Visual Informatics 1 (1) (2017) 48–56. arXiv:1702.01226, doi:10.1016/j.visinf.2017.01.006.

[8] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536. doi:10.1038/323533a0.

[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2323. doi:10.1109/5.726791.

[10] Y. Bengio, Deep Learning of Representations for Unsupervised and Transfer Learning, in: JMLR: Workshop and Conference Proceedings, Vol. 7, 2011, pp. 1–20. doi:10.1109/IJCNN.2011.6033302.

[11] P. Baldi, Autoencoders, unsupervised learning and deep architectures, in: Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27, UTLW'11, JMLR.org, 2011, p. 37–50.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys 51 (5) (2018) 1–42. doi:10.1145/3236009.

[13] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, ArXiv e-printsarXiv:1702.08608.

[14] C. Molnar, Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/ (2019).

[15] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. Müller, How to Explain Individual Classification Decisions, Journal of Machine Learning Research 11 (2010) 1803–1831.

[16] S. Bazen, X. Joutard, The taylor decomposition: A unified generalization of the oaxaca method to nonlin-

ear models (05 2013).

[17] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLOS ONE 10 (7) (2015) 1–46. doi:10.1371/journal.pone.0130140.

[18] H. Lee, S. T. Kim, Y. M. Ro, Generation of Multimodal Justification Using Visual Word Constraint Model for Explainable Computer-Aided Diagnosis, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11797 LNCS (2019) 21–29. arXiv:1906.03922, doi:10.1007/978-3-030-33850-3_3.
URL http://arxiv.org/abs/1906.03922

[19] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding Neural Networks Through Deep Visualization, in: International Conference on Machine Learning - Deep Learning Workshop, 2015, p. 12.

[20] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, ArXiv e-printsarXiv:1802.03426.

[21] V. A. V. Y. J. P. C. D. E. S. I. M. F. C. J. K.-C. H. M. Mara Graziani, James M. Brown, Improved interpretability for computer-aided severity assessment of retinopathy of prematurity, in: Proceedings of Computer-Aided Diagnosis, 2019, p. 63. doi:10.1117/12.2512584.
URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10950/2512584/Improved-interpretability-for-computer-aided-severity-assessment-of-retinopathy-of/10.1117/12.2512584.full

[22] S. Hsieh, C. D. Lehman, V. Dialani, B. Zhou, D. Peck, G. Patterson, L. Mackey, V. Syrgkanis, J. Wu, Expert identification of visual primitives used by CNNs during mammogram classification, in: Proceedings of Computer-Aided Diagnosis, 2018, p. 100. arXiv:1803.04858, doi:10.1117/12.2293890.

[23] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., 2016, pp. 3395–3403.

[24] C. Olah, A. Mordvintsev, L. Schubert, Feature Visualization, DistillHttps://distill.pub/2017/feature-visualization. doi:10.23915/distill.00007.

[25] R. Caruana, H. Kangarloo, J. Dionisio, U. Sinha, D. Johnson, Case-based explanation of non-case-based learning methods, in: Proceedings of the AMIA Symposium, 1999, pp. 212–5.

[26] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples (2014). arXiv:1412.6572.

[27] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Computing and Applications 0123456789 (2019) 1–15. doi:10.1007/s00521-019-04051-w.
URL https://doi.org/10.1007/s00521-019-04051-w

[28] F. Cabitza, R. Rasoini, G. F. Gensini, Unintended Consequences of Machine Learning in Medicine, JAMA 318 (6) (2017) 517–518. doi:10.1001/jama.2017.7797.
URL https://doi.org/10.1001/jama.2017.7797

[29] M. Graziani, V. Andrearczyk, H. Müller, Visual interpretability for patch-based classification of breast cancer histopathology images, in: Proceedings of Medical Imaging with Deep Learning, 2018, pp. 1–4.

[30] M. Graziani, V. Andrearczyk, H. Möuller, Regression Concept Vectors for Bidirectional Explanations in Histopathology, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer International Publishing, 2018, pp. 124–132.

[31] M. Graziani, V. Andrearczyk, S. Marchand-maillet, H. Müller, Concept attribution: Explaining CNN decisions to physicians, Computers in Biology and Medicine 123 (2020) 103865. doi:10.1016/j.compbiomed.2020.103865.
URL https://doi.org/10.1016/j.compbiomed.2020.103865

[32] S. T. Kim, J. H. Lee, H. Lee, Y. M. Ro, Visually interpretable deep network for diagnosis of breast masses on mammograms, Physics in Medicine and Biology 63 (23). doi:10.1088/1361-6560/aaef0a.

