Interpreting Deep Machine Learning Models: An Easy Guide for Oncologists

José P. Amorim, Pedro H. Abreu, Alberto Fernández, Mauricio Reyes, João Santos, and Miguel H. Abreu

(Clinical Application Review)

***Abstract*—Healthcare agents, in particular in the oncology field, are currently collecting vast amounts of diverse patient data. In this context, some decision-support systems, mostly based on deep learning techniques, have already been approved for clinical purposes. Despite all the efforts in introducing artificial intelligence methods in the workflow of clinicians, its lack of interpretability - understand how the methods make decisions - still inhibits their dissemination in clinical practice. The aim of this article is to present an easy guide for oncologists explaining how these methods make decisions and illustrating the strategies to explain them. Theoretical concepts were illustrated based on oncological examples and a literature review of research works was performed from PubMed between January 2014 to September 2020, using "deep learning techniques," "interpretability" and "oncology" as keywords. Overall, more than 60% are related to breast, skin or brain cancers and the majority focused on explaining the importance of tumor characteristics (e.g. dimension, shape) in the predictions. The most used computational methods are multilayer perceptrons and convolutional neural networks. Nevertheless, despite being successfully applied in different cancers scenarios, endowing deep learning techniques with interpretability, while maintaining their performance, continues to be one of the greatest challenges of artificial intelligence.**

***Index Terms*—Big Data, interpretability, deep learning, decision-support systems, oncology.**

I. Introduction

TODAY, in healthcare scenarios, we are living in 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 application. During training, the algorithm builds a model which fits the data received as input, while in application, 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 US Food and Drug Administration approved 46 ML algorithms [1] for clinical purposes 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. When this target variable is discrete we are present to a classification problem (benign or malignant), or a regression problem in case the variable is continuous (overall survival - measured in months).

Among different ML paradigms that are used in medical contexts, 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

Manuscript received 26 May 2021; revised 26 September 2021; accepted 15 November 2021. Date of publication 30 November 2021; date of current version 6 January 2023. This work was supported in part by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project CISUC - UID/CEC/00326/2020, in part by European Social Fund, through the Regional Operational Program Centro 2020, and in part by the FCT Research under Grant SFRH/BD/136786/2018. *(Corresponding author: José P. Amorim.)*

José P. Amorim is with the Department of Informatics Engineering, University of Coimbra, CISUC, 3030-290 Coimbra, Portugal, and also with IPO-Porto Research Center - Portuguese Institute of Oncology of Porto, 4200-072 Porto, Portugal (e-mail: jpamorim@dei.uc.pt).

Pedro H. Abreu is with the Department of Informatics Engineering, University of Coimbra, CISUC, 3030-790 Coimbra, Portugal (e-mail: pha@dei.uc.pt).

Alberto Fernández is with the DaSCI Andalusian Research Institute, University of Granada, 18071 Granada, Spain (e-mail: alberto@decsai.ugr.es).

Mauricio Reyes is with Insel Data Science Center, Inselspital, Bern University Hospital, 3010 Bern, Switzerland, and also with the ARTORG Center for Biomedical Research, University of Bern, 3008 Bern, Switzerland (e-mail: mauricio.reyes@med.unibe.ch).

João Santos is with IPO-Porto Research Center – Portuguese Institute of Oncology of Porto, 4200-072 Porto, Portugal, and also with the ICBAS - Instituto de Ciências Biomédicas Abel Salazar, 4050-313 Porto, Portugal (e-mail: joao.santos@ipoporto.min-saude.pt).

Miguel H. Abreu is with the Department of Medical Oncology, Portuguese Oncology Institute of Porto, 4200-072 Porto, Portugal (e-mail: p_abreu@sapo.pt).

Digital Object Identifier 10.1109/RBME.2021.3131358

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

Due to the increasing computational power, the complexity of these networks has substantially grown, materializing 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 makes it difficult to interpret their decisions and prevents their dissemination in clinical practice.

The objective of this study is to present an easy guide for oncologists explaining how DL techniques make decisions and illustrating the strategies that can be used in the oncological field to explain them, as it is an essential step towards the integration of DL in the workflow of physicians in the field of oncology. To better illustrate these strategies to oncologists and other healthcare agents we give self-explanatory oncological examples. Other reviews already covered specific medical areas such as radiology [6] which only equate to a small set of image modalities and does not cover other patient data such as genomic data. Others expand the review to the medical field but do not focus on DL techniques [7]. 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 September 2020, searching individually and in combination search terms such as “interpretability,” “deep learning,” “oncology”, “cancer” and “decision support systems”.

Overall, from this selection, more than 60% are related to breast, skin or brain cancers and the majority focused on explaining the importance of tumor characteristics (e.g. dimension, shape) in the disease behavior prediction. Among the DL techniques used in the oncology field which were interpreted, the majority are multilayer perceptons and convolutional neural networks. In this study we also 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 most prevalent diseases and also the dissemination of well curated datasets and challenges target at those diseases. Overall, most works focus on the validation of the knowledge acquired by the DL model for the diagnosis of malignancy or detection of a cancer disease.

Despite being successfully applied in different cancer scenarios, endowing deep learning techniques with the ability to explain their predictions, while maintaining their exceptional performance, will continue to be one of the greatest challenges faced by artificial intelligence. Future work includes the extension of interpretability methods for debugging model misbehavior and acquire new knowledge about disease, as well as largely overlooked cancer tasks such as tumor segmentation and image registration. Also, the evaluation of interpretability methods so that they can be compared and validated.

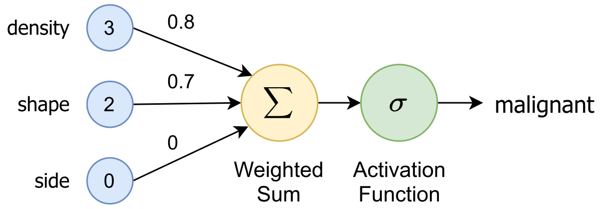


Fig. 1. The Perceptron computes the weighted sum of the breast cancer tumor input variables, and an activation function turns the output into a binary prediction of malignancy.

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 (benign - 0 or malignant - 1). In other cases, examples from actual DL works in the field of oncology will be used to illustrate the techniques.

II. ANN Techniques Overview

Artificial Neural Networks (ANN) are a set of algorithms, inspired by the human brain, which are used to approximate unknown functions. They are sometimes called “universal approximators,” because they can learn to approximate mappings between any input x and any output y, assuming they are correlated. ANNs are composed of layers of neurons, which combine input from the data with a set of coefficients, or weights, assigning significance to inputs with regard to the output label.

Perceptron: The Perceptron [8] is the the precursor to the ANN techniques. In this binary classification algorithm, the linear predictor chooses to “fire” based on a function combining a set of weights with the input vector.

Training process: As seen in Fig. 1, 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 that will use the weighted sum of the input variables 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 Fig. 1, the perceptron is given the breast cancer tumor variables density, shape, and side and given the weights obtained during training (0.8, 0.7 and 0 respectively), predicts the tumor to be malignant.

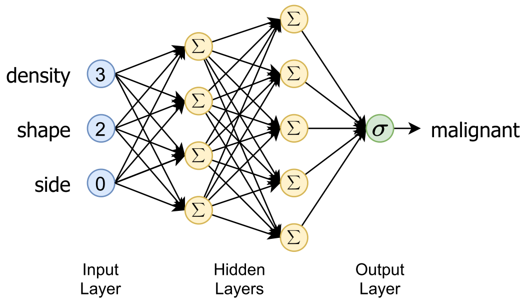


Fig. 2. A Multilayer Perceptron (MLP) composed by an input layer, an output layer and two hidden layers similar to perceptrons that predict malignancy based on breast cancer tumor variables.

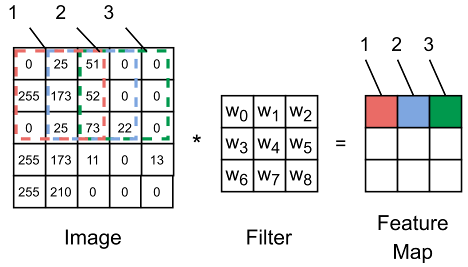
Multilayer Perceptron: The Multilayer Perceptron (MLP) [8] is the natural extension of the perceptron to solve more complex problems. Rather than having a single unit, or neuron, the MLP has multiple layers with multiple neurons each, as can be seen in Fig. 2. Also, the linear activation function of the perceptron is replaced by a non-linear activation function which helps to solve non-linear problems. 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 present in the hidden layers acts like a perceptron, performing the weighted combination of its inputs and applying a non-linear activation function. The output of activations function of each neuron, also known as activation, acts as input for the neurons of the next layer. The combination of activations of the last intermediate layer produces a desired output (y).

MLPs have been explored on multiple public datasets for breast cancer diagnosis based on tumor characteristics such as density, shape, and side with high accuracy (>97%). Fig. 2 illustrates the approach used in [9] based on the public Wisconsin Breast Cancer dataset. In the example, given the tumor variables (density, shape, and side) the model learns to optimal weight's values during training, to predict the malignancy. Due to their 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.

Convolutional Neural Networks: Convolutional Neural Networks (CNN) [10], [11] techniques emerged as a solution to addresses the previous computational problem.

Training process: CNNs treat the image as a matrix (Fig. 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 a 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.



0	25	51	0	0
255	173	52	0	0
0	25	73	22	0
255	173	11	0	13
255	210	0	0	0

w_0	w_1	w_2
w_3	w_4	w_5
w_6	w_7	w_8

Fig. 3. The convolution operation produces a feature map, where each element is the result of the element-wise multiplication between the region of the image and the filter (shown in the same shade).

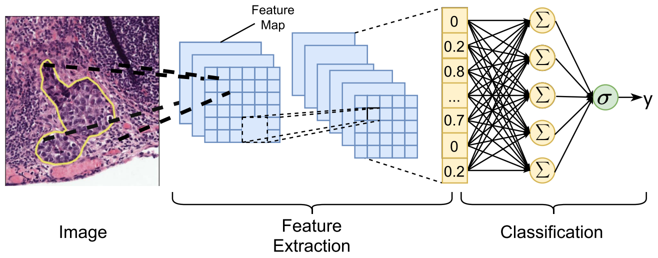


Fig. 4. Representation of Convolutional Neural Network (CNN) used in [5] for the detection of lymph node metastases of breast cancer in histopathological images. First, each convolutional layer produces features maps using the convolution operator across the previous layers' output. The output of the feature extraction is concatenated into a feature vector which serves as input for the classification MLP which predicts the presence of metastases.

During feature extraction, each convolutional layer is composed by n filters resulting in n feature maps. The values of the feature maps of the last convolutional layer are concatenated into a single vector and used as an input for a MLP 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.

CNNs were used for example in the context of detection of lymph node metastases of breast cancer based on whole-slide images of digitally scanned tissue sections of over two hundred patients [5]. Fig. 4 illustrates the approach which led to a performance comparable with an expert pathologist interpreting the slides. The CNN learns the weights of the filters, and during the feature extraction is able to extract features which may include the color and shape of the nuclei. The features are used to make the classification, which predicts the tissue 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: Recurrent Neural Networks (RNN) techniques solve this issue by having a small network looped for each element of the sequence, allowing information to persist. A simple RNN contains an hidden state, h_t, at time t which depends on the input of the current step t and the state of the previous step.

Training process: RNNs are usually composed of only a layer of neurons, which taking an input (x_i) predicts the output (o_i) in a recurrent way (Fig. 5 a).

This refers to the fact that its processing unit (P) 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

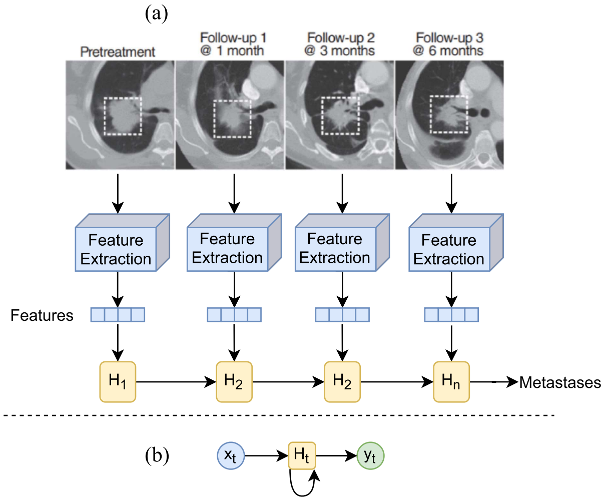


Fig. 5. a) The Recurrent Neural Network (RNN) first extracts set of visual features from CT slides from multiple stages using a CNN [13]. The hidden units optimize their weights to learn useful information from the features and pass stage-specific context sequentially until a final metastases prediction is made. b) Hidden unit (H_t) shared between steps (t) and receives the context of previous CT scan (x_t) and predicts the prognosis (y_t).

sequenced, and the predicted one. As represented by the self-arrow in Fig. 5 b, the processing unit shares information among steps allowing the context and information from each slice to be passed on until a final diagnosis is given (y_t) [12].

In the example provided in Fig. 5 a, the RNN is presented in an unfolded version, where the processing unit is repeated for each step in the sequence. It corresponds to an approach for the treatment prognosis of patients with lung cancer based on Computerized Tomography (CT) of four different stages (pre-treatment, 1 months follow-up, 3 months follow-up and 6 months follow-up) [13]. Outcomes such as survival and metastases were predicted using a RNN based on the a set of features extracted from the CT using a CNN. At each step, and based on the context that is passed from the previous step, it learned to extract and memorize useful context and pass it to subsequent steps until a final prognosis was made.

Autoencoder: The autoencoder [14] is a unsupervised algorithm, which means that unlike the previous supervised algorithms it does not require labelled data in the training process. 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. Depending on the characteristics of the input, the encoder and decoder can have different architectures, some based on multi-layer perceptrons and other on convolutional neural networks.

Training process: The denoising autoencoder (Fig. 6) contains an encoder which receives the noisy input, compresses into a small representation, called code, and is reconstructed by a decoder into the original noiseless input. Due to the small size of the code, the autoencoder learns the distinctive features of the image and learns to ignore random noise. During training, the weights of the neurons present in the encoder and decoder are continuously updated to reduce the difference between the original input and the output, called reconstruction error, to find useful patterns in the data.

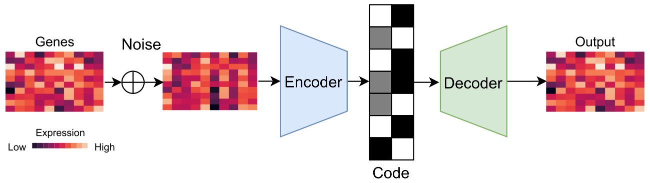


Fig. 6. In a Denoising Autoencoder an encoder transforms a noisy gene expression data into a compressed representation (code) and the decoder transforms the code back into denoised version of original data.

One frequent use of denoising autoencoders is the extraction and compression of relevant features for the detection of genes correlated with the ER status of patients with breast cancer [15]. Fig. 6 illustrates how the autoencoder is given a set of gene expression data with some noise with the task of compressing the data into an relevant representation (code).

III. Interpretability Concepts Overview: Desiderata, Dimensions, and Strategies

The significance of interpretability when developing ML solutions is well-known in academia and corporations. However, there is no consensus upon the definition of interpretability [16]. One of the most used definitions was presented by [17] which defined interpretability as the "ability to explain or to present in understandable terms to a human," and will be used in this work.

A. Desiderata of Interpretability

The demand of interpretability arises due to a mismatch between the objectives of the model and of the users - clinicians and patients. Although DL techniques have 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 to interpret them emerges, especially in healthcare contexts.

In addition to high accuracy of ML algorithms, users have additional desiderata. Doshi-Velez and Kim [17] specified five main desiderata for interpretability:

- *Fairness:* Assure that protected groups (e.g. gender, ethnicity) are not somehow discriminated against (explicit or implicit);
- *Privacy:* Assure that sensitive information is protected;
- *Reliability/Robustness:* Assure high algorithmic performance despite variation of parameter or input;
- *Causality:* Assure that the predicted change in output due to a perturbation will occur in the real system;
- *Trust:* Allow users to trust a system capable of explaining its decisions rather than a black box that just outputs the decision itself.

B. Dimensions of Interpretability

Interpretability methods can be characterized by a set of dimensions [18]: global and local interpretability, intrinsic and post-hoc interpretability and model-specific and model-agnostic interpretability. These will be described in what follows.

a) Global and Local Interpretability: This dimension reflects the scope of interpretability of a model and depicts the portion of predictions that the model can explain. To perform a classification task an ML 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. Global interpretability aims to analyze 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. Local interpretability aims to understand the relationship between the set of input features of a specific case and the model decision.

In our MLP example (Fig. 2), based on the instances provided, the network learned relationships that help predict the tumor malignancy, 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 help understand which relationships the network learned, and for the example of breast side confirm that it was not used. Global interpretability can also help as know if non-random sources of noise which have been not been removed have affected the model's learning (e.g. artifacts). Local interpretability could help understand the importance of the input features in the malignancy prediction of a particular patient.

b) 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. Intrinsic interpretability refers to models which due to their simplicity are interpretable by themselves, such as decision trees or sparse linear models [18]. Complex models can increase their intrinsic interpretability by constraining their complexity or simplifying their behavior. Examples of these constraints are sparsity, monotonicity, adding domain knowledge, or even constraints on the complexity of the network by limiting the number of neurons or layers.