[33] N. Antropova, B. Huynh, M. Giger, A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets, Medical Physics 44. doi:10.1002/mp.12453.

[34] D. Alvarez-Melis, T. Jaakkola, Towards Robust Interpretability with Self-Explaining Neural Networks, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, pp. 7775–7784.

[35] J. P. Amorim, I. Domingues, P. Abreu, J. Santos, Interpreting deep learning models for ordinal problems, in: ESANN, 2018, pp. 373–377.

[36] G. Bologna, Y. Hayashi, Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning, Journal of Artificial Intelligence and Soft Computing Research 7 (4) (2017) 265–286. doi:10.1515/jaiscr-2017-0019.

[37] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, K. Huang, Salmon: Survival analysis learning with multi-omics neural networks on breast cancer, Frontiers in Genetics 10 (4) (2019) 1–13. doi:10.3389/fgene.2019.00166.

[38] F. M. Alakwaa, K. Chaudhary, L. X. Garmire, B. G. Program, Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data, Journal of Proteome Research (2018) 337–347doi:10.1021/acs.jproteome.7b00595.

[39] Q. Liu, P. Hu, Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer, Cancers 11 (4). doi:10.3390/cancers11040494.

[40] P. Van Molle, M. De Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, B. Dhoedt, Visualizing Convolutional Neural Networks to Improve Decision Support for Skin Lesion Classification, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer International Publishing, 2018, pp. 115–123.

[41] A. Cruz-Roa, et al., Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks, in: Medical Imaging: Digi-

tal Pathology, Vol. 9041, International Society for Optics and Photonics, 2014, p. 904103. doi:10.1117/12.2043872.

[42] A. Radhakrishnan, C. Durham, A. Soylemezoglu, C. Uhler, Patchnet: Interpretable Neural Networks for Image Classification, ArXiv e-printsarXiv:1705.08078.

[43] M. Paschali, M. Ferjad Naeem, W. Simson, K. Steiger, M. Mollenhauer, N. Navab, M. F. Naeem, W. Simson, K. Steiger, M. Mollenhauer, N. Navab, Deep Learning Under the Microscope: Improving the Interpretability of Medical Imaging Neural Networks, ArXiv e-prints (2019) 1–9arXiv:1904.03127.
URL http://arxiv.org/abs/1904.03127

[44] I. Gonzalez Diaz, DermaKNet: Incorporating the knowledge of dermatologists to Convolutional Neural Networks for skin lesion diagnosis, IEEE Journal of Biomedical and Health Informatics 2194 (2018) 1–14. doi:10.1109/JBHI.2018.2806962.

[45] A. A. Cruz-Roa, J. E. Arevalo Ovalle, A. Madabhushi, F. A. González Osorio, A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer Berlin Heidelberg, 2013, pp. 403–410.

[46] M. Sadeghi, P. K. Chilana, M. S. Atkins, How Users Perceive Content-Based Image Retrieval for Identifying Skin Images, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer International Publishing, 2018, pp. 141–148.

[47] W. Silva, K. Fernandes, M. J. Cardoso, J. S. Cardoso, Towards Complementary Explanations Using Deep Neural Networks, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer International Publishing, 2018, pp. 133–140.

[48] W. Silva, K. Fernandes, J. S. Cardoso, How to produce complementary explanations using an ensemble model, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.

[49] N. C. F. Noel C. F. Codella, C.-C. Chung-Ching Lin, A. Halpern, M. Hind, R. Feris, J. R. Smith, Collaborative Human-AI (CHAI): Evidence-Based Interpretable Melanoma Classification in Dermoscopic Images, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer International Publishing, 2018, pp. 97–105.

[50] P. Rajpurkar, el al., Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLOS Medicine 15 (11) (2018) 1–17. doi:10.1371/journal.pmed.1002686.

[51] R. Paul, Y. Liu, Q. Li, L. Hall, D. Goldgof, Representation of Deep Features using Radiologist defined Semantic Features, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1429–1435. doi:10.1109/IJCNN.2018.8489440.

[52] S. Shen, S. X. Han, D. R. Aberle, A. A. T. Bui, W. Hsu, An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification, ArXiv e-printsarXiv:1806.00712.

[53] S. Cui, Y. Luo, H.-H. Tseng, R. Ten Haken, I. El Naqa, Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage, Medical Physics 46. doi:10.1002/mp.13497.

[54] S. Pereira, R. Meier, V. Alves, M. Reyes, C. A. Silva, Automatic Brain Tumor Grading from MRI Data Using Convolutional Neural Networks and Quality Assessment, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer International Publishing, 2018, pp. 106–114.