Post-hoc interpretability refers to the application of interpretability methods after the model's training [18]. Post-hoc methods help elucidate how the model works without constraining it.

In our MLP example, we could instead use a short decision tree or a small sparse MLP to achieve intrinsic interpretability or choose to maintain the complexity of the MLP and use a post-hoc method such as feature importance to understand the importance of the input features.

c) Model-specific and Model-agnostic: Another way to classify interpretability methods is based on the dependency 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 [18].

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.

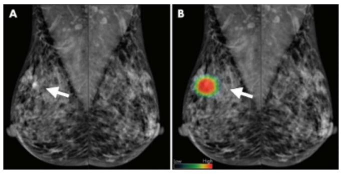


Fig. 7. Example of a saliency map depicting the important pixels for malignancy prediction based on mammograms. Left: ground-truth expert segmentation. Right: saliency map, where the pixel intensity indicates the importance of the pixel in the classification.) [23].

C. Interpretability Strategies

During the training phase, DL algorithms create data-driven models that can be interpreted using different strategies producing different types of explanations. Namely feature importance, saliency map, model visualization, surrogate model, domain knowledge and example-based explanations, which will be introduced next.

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 [19] and decomposition [20], [21].

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. 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.

2) Saliency Map: When dealing with images, saliency maps [17] (or heatmaps) can be used to visually illustrate variations in the importance of different features, using color to convey the weight of pixel in a given prediction.

Similarly to feature importance, the pixel values of saliency maps can be obtained following two main approaches: Backpropagation methods compute the relevance of a pixel by propagating a signal from the output neuron backward through the layers to the input image in a single pass [21]. Perturbation methods compute pixel relevance by making small changes in the pixel value of the input image and compute how the changes affect the prediction [22].

An example of a saliency map, extracted from a CNN trained to predict the malignancy based on mammogram patches is seen in Fig. 7. The red and yellow regions correspond to the most important regions of the image. The method correctly

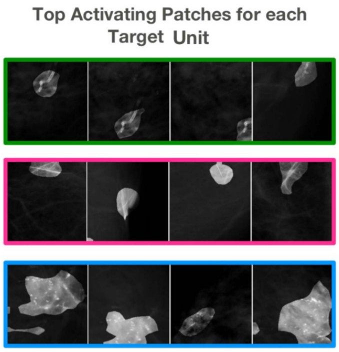


Fig. 8. Illustration of the internal behaviour of a network unit by visualizing regions of mammograms with patterns detected by individual units of the network [27].

Rule 1: IF (density = 'high' or 'equal') and (shape = 'irregular')
THEN **malignant**

Rule 2: IF (density = 'high')
THEN **malignant**

Rule 3: IF (density = 'fat-containing') and (shape = 'irregular')
THEN **benign**

Rule 4: OTHERWISE **benign**

Fig. 9. Rule list extracted from a MLP trained to predict the malignancy of a breast tumor using a surrogate model strategy.

focus on the mass, supporting our confidence in the model's decisions.

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 visualizing patterns detected in an image [24], whereas others help visualizing the feature distribution in the dataset [25], [26]. Also, whereas some strategies may ease to find the image that contains a pattern detected by the network [27], others artificially create images that accentuate the same patterns [28], [29].

In Fig. 8 we can see regions of mammograms which contain patterns detected by individual filters of the CNN trained to diagnose the tumor malignancy.

4) Surrogate Model: Surrogate models are interpretable models that are trained to explain predictions of a black-box model. In the example of oncology, a rule list [30] can be extracted from a network allowing the clinician to understand the knowledge produced by the algorithm. Each rule specifies a condition which when evaluated as true produces on result (benign/malignant in malignancy diagnosis). 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.

While global surrogates models approximate the model in all the input space, local surrogate models approximate single predictions, which makes them more accurate and faithful to the model being explained.

To better understand what is a surrogate model, let's consider the example in Fig. 9, where we can see a rule list extracted from a MLP that demonstrate its decisions. This surrogate model was built by iterating through the MLP neurons and inspecting the connections between the input features and the output label, so that they can be represented by rules. Decision tree is another appropriate type of surrogate model. This method could be seen as an unordered rule list where each leaf is a separate rule where the condition is the labels of the path from the root to the leaf.

TABLE I
ASSOCIATION BETWEEN INTERPRETABILITY STRATEGIES AND DIMENSIONS OF INTERPRETABILITY

	Dimensions	
Strategy	**Scope**	**Intrinsic vs. Post-hoc**
Feature Importance	Local	Post-hoc
Saliency Map	Local	Post-hoc
Model Visualization	Global	Post-hoc
Surrogate Model	Local/Global	Post-hoc
Domain Knowledge	Global	Intrinsic
Example-based	Global	Post-hoc

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.

The introduction of domain knowledge from medical doctors on training can help produce models that resemble how medical doctors diagnose or focus on the features or areas they pay particular attention to [31].

In the case of malignancy diagnosis, domain knowledge can be introduced directly as an input feature, for example a discrete value indicating the shape of the tumor. Domain knowledge can also be used as an additional target variable (e.g. shape, density), besides malignancy, allowing to evaluate how well the model predicts both target variables similarly to how clinicians also take those variables into account.

6) Example-Based Explanation: Example-based explanation methods select examples of the dataset that explain the behavior of the network [18]. 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 [32].

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.

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.

Table I associates the interpretability strategies previously introduced with the dimensions of interpretability, namely scope and intrinsic vs. post-hoc. The dimensions of model specificity vs. agnostic was omitted as it depends on the actual algorithms used and not on the broader interpretability strategy.

IV. 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 [33], [34]. In this section, an overview of interpretability strategies applied to oncological

TABLE II
SUMMARY OF PAPERS REVIEWED

Ref	Disease	Task	Modality	Explanation	Architecture	Dataset
[35]	Breast Cancer	Metastases Detection	WSI H&E	Model Visualization, Saliency Map	CNN	Public
[27]	Breast Cancer	Malignancy Diagnosis	Mammogram	Model Visualization	CNN	Public
[36], [37]	Breast Cancer	Malignancy Diagnosis	WSI H&E	Feature Importance, Domain Knowledge	CNN	Public
[38]	Breast Cancer	Malignancy Diagnosis	Mammogram	Domain Knowledge, Saliency Map	CNN	Public
[39]	Breast Cancer	Malignancy Diagnosis	Mammogram, Ultrasound, MRI	Domain Knowledge	CNN	Public
[40]	Breast Cancer	Malignancy Diagnosis	Mammogram	Saliency Map, Text	CNN + RNN	Public
[41]	Breast Cancer	Malignancy Diagnosis	Hand-crafted	Feature Importance	CNN	Public
[42]	Breast Cancer	Malignancy Diagnosis	Hand-crafted from H&E	Surrogate	MLP	Private
[43]	Breast Cancer	Malignancy Diagnosis	Hand-crafted from H&E	Surrogate	MLP	Public
[44]	Breast Cancer	Survival Prediction	Gene expression, Biomarkers	Feature Importance	MLP	Public
[45]	Breast Cancer	ER+ Prediction	Metabolomics Data	Feature Importance	AE + MLP	Public
[46]	Breast Cancer	Clustering	Gene expression, CNA data	Model Visualization	AE	Public
[47]	Skin Cancer	Malignancy Diagnosis	Dermoscopic images	Model Visualization	CNN	Public
[48]	Skin Cancer	Malignancy Diagnosis	WSI H&E	Saliency Map	CNN	Private
[49]	Skin Cancer	Malignancy Diagnosis	Dermoscopic images	Saliency Map	CNN	Public
[50]	Skin Cancer	Diagnosis of Skin Lesion	WSI H&E	Saliency Map	CNN	Public
[51]	Skin Cancer	Diagnosis of Skin Lesion	Dermoscopic images	Saliency Map	CNN	Public
[52]	Skin Cancer	Malignancy Diagnosis	WSI H&E	Saliency Map	CNN	Public
[53]	Skin Cancer	Diagnosis of Skin Lesion	Dermoscopic images	Example	CNN	Public
[54], [55]	Skin Cancer	Malignancy Diagnosis	Dermoscopic images	Feature Importance, Example, Surrogate	MLP	Public
[56]	Skin Cancer	Diagnosis of Skin Lesion	Dermoscopic images	Example, Saliency Map	CNN	Public
[57]	Lung Cancer	Disease Diagnosis	Chest Radiograph	Saliency Map	CNN	Public
[58]	Lung Cancer	Malignancy Diagnosis	CT	Domain knowledge	CNN	Public
[59]	Lung Cancer	Malignancy Diagnosis	CT	Domain knowledge	CNN	Public
[60]	Lung Cancer	Prognosis Radiation	Biomarker, clinical data	Domain knowledge	AE + MLP	Private
[61]	Brain Cancer	Tumor Grading	MRI	Saliency Map	CNN	Public
[62]	Brain Cancer	Tumor Grading	MRI	Feature Importance, Saliency Map	MLP	Public
[63]	Brain Cancer	Predict Methylation State	MRI	Model Visualization	CNN + RNN	Public
[64]	Brain Cancer	Survival Prediction	MRI	Feature Importance	CNN	Public
[65]	Brain Cancer	Survival Prediction	WSI H&E, Biomarkers	Saliency Map	CNN	Public
[66]	Other	Malignancy Diagnosis	Gene expression	Feature Importance	MLP	Public
[67]	Other	Survival Prediction	Gene and protein expression	Feature Importance	MLP	Public
[68]	Other	Disease Diagnosis	RNA-seq expression, SVN data	Feature Importance, Surrogate	MLP	Private
[69]	Other	Disease Diagnosis	Volumetric Laser Endomicroscopy	Saliency Map	CNN	Private
[70]	Other	Disease Diagnosis	Endoscopic images	Saliency Map	CNN	Public
[71]	Other	Disease Diagnosis	WSI H&E	Saliency Map	CNN	Private
[72]	Other	Disease Diagnosis	DESI	Cluster	AE	Private
[73]	Other	Disease Diagnosis	Ophtalmic images	Domain Knowledge	CNN	Private
[74]	Other	Malignancy Diagnosis	Ultrasound	Domain knowledge	CNN	Private
[75]	Other	Malignancy Diagnosis	WSI H&E	Text, Saliency Map	CNN + RNN	Public
[76]	Other	Disease Diagnosis	Chest Radiograph	Text, Saliency Map, Text	CNN + RNN	Public
[77]	Other	Tumor Grading	WSI H&E	Text, Saliency Map	CNN + RNN	Private

diseases will be presented. The section will be divided into different diseases, namely breast cancer, skin cancer, lung cancer, brain cancer and other. This division was chosen to promote the best understanding of the area by the main target audience of this paper - oncologist, clinicians and other practitioners.

We conducted a search of papers in the PubMed database published between January 2014 and September 2020 with individual and combination of search terms such as “interpretability,” “deep learning,” “oncology”, “cancer” and “decision support systems,” and compiled the results in Table II. 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 common interpretability strategies were saliency maps (32%) and feature importance (20%) and among the prediction tasks, most works focused on diagnosis of malignancy (45%) and of different pathologies (27%).

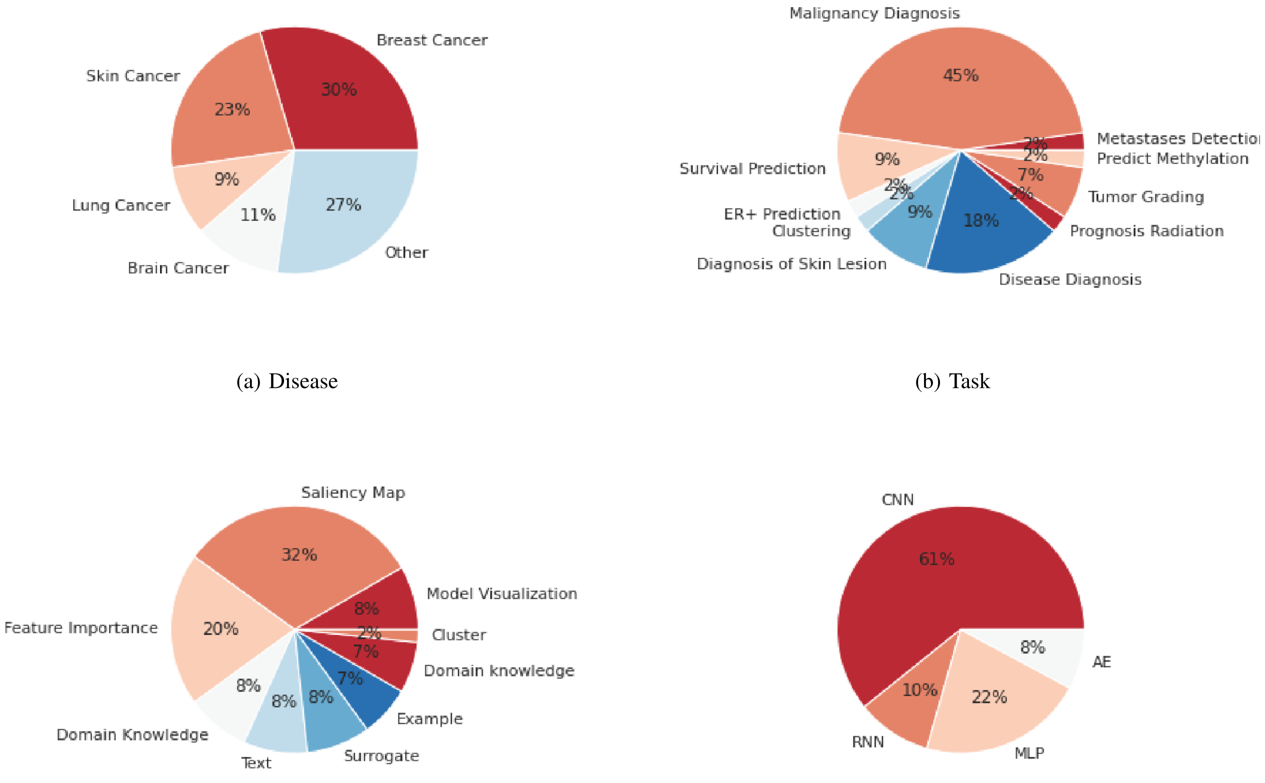
Fig. 10 helps visualize the distribution of papers based on different classifications present in Table II, namely the target disease and task as well as the interpretability strategy (explanation) and ANN technique (architecture).

A. Breast Cancer

Prediction of breast cancer malignancy has been one the most successful applications of deep learning in oncology, achieving 87% sensitivity and 96% specificity when diagnosing mammograms [78]. It also is the main task on interpretability work (69% of breast cancer studies). Due to the availability of well-curated public datasets on breast cancer, mainly mammograms and hematoxylin and eosin (H&E) stained histological images, research in this area has 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, please refer to section III-C2. 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). In what follows, we analyze with detail some of the main selected works on the topic.

Graziani *et al.* [35] 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 [28], [29]) and by extracting saliency maps [79]. They found that the network detected nuclei-resembling shapes and regions of nuclei with marked variations in size and irregular shapes. Hsieh *et al.* [27] used Network Dissection method [80] 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. Fig. 11 shows an example of a pattern which was labeled as ‘Calcified Vessels’. Also, other BI-RADS [81] medical concepts (e.g. mass margin) were found to overlap with patterns detected by the network.



(a) Disease

(b) Task

(c) Interpretability strategy

(d) ANN techniques

Fig. 10. Distribution of papers reviewed based on characteristics of Table II

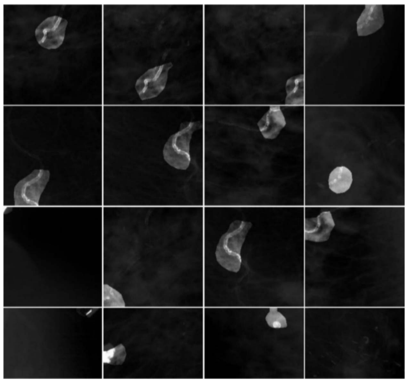


Fig. 11. Example of pattern detected by the network and labeled by an expert as ‘Calcified Vessels’ in the web-based labeling tool [27].

Rather than being validated by experts, Graziani *et al.* [36], [37] introduced Regression Concept Vectors (an extension of Concept Activation Vectors [82]) 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. Contrast was found to be positively correlated with malignancy, while correlation was negatively correlated. Kim *et al.* [38] used medical concepts during training, computing their importance alongside saliency maps to help explain the malignancy diagnosis of mammograms.

Antropova *et al.* [39] visualized the values of both 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.* [40] trained a malignancy diagnosis network able to justify its decisions both visually and textually. It trained a a language model that composes text description [38], [76], [77], [83] 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 [41] making it easier to compute the feature importance, or extract simpler classifiers that could present physicians with simple rules (i.e. decision rules [42] and symbolic rules [43]) increasing interpretability.