[55] S. Pereira, R. Meier, R. McKinley, R. Wiest, V. Alves, C. A. Silva, M. Reyes, Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation, Medical Image Analysis 44 (2018) 228–244. doi:10.1016/j.media.2017.12.009.

[56] L. Han, M. R. Kamdar, MRI to MGMT: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks, Analytical Chemistry 25 (4) (2015) 368–379. doi:10.1016/j.cogdev.2010.08.003.Personal.

[57] J. Lao, Y. Chen, Z. C. Li, Q. Li, J. Zhang, J. Liu, G. Zhai, A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme, Scientific Reports 7 (1). doi:10.1038/s41598-017-10649-8.

[58] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, L. A. D. Cooper, Predicting cancer outcomes from histology and genomics using convolutional networks, Proceedings of the National Academy of Sciences 115 (13) (2018) E2970–E2979. doi:10.1073/pnas.1717139115.

[59] T. Ahn, T. Goo, C. H. Lee, S. Kim, K. Han, S. Park, T. Park, Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data, Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2019) 1748–1752doi:10.1109/BIBM.2018.8621108.

[60] S. Yousefi, F. Amrollahi, M. Amgad, C. Dong, J. E. Lewis, C. Song, D. A. Gutman, S. H. Halani, J. Enrique Velazquez Vega, D. J. Brat, L. A. D. Cooper, Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models, Scientific Reports 7 (2017) 11707. doi:10.1038/s41598-017-11817-6.

[61] O. Oni, S. Qiao, Model-Agnostic Interpretation of Cancer Classification with Multi-Platform Genomic Data, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '19, ACM, New York, NY, USA, 2019, pp. 34–41. doi:10.1145/3307339.3342189.

[62] R. Fonollà, et al., Ensemble of Deep Convolutional Neural Networks for Classification of Early Barrett's Neoplasia Using Volumetric Laser Endomicroscopy, Applied Sciences 9 (11) (2019) 2183. doi:10.3390/app9112183.

[63] L. C. Garcia-Peraza-Herrera, M. Everson, W. Li, I. Luengo, L. Berger, O. Ahmad, L. Lovat, H.-P. Wang, W.-L. Wang, R. Haidry, D. Stoyanov, T. Vercauteren, S. Ourselin, Interpretable Fully Convolutional Classification of Intrapapillary Capillary Loops for Real-Time Detection of Early Squamous Neoplasia, ArXiv e-prints (2018) 1–8arXiv:1805.00632.

[64] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, S. Hassanpour, Looking Under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 821–827. doi:10.1109/CVPRW.2017.114.

[65] P. Inglese, et al., Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer, Chemical Science (2017) 3500–3511doi:10.1039/c6sc03738k.

[66] K. Zhang, X. Liu, F. Liu, L. He, L. Zhang, Y. Yang, W. Li, S. Wang, L. Liu, Z. Liu, X. Wu, H. Lin, An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study, Journal of Medical Internet Research 20 (11) (2018) 1–13. doi:10.2196/11144.
URL http://www.ncbi.nlm.nih.gov/pubmed/30429111http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6301833

[67] S. Zhang, Y. Luo, H. Du, Z. Jin, Y. Zhu, Y. Zhang, F. Xie, M. Zhang, X. Tian, J. Zhang, A Novel Interpretable Computer-Aided Diagnosis System of Thyroid Nodules on Ultrasound based on Clinical Experience, IEEE Access (2020) 1–1doi:10.1109/ACCESS.2020.2976495.
URL https://ieeexplore.ieee.org/document/9016204/

[68] Z. Zhang, P. Chen, M. Mcgough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F. K. Khalil, S. I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, L. Yang, Diagnosis With Deep Learning, Nature Machine Intelligence 1 (May). doi:10.1038/s42256-019-0052-1.

[69] X. Wang, Y. Peng, L. Lu, Z. Lu, R. M. Summers, TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2018) 9049–9058arXiv:1801.04334, doi:10.1109/CVPR.2018.00943.

[70] Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3549–3557.

[71] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography, Scientific Reports 9 (1) (2019) 12495. doi:10.1038/s41598-019-48995-4.
URL https://doi.org/10.1038/s41598-019-48995-4

[72] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[73] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua (2017) 3319–3327. arXiv:1704.05796, doi:10.1109/CVPR.2017.354.

[74] A. A. Kabbani, Y. Weerakkody, et. al., Breast imaging-reporting and data system (BI-RADS), Reston VA: American College of Radiology.