Feature importance was the focus of most works dealing with gene expression data. For example, SALMON [44] 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 co-expression modules play pivotal roles in patient prognosis. Similar methods, using the H2O [84] library, were used to detect the important features in the detection of estrogen-receptor-positive (ER+) patients based on the classification of the Estrogen Receptor Status of breast cancer patients based on metabolomics data [45]. 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.* [46] 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.

B. 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 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 [85]. 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 [47]. Inspection of activations led 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.* [48] 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 [49] and diagnosis of skin diseases based on WSI H&E images [50]. PatchNet [49] 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.* [50] also found that smaller convolutional filters resulted in more fine-grained saliency maps. Gonzalez-Diaz *et al.* [51] 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.* [53] 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% to 61% when the 15 most similar cases were provided to the users. Silva *et al.* [54], [55] 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.* [56] explained the decision with similar examples using k-nearest neighbors on the deep features and highlighted the most salient regions of the image.

C. 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 [57]. Radiologists confirmed, by inspecting saliency maps [86], 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.* [58] 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.* [59] 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.* [60] used a combination of hand-crafted features composed of clinical features and cancer biomarkers in a non-small 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 [87].

D. 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 - researchers have focused on producing saliency maps from the 3D MRI scans or Region of Interest (ROI) annotated by experts. Pereira *et al.* [61] extended existing saliency map methods for three dimensional inputs [79], [88]. The ROI classifier achieved better performance than the 3D scan 93% and 90% accuracy), but they were both able to locate the tumor. Pereira *et al.* [62] also used a feature importance method [89] 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 *et al.* [63] trained 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) [90] and the methylation data from the Cancer Genome Atlas (TCGA) [91]. 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.* [64] 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 [65] with validation of produced saliency maps by expert pathologists.

E. 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.* [66] 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.* [67] 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.* [68] diagnosed eight different cancer types from RNA-seq expression and single nucleotide variation (SNV) data. To explain its decisions, a linear surrogate model [89] 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 [69], saliency maps [86] 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.* [70] extended the same saliency map method to interpret the diagnosis of 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.* [71] interpreted the diagnosis of colorectal polyps based on histological images using saliency maps [79] [86] and found that by adding a boundary box around them increased their similarity with pathologists' segmentations.

Inglese *et al.* [72] 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 [92] unveiling clusters with different chemical and biological interactions occurring.

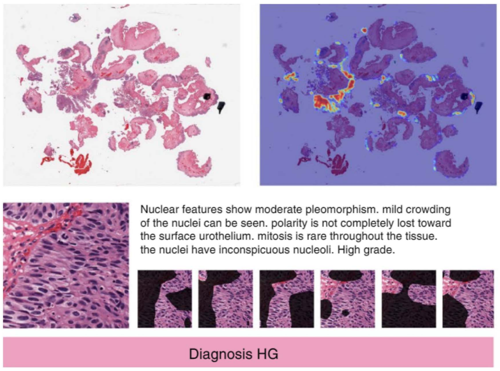

Tumour
1
0.5
0
Normal
Nuclear features show moderate pleomorphism. mild crowding of the nuclei can be seen. polarity is not completely lost toward the surface urothelium. mitosis is rare throughout the tissue. the nuclei have inconspicuous nucleoli. High grade.
Diagnosis HG

Fig. 12. Left: H&E stained whole-slide tissue image. Right: saliency map generated [75]. Bottom: description generated for the image and feature-aware attention maps.

Zhang *et al.* [73] 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.* [74] 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.* [75] 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 (Fig. 12). Similarly, TieNet [76] provided the same explanation for the network which diagnoses diseases based on chest radiographs and generates text descriptions with similar visual attention. MDNet [77] 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.

V. Open Issues and Promising Research Directions

As DL grows in popularity, so does the need for interpretability in the dichotomy between ML and medical practice. From this survey, it becomes clear that are four main issues that needs more attention: (1) limitation on the applications of interpretability methods; (2) limitation on medical tasks explored; (3) lack of reliability of some interpretability method; and (4) lack of evaluation metrics for interpretability methods. Throughout this section, we will provide a discussion on the former four issues.

A. Limitation on the Applications of Interpretability Methods

Du *et al.* [93] classified three major application of interpretability strategies: model validation, model debugging and knowledge discovery.

Model validation verifies that the model was able to learn useful knowledge and avoid learning bias information. The majority of works reviewed follow in this category, for example works which explored the use of saliency maps mainly focused on verifying that the region highlighted corresponded to regions segmented by experts.

Other applications for the interpretation of deep learning models, such as model debugging and knowledge discovery, were overlooked by the current literature and constitute promising directions to further improve the diagnostic capabilities of models and discover new insights on the biology of different cancer diseases.

Model debugging aims at analyzing what leads to the misbehaviour of models and erroneous predictions. Interpretability can help to uncover the reason for this misbehaviour, by inspecting the examples what were misclassified by the model, examples that have artifacts from the data collection (e.g. metal tools in a CT scan, hairs in a dermoscopic image), in addition to difficult to diagnose cases. Model debugging is also extremely relevant when generalizing the model for other hospital data or for clinical use where the risk for misbehavior is much bigger. This application is still overlooked in current works in the field of oncology.

Carlini *et al.* [94] demonstrated that standard models can make perfect predictions in random training set while performing poorly on the test set. This proves the model's ability to memorize the input data even if it is random which causes low generalization to unseen data. The lack of generalization of models which can be caused by overfitting to the training dataset must be an active concern of all ML practitioners, especially deep learning techniques as the high complexity of the models coupled with an low data size increases risk of overfitting.

Another issue related with model debugging is adversarial attacks which consist on inputs that are intentionally crafted to force the model to make a mistake. Finlayson [95] demonstrated how an adversarial noise added to a dermoscopic image previously diagnosed as benign with over 99% confidence by a highly accurate model resulted with the model predicting malignant with 100% of confidence even though the difference is imperceptible to the human eye. Finlayson [95] also pointed at insurance claims approvals as a possible motivation for adversarial attacks.

Another problem with generalizability is discriminatory bias where models learn unintended associations regarding minority subgroups due to bias in the data used to train the model [96]. An example is how malignancy diagnosis systems with accuracy similar to that of board-certified dermatologists underperformed on images of lesions in skin of color due to the majority of training examples represent fair skinned patients [96].

Discriminatory bias is not the only type of bias which can cause problems as there have been several instances where exceptional results have been obtained from the model learning to distinguish slides based on the hospital they came from or the clinicians that generated the ground truth rather than actual evidence in the slide [96]. For example, a system for the detection of pneumonia on chest x-rays was able learn to associate the use of a portable x-ray machine with pneumonia [57].

Knowledge discovery allows physicians and researchers to obtain new insights on the physiology of the disease by interpreting the deep learning model and its decision process, such as finding that HER2 receptor over-expression is related to breast cancer. Knowledge discovery could lead to finding other receptors, thus helping in the characterization of cancer diseases that are still unknown to this date. While some visualization methods have been used to discover cluster of patients with specific characteristics [46], [72], this direction of research is still mostly unexplored.

B. Limitation on Medical Tasks Explored

Analysis on the results of the review (Fig. 10) shows that 72% of works focus on some type of disease classification (45% malignancy diagnosis, 18% disease diagnosis, 9% diagnosis of skin lesion). This shows a great imbalance as there exists many more medical tasks in the oncology field with promising results but still lack interpretability. In the following sub-section relevant work on other medical tasks will be briefly reviewed. Those medical tasks are:

- Tumor or lesion segmentation: identify the set of voxels which make up the lesions or tumors present [97], [98];
- Organ and substructure segmentation: identify the set of voxels which make up either the contour or the interior of the objects of interest [99];
- Cancer prognosis: estimate the likely course and outcome of a disease [13], [100];
- Radiation treatment planning: determine location and dosage to deliver the most desirable dose distribution of radiotherapy [101];
- Image registration: seeks to determine a transformation that will map two volumes (source and reference) to the same coordinate system [102];
- Image generation and enhancement: includes many different tasks to improve quality of the input from removing obstructing artifacts or noise in images to complete missing data [103]–[105].

Tumor and lesion segmentation is an important first step for numerous other tasks such as diagnosis and treatment planning, in order to evaluate the extend of the diseased tissue. DL techniques have achieved state-of-the-art results is in brain tumor segmentation from MRI scans [97], [98]. The same type of networks have also been used in the segmentation of different lesion of the skin based on dermoscopic images [106], [107].

Segmentation of organs and substructures is also an critical step before radiotherapy in order to decide the which regions to avoid targeting with radiation. One example of it is the segmentation of organs from abdominal CT scans [99].

Cancer prognosis is comprised of a large number of sub-tasks such as survival prediction and prediction of likelihood of metastases. Zhu *et al.* [100] for example, reviewed a large number of studies which applied DL techniques to different cancer prognosis tasks such as cancer recurrence, progression and survival prediction [100]. Other studies focused on sub-tasks which concern with the progression of the disease after treatment, from the prediction of future distant metastases and local-regional recurrence using pre-treatment, post-treatment and follow-up medical imaging scans [13].

Radiation treatment planning requires not only the segmentation of diseased tissue but also the dosage that should be used. An CNN-based model was used to MRI to accurately transfer contrast into CT images with clearly identified air, brain soft tissue, and bone highly similar to that of current methods based on CT and used in medical practice [101].

Image registration, also known as image fusion, is commonly used to combine two modalities - for example PET-CT is obtained by combining two different modalities (PET and CT), but also multiple images of the same modalities. Fu *et al.* [102] review a large number of DL techniques proposed for the image registration of different modalities such as T1 and T2 MRIs and MRI and CT.

In addition, DL approaches also has seen success in restoring medical images corrupted with noise or artifacts, but the interpretation of the reasoning behind this process has also been pointed out as a challenge [108]. The extensive use of CT in medical analysis has raise some concerns due to the large dose of radiation that it delivers to the patient. Low dose CT is a solution for this problem, but by using lower radiation amounts, noise and artifacts become a problem. DL techniques have been proposed to reconstruct low dose CT images and recover from noise and streaking artifacts caused by metal objects [103]–[105].

Even though DL techniques have help the numerous problems pointed out above, they all face the same obstacle which prevents their use in clinical practice, the lack of interpretability. Future research efforts should then be targeted in the exploration of other application of interpretability methods other than model validation and different cancer tasks than disease diagnosis. With the expansion of cancer applications, other interpretability strategies will emerge based on images (most used modality) and other modalities that may be more associated with other problems.

C. Lack of Reliability of Some Interpretability Methods

Some post-hoc interpretability methods can present bias [109], [110] and might not be representative of the behavior of the model they are trying to explain [111]. This happens because although explanations should approximate as much as possible the actual behaviour of the model, during the process of optimization (e.g. backpropagation) some inputs given to the network are outside the distribution of the training data and can trigger artifacts of the deep learning model.

As different interpretation methods sometimes focus on distinct aspects of the model [105], a promising direction to improve the reliability of the interpretations is deploy an ensemble of complementary interpretability methods. Furthermore, interpretability methods should also be provided with imperfect data (i.e. noisy) to guarantee robustness to noise.

D. Lack of Evaluation Metrics for Interpretability Methods

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 [17]. The lack of ground-truth explanations, for example the 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, and generalize the obtained results. One of possible solutions to solve this issue is to conduct a comparison study between the interpretation produced by the deep model and one produced by a set of physicians. However, and once again, this solution may not be generalizable, hence most studies conducting 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 based on three factors. First, evaluate how faithful the explanations are to the actual model's behaviour. Second, evaluate how easily the explanations are understood by the physician. Third, evaluate the usefulness of the explanation of its target application (i.e. model validation). Only by evaluating these factors can explanations extracted from deep learning models be truly trusted and applied in clinical practice.

VI. Conclusion

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 [33], [34], [112].

In this work, we presented an easy guide for oncologists where we introduced various deep learning techniques and illustrated how the decisions of these could be interpreted with self-explanatory oncological cases to better illustrate. We also review the related research on the application of interpretability methods for cancer diseases, summarizing their main conclusions.

To the extent of the authors' knowledge, such comprehensive review on the interpretability of DL models for cancer diseases has not been previously performed. Overall, a high number of studies focused on breast, skin and brain cancers (60%) and on the explanation of the importance of tumor characteristics like tumor dimensions and shape, in the prediction of decision system. The majority of DL techniques interpreted were multilayer perceptrons and convolutional neural networks, often used to predict based on raw images or handcrafted features extracted from them.

As discussed in the previous section, three main issues were identified: (1) limitation on the applications of interpretability methods; (2) lack of reliability of some interpretability method; and (3) lack of evaluation metrics for interpretability methods.

Future research should go beyond model validation and apply interpretability to understand how models misbehave, as well as discover new knowledge about different cancer diseases. Also, although DL has been successful in many cancer tasks (e.g. tumor segmentation, cancer prognosis and image registration), works aim at interpreting models on these tasks remain unexplored. 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.

Conflict of Interest

The authors declare no competing interests.

References

[1] *Futurist*, “FDA approvals for smart algorithms in medicine in one giant infographic,” 2019. Accessed: Jan. 12, 2019. [Online]. Available: https://medicalfuturist.com/fda-approvals-for-algorithms-in-medicine/

[2] W. Penny and D. Frost, “Neural networks in clinical medicine,” *Med. Decis. Making : Int. J. Soc. Med. Decis. Mak.*, vol. 16, pp. 386–98, 1996.

[3] F. Rosenblatt, “ The perceptron, a perceiving and recognizing automaton project para,” Cornell Aeronautical Lab., Buffalo, New York, NY, USA, Tech. Rep. 85-60-1, 1957.

[4] T. J. Brinker *et al.*, “A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task,” *Eur. J. Cancer*, vol. 111, pp. 148–154, 2019.

[5] B. E. Bejnordi *et al.*, “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” *J. Amer. Med. Assoc.*, vol. 318, no. 22, pp. 2199–2210, 2017.

[6] M. Reyes *et al.*, “On the interpretability of artificial intelligence in radiology: Challenges and opportunities,” *Radiol.: Artif. Intell.*, vol. 2, 2020, Art. no. e 190043.

[7] S. Liu, X. Wang, M. Liu, and J. Zhu, “Towards better analysis of machine learning models: A visual analytics perspective,” *Vis. Informat.*, vol. 1, no. 1, pp. 48–56, 2017.

[8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 10 1986.

[9] F. Shahidi, S. Mohd Daud, H. Abas, N. A. Ahmad, and N. Maarop, “Breast cancer classification using deep learning approaches and histopathology image: A comparison study,” *IEEE Access*, vol. 8, pp. 187 531–187552, 2020.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, Nov. 1998.

[11] Y. Bengio, “Deep learning of representations for unsupervised and transfer learning,” in *Proc. J. Mach. Learn. Res.: Workshop Conf.*, 2011, pp. 1–20.

[12] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, “Deep learning techniques for medical image segmentation: Achievements and challenges,” *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, Aug. 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31144149

[13] Y. Xu *et al.*, “Deep learning predicts lung cancer treatment response from serial medical imaging,” *Clin. Cancer Res.*, vol. 25, no. 11, pp. 3266–3275, 2019. [Online]. Available: https://clincancerres.aacrjournals.org/content/25/11/3266

[14] P. Baldi, “Autoencoders, unsupervised learning and deep architectures,” in *Proc. Int. Conf. Unsupervised Transfer Learn. Workshop*, 2011, pp. 37–50.

[15] J. Tan, M. Ung, C. Cheng, and C. Greene, “Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders,” *Pacific Symp. Biocomputing*, vol. 20, pp. 132–43, 2015.

[16] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[17] F. Doshi-Velez and B. Kim, “Towards A rigorous science of interpretable machine learning,” Feb. 2017, *arXiv:1702.08608*.

[18] C. Molnar, “Interpretable machine learning,” 2019. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[19] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, “How to explain individual classification decisions,” *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, 2010.

[20] S. Bazen and X. Joutard, “The Taylor decomposition: A unified generalization of the Oaxaca method to nonlinear models,” 2013. [Online]. Available: https://halshs.archives-ouvertes.fr/file/index/docid/828790/filename/WP_2013_-_Nr_32.pdf

[21] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLoS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[22] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *Proc. Workshop Int. Conf. Learn. Representations*, 2014, pp. 1–8.

[23] K. J. Geras, R. M. Mann, and L. Moy, “Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives,” *Radiology*, vol. 293, no. 2, pp. 246–259, Nov. 2019. [Online]. Available: https://europepmc.org/articles/PMC6822772

[24] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, “Understanding neural networks through deep visualization,” in *Proc. Int. Conf. Mach. Learn. - Deep Learn. Workshop*, 2015, pp. 1–12.

[25] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2018, *arXiv:1802.03426*.

[26] M. Graziani *et al.*, “Improved interpretability for computer-aided severity assessment of retinopathy of prematurity,” in *Proc. Comput.-Aided Diagnosis*, 2019, pp. 62–73.

[27] J. Wu *et al.*, “Expert identification of visual primitives used by CNNs during mammogram classification,” in *Proc. SPIE*, 2018, pp. 633–641.

[28] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks,” in *Adv. Neural Inf. Process. Syst.*, 2016, pp. 3387–3395.