[75] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ArXiv e-printsarXiv:1711.11279v5.

[76] S. T. Kim, J.-H. Lee, Y. Ro, Visual evidence for interpreting diagnostic decision of deep neural network in computer-aided diagnosis, in: Proceedings of Computer-Aided Diagnosis, 2019, p. 19. doi:10.1117/12.2512621.

[77] H2O.ai, https://www.h2o.ai/, accessed: 2019-07-01.

[78] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118. doi:10.1038/nature21056.

[79] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929. doi:10.1109/CVPR.2016.319.

[80] Y. Luo, H.-H. Tseng, S. Cui, L. Wei, R. K. Ten Haken, I. El Naqa, Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling, BJR—Open 1 (1) (2019) 20190021. doi:10.1259/bjro.20190021.
URL https://www.birpublications.org/doi/10.1259/bjro.20190021

[81] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for Simplicity: The All Convolutional Net, ArXiv e-printsarXiv:1412.6806.

[82] M. T. Ribeiro, S. Singh, C. 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, ACM, New York, NY, USA, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[83] K. Clark, et al., The cancer imaging archive (TCIA): Maintaining and operating a public information repository, Journal of Digital Imaging 26 (6) (2013) 1045–1057. doi:10.1007/s10278-013-9622-7.

[84] The cancer genome atlas, https://www.cancer.gov/tcga, accessed: 2019-07-01.

[85] L. Van Der Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

[86] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, B. Kim, The (Un)reliability of Saliency Methods, Springer International Publishing, 2019, Ch. 4, pp. 267–280. doi:10.1007/978-3-030-28954-6_14.

[87] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, pp. 9505–9515.

[88] J. Fauw, J. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, O. Ronneberger, Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature

Medicine 24. doi:10.1038/s41591-018-0107-6.

[89] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, N. K. Jha, Systematic poisoning attacks on and defenses for machine learning in healthcare, IEEE Journal of Biomedical and Health Informatics 19 (6) (2015) 1893–1905. doi:10.1109/JBHI.2014.2344095.

[90] A. Ghorbani, A. Abid, J. Zou, Interpretation of Neural Networks Is Fragile, Proceedings of the AAAI Conference on Artificial Intelligence 33 (Lipton 2016) (2019) 3681–3688. arXiv:1710.10547, doi:10.1609/aaai.v33i01.33013681.

[91] W. Zhu, L. Xie, J. Han, X. Guo, The application of deep learning in cancer prognosis prediction, Cancers 12 (3) (2020) 603.

[92] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, H. J. Aerts, Deep learning predicts lung cancer treatment response from serial medical imaging, Clinical Cancer Research 25 (11) (2019) 3266–3275. arXiv:https://clincancerres.aacrjournals.org/content/25/11/3266.full.pdf, doi:10.1158/1078-0432.CCR-18-2495.
URL https://clincancerres.aacrjournals.org/content/25/11/3266

[93] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, Medical Image Analysis 35 (2017) 18–31. doi:10.1016/j.media.2016.05.004.
URL http://dx.doi.org/10.1016/j.media.2016.05.004

[94] M. Mittal, L. M. Goyal, S. Kaur, I. Kaur, A. Verma, D. J. Hemanth, Deep learning based enhanced tumor segmentation approach for mr brain images, Appl. Soft Comput. 78 (2019) 346–354.

[95] H.-M. Zhang, B. Dong, A review on deep learning in medical image reconstruction, Journal of the Operations Research Society of Chinadoi:10.1007/s40305-019-00287-4.
URL https://doi.org/10.1007/s40305-019-00287-4

[96] D. Sun, M. Wang, A. Li, A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data, IEEE/ACM Transactions on Computational Biology and Bioinformatics 16 (03) (2019) 841–850. doi:10.1109/TCBB.2018.2806438.

[97] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, D. L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Scientific Data 4 (1) (2017) 170177. doi:10.1038/sdata.2017.177.
URL https://doi.org/10.1038/sdata.2017.177

[98] iMIMIC – Interpretability of Machine Intelligence in Medical Image Computing, http://imimic-workshop.com/, accessed: 2019-07-01.

[99] ISIC Archive, https://www.isic-archive.com/, accessed: 2019-07-01.

[100] Multimodal Brain Tumor Segmentation Challenge 2019, https://www.med.upenn.edu/cbica/brats2019.html, accessed: 2019-12-01.

[101] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu, B. Lo, G.-Z. Yang, Deep learning for health informatics, IEEE journal of biomedical and health informatics PP. doi:10.1109/JBHI.2016.2636665.