[29] C. Olah, A. Mordvintsev, and L. Schubert, “Feature visualization,” *Distill*, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–7.

[31] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, “A survey on incorporating domain knowledge into deep learning for medical image analysis,” *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101985. [Online]. Available: http://dx.doi.org/10.1016/j.media.2021.101985

[32] R. Caruana, H. Kangarloo, J. Dionisio, U. Sinha, and D. Johnson, “Case-based explanation of non-case-based learning methods,” in *Proc. AMIA Symp.*, 1999, pp. 212–215.

[33] A. Vellido, “The importance of interpretability and visualization in machine learning for applications in medicine and health care,” *Neural Comput. Appl.*, vol. 0123456789, pp. 1–15, Feb. 2019. [Online]. Available: https://doi.org/10.1007/s00521-019-04051-w

[34] F. Cabitza, R. Rasoini, and G. F. Gensini, “Unintended consequences of machine learning in medicine,” *J. Amer. Med. Assoc.*, vol. 318, no. 6, pp. 517–518, 2017. [Online]. Available: https://doi.org/10.1001/jama.2017.7797

[35] M. Graziani, V. Andrearczyk, and H. Müller, “Visual interpretability for patch-based classification of breast cancer histopathology images,” in *Proc. Med. Imag. Deep Learn.*, 2018, pp. 1–4.

[36] M. Graziani, V. Andrearczyk, and H. Möuller, “Regression Concept Vectors for Bidirectional Explanations in Histopathology,” in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Berlin, Germany: Springer, 2018, pp. 124–132.

[37] M. Graziani, V. Andrearczyk, S. Marchand-maillet, and H. Müller, “Concept attribution: Explaining CNN decisions to physicians,” *Comput. Biol. Med.*, vol. 123, 2020, Art. no. 103865. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2020.103865

[38] S. T. Kim, J. H. Lee, H. Lee, and Y. M. Ro, “Visually interpretable deep network for diagnosis of breast masses on mammograms,” *Phys. Med. Biol.*, vol. 63, no. 23, pp. 1–14, Dec. 2018.

[39] N. Antropova, B. Huynh, and M. Giger, “A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets,” *Med. Phys.*, vol. 44, pp. 5162–5171, 2017.

[40] H. Lee, S. T. Kim, and Y. M. Ro, “Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis,” in *Proc. Lecture Notes Comput. Sci. (Including Subseries Lecture Notes Artif. Intell. Lecture Notes Bioinf.)*, vol. 11797, Oct. 2019, pp. 21–29. [Online]. Available: http://arxiv.org/abs/1906.03922

[41] D. Alvarez-Melis and T. Jaakkola, “Towards robust interpretability with self-explaining neural networks,” in *Advances in Neural Information Processing Systems* 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2018, pp. 7775–7784.

[42] J. P. Amorim, I. Domingues, P. Abreu, and J. Santos, “Interpreting deep learning models for ordinal problems,” in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2018, pp. 373–377.

[43] G. Bologna and Y. Hayashi, “Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning,” *J. Artif. Intell. Soft Comput. Res.*, vol. 7, no. 4, pp. 265–286, 2017.

[44] Z. Huang *et al.*, “Salmon: Survival analysis learning with multi-omics neural networks on breast cancer,” *Front. Genet.*, vol. 10, no. 4, pp. 1–13, 2019.

[45] F. M. Alakwaa, K. Chaudhary, L. X. Garmire, and B. G. Program, “Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data,” *J. Proteome Res.*, vol. 17, pp. 337–347, 2018.

[46] Q. Liu and P. Hu, “Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer,” *Cancers*, vol. 11, no. 4, pp. 1–13, 2019.

[47] P. Van Molle, M. De Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, and 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*. Berlin, Germany: Springer, 2018, pp. 115–123.

[48] A. Cruz-Roa *et al.*, “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks,” in *Proc. SPIE*, 2014, vol. 9041, Art. no. 904103.

[49] A. Radhakrishnan, C. Durham, A. Soylemezoglu, and C. Uhler, “PatchNet: Interpretable neural networks for image classification,” 2017, *arXiv:1705.08078*.

[50] M. Paschali *et al.*, “Deep learning under the microscope: Improving the interpretability of medical imaging neural networks,” 2019, *arXiv:1904.03127*.

[51] I. Gonzalez Diaz, “DermaKNet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis,” *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 547–559, Mar. 2019.

[52] A. A. Cruz-Roa, J. E. Arevalo Ovalle, A. Madabhushi, and F. A. González Osorio, “A Deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection,” in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2013, pp. 403–410.

[53] M. Sadeghi, P. K. Chilana, and 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*. Berlin, Germany: Springer, 2018, pp. 141–148.

[54] W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso, “Towards complementary explanations using deep neural networks,” in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Berlin, Germany: Springer, 2018, pp. 133–140.

[55] W. Silva, K. Fernandes, and J. S. Cardoso, “How to produce complementary explanations using an ensemble model,” in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[56] N. C. F. Codella, C.-C. Chung-Ching Lin, A. Halpern, M. Hind, R. Feris, and 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*. Berlin, Germany: Springer, 2018, pp. 97–105.

[57] P. Rajpurkar *et al.*, “Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists,” *PLOS Med.*, vol. 15, no. 11, pp. 1–17, 2018.

[58] R. Paul, Y. Liu, Q. Li, L. Hall, and D. Goldgof, “Representation of deep features using radiologist defined semantic features,” in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1429–1435.

[59] S. Shen, S. X. Han, D. R. Aberle, A. A. T. Bui, and W. Hsu, “An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification,” *Expert Syst. Appl.*, vol. 128, pp. 84–95, 2018.

[60] S. Cui, Y. Luo, H.-H. Tseng, R. T. Haken, and I. El Naqa, “Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage,” *Med. Phys.*, vol. 46, no. 5, pp. 2497–2511, 2019.

[61] S. Pereira, R. Meier, V. Alves, M. Reyes, and 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*. Berlin, Germany: Springer, 2018, pp. 106–114.

[62] S. Pereira *et al.*, “Enhancing interpretability of automatically extracted machine learning features: Application to a RBM-Random forest system on brain lesion segmentation,” *Med. Image Anal.*, vol. 44, pp. 228–244, 2018.

[63] L. Han and M. R. Kamdar, “MRI to MGMT: Predicting methylation status in glioblastoma patients using convolutional recurrent neural networks,” in *Pac. Symp. Biocomput.*, 2018, pp. 331–342.

[64] J. Lao *et al.*, “A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–8, 2017.

[65] P. Mobadersany *et al.*, “Predicting cancer outcomes from histology and genomics using convolutional networks,” *Proc. Nat. Acad. Sci.USA*, vol. 115, no. 13, pp. E 2970–E2979, 2018.

[66] T. Ahn *et al.*, “Deep learning-based identification of cancer or normal tissue using gene expression data,” in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 1748–1752.

[67] S. Yousefi *et al.*, “Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models,” *Sci. Rep.*, vol. 7, 2017, Art. no. 11707.

[68] O. Oni and S. Qiao, “Model-agnostic interpretation of cancer classification with multi-platform genomic data,” in *Proc. 10th ACM Int. Conf. Bioinf. Comput. Biol. Health Inform.*, New York, NY, USA: ACM, 2019, pp. 34–41.

[69] R. Fonollà *et al.*, “Ensemble of deep convolutional neural networks for classification of early Barrett’s neoplasia using volumetric laser endomicroscopy,” *Appl. Sci.*, vol. 9, no. 11, 2019, Art. no. 2183.

[70] L. C. Garcia-Peraza-Herrera *et al.*, “Interpretable fully convolutional classification of intrapapillary capillary loops for real-time detection of early squamous neoplasia,” pp. 1–8, 2018, *arXiv:1805.00632*.

[71] B. Korbar *et al.*, “Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 821–827.

[72] P. Inglese *et al.*, “Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer,” *Chem. Sci.*, vol. 8, pp. 3500–3511, 2017.

[73] K. Zhang *et al.*, “An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study,” *J. Med. Internet Res.*, vol. 20, no. 11, pp. 1–13, Nov. 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30429111http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6301833

[74] S. Zhang *et al.*, “A novel interpretable computer-aided diagnosis system of thyroid nodules on ultrasound based on clinical experience,” *IEEE Access*, vol. 8, pp. 53223–53231, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9016204/

[75] Z. Zhang *et al.*, “Pathologist-level interpretable whole-slide cancer diagnosis with deep learning,” *Nature Mach. Intell.*, vol. 1, pp. 236–245, May 2019.

[76] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, “TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9049–9058.

[77] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, “MDNet: A semantically and visually interpretable medical image diagnosis network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3549–3557.

[78] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep learning to improve breast cancer detection on screening mammography,” *Sci. Rep.*, vol. 9, no. 1, Aug. 2019, Art. no. 12495. [Online]. Available: https://doi.org/10.1038/s41598-019-48995-4

[79] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[80] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network dissection: Quantifying interpretability of deep visual representations,” in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3319–3327.

[81] A. A. Kabbani *et al.*, “Breast imaging-reporting and data system (BI-RADS)” Reston VA: American College of Radiology, 1998.

[82] B. Kim *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV),” in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2668–2677.

[83] S. T. Kim, J.-H. Lee, and Y. Ro, “Visual evidence for interpreting diagnostic decision of deep neural network in computer-aided diagnosis,” in *Proc. SPIE*, 2019, pp. 1–19.

[84] H2O, “H2O.ai,” 2019. Accessed: Jan. 7, 2019. [Online]. Available: https://www.h2o.ai/

[85] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[86] B. Zhou, A. Khosla, Á. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[87] Y. Luo, H.-H. Tseng, S. Cui, L. Wei, R. K. Ten Haken, and I. El Naqa, "Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling," *BJR| Open*, vol. 1, no. 1, Jul. 2019, Art. no. 20190021. [Online]. Available: https://www.birpublications.org/doi/10.1259/bjro.20190021

[88] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[89] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA: ACM, 2016, pp. 1135–1144.

[90] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013.

[91] K. Tomczak *et al.*, "The cancer genome atlas(TCGA): An immeasurable source of knowledge," *Contemporary Oncol.*, vol. 19, no. 1A, pp. 68–77, 2015.

[92] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[93] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 65, no. 1, pp. 68–77, 2020.

[94] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *Proc. 28th USENIX Secur. Symp.*, Santa Clara, CA: USENIX Association, Aug. 2019, pp. 267–284. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

[95] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aaw4399

[96] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, pp. 1–9, Oct. 2019. [Online]. Available: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1426-2

[97] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017. [Online]. Available: http://dx.doi.org/10.1016/j.media.2016.05.004

[98] M. Mittal, L. M. Goyal, S. Kaur, I. Kaur, A. Verma, and D. J. Hemanth, "Deep learning based enhanced tumor segmentation approach for MR brain images," *Appl. Soft Comput.*, vol. 78, pp. 346–354, 2019.

[99] H. Kim *et al.*, "Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network," *Sci. Rep.*, vol. 10, no. 1, Apr. 2020, Art. no. 6204. [Online]. Available: https://doi.org/10.1038/s41598-020-63285-0

[100] W. Zhu, L. Xie, J. Han, and X. Guo, "The application of deep learning in cancer prognosis prediction," *Cancers*, vol. 12, no. 3, pp. 1–19, Mar. 2020.

[101] F. Liu, P. P. Yadav, A. M. Baschnagel, and A. B. McMillan, "Mr-based treatment planning in radiation therapy using a deep learning approach," *J. Appl. Clin. Med. Phys.*, vol. 20, pp. 105–114, 2019.

[102] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: A review," *Phys. Med. Biol.*, 2019, Art. no. 20TR01.

[103] H. Chen *et al.*, "Low-dose CT with a residual encoder-decoder convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017. [Online]. Available: http://dx.doi.org/10.1109/TMI.2017.2715284

[104] Q. Yang *et al.*, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018. [Online]. Available: http://dx.doi.org/10.1109/TMI.2018.2827462

[105] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang, "Interpretable deep learning under fire," in *Proc. 29th USENIX Secur. Symp.*, 2018, pp. 1659–1676. [Online]. Available: http://arxiv.org/abs/1812.00891

[106] A. Adegun and S. Viriri, "Deep learning model for skin lesion segmentation: Fully convolutional network," in *Image Analysis and Recognition*, F. Karray, A. Campilho, and A. Yu, Eds. Cham, Switzerland: Springer, 2019, pp. 232–242.

[107] J. Lameski, A. Jovanov, E. Zdravevski, P. Lameski, and S. Gievska, "Skin lesion segmentation with deep learning," in *Proc. 18th Int. Conf. Smart Technol.*, 2019, pp. 1–5.

[108] H.-M. Zhang and B. Dong, "A review on deep learning in medical image reconstruction," *J. Operations Res. Soc. China*, Jan. 2020. [Online]. Available: https://doi.org/10.1007/s40305-019-00287-4

[109] P.-J. Kindermans *et al.*, *The (Un)reliability of Saliency Methods*. Berlin. Germany: Springer, 2019, ch. 4, pp. 267–280.

[110] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Advances in Neural Information Processing Systems 31*. Red Hook, NY, USA: Curran Associates, Inc., 2018, pp. 9505–9515.

[111] J. Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Med.*, vol. 24, no. 9, pp. 1342–1350, 2018.

[112] D. Ravì *et al.*, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.

José P. Amorim received the M.S. degree in informatics engineering from the Faculty of Engineering, University of Porto, Porto, Portugal, in 2017. He is currently working toward the Ph.D. degree with the University of Coimbra, Coimbra, Portugal. He is currently affiliated with the IPO-Porto Research Centre (CI-IPOP) and the Centre for Informatics and Systems, University of Coimbra and his research focuses on creating interpretable models in order to assist physicians in the field of oncology.

Pedro H. Abreu received the Informatics Engineering degree and the Ph.D. degree in soccer teams modeling from Porto University, Porto, Portugal, in 2006 and 2011, respectively. He is currently an Assistant Professor with the DEI, University of Coimbra, Coimbra, Portugal. He is the author of more than 60 publications in international conferences and journals. His research interests include artificial intelligence applied to medical informatics contexts.

Alberto Fernández received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain, in 2005 and 2010, respectively. Since 2017, he has been a University Professor with the Department of Computer Science and Artificial Intelligence and teaches with the Faculty of Science and with the E.T.S. of Computer Engineering and Telecommunications, the University of Granada. His research interests include classification in unbalanced domains, learning fuzzy rules, evolutionary algorithms, multiclassification problems with Ensembles and decomposition techniques and data science and Big Data applications, in which he has published more than 50 works in international journals ISI, collaborated in more than 20 projects and research contracts and three doctoral thesis. He is a Member of the editorial committee of a number of international journals, such as the Applied Intelligence, PLOS-One, Progress in Artificial Intelligence, Cognitive Computation and Big Data, and a Member of the "referees" committee of other ISI journals, such as the *Fuzzy Sets and Systems*, IEEE Transactions on Fuzzy Systems, Information Sciences, and among others. He was the recipient of the Extraordinary Doctoral Thesis Award 2010, on three occasions the Award for Works of Excellence of the University of Granada, and also the Lofti Zadeh Award for best publication in 2009.

Mauricio Reyes received the bachelor's degree from the University of Santiago de Chile, Chile, in 2001. He is currently the Head of the Healthcare Imaging A.I. Group with the University of Bern, Switzerland, and also the Head of the Data Science Team of the Insel Data Science Center (IDSC), University Hospital Bern. His thesis Three-dimensional Reconstruction of a Human Embryo Hand Using Artificial Vision Techniques was awarded best Electrical Engineering bachelor thesis work. Later on, during 2002- 2004, he conducted studies to obtain a Ph.D.. degree in computer sciences from the University of Nice, France on the topic of lung cancer imaging and breathing compensation in emission tomography, under the supervision of Dr. Grégoire Malandain, Asclepios research project (formerly known as Epidaure). In 2006, he joined the Medical Image Analysis Group, the MEM Research Center as a Postdoctoral Fellow focusing on topics related to medical image analysis and statistical shape models for orthopaedic research. In 2008, he took more than the lead of the Medical Image Analysis group at the Institute for Surgical Technology and Biomechanics, Switzerland.

João Santos received the M.Sc. degree in physics (Optoelectronics and Lasers) and is a Physicist with the Ph.D. degree in physics (Condensed Matter). He concluded his Residency in Medical Physics in 2005 with the Portuguese Institute of Oncology of Porto (IPOPFG, EPE). He is currently working as a Medical Physicist Expert with IPOPFG, EPE and he has been the Coordinator of the Medical Physics, Radiobiology and Radiation Protection group of the IPOPFG, EPE Research Center since 2008.

Miguel H. Abreu Graduated in medicine from the University of Porto, Porto, Portugal, in 2007 and the Ph.D. degree in medical sciences from the same Institution in 2016. He is currently the M.D. of oncologist with Instituto Português Oncologia do Porto Francisco Gentil. He completed his residence training in medical oncology with the Portuguese Institute of Oncology of Porto, Porto, Portugal, in 2013 and has been a Medical Assistant with the same Institution since 2014, being dedicated to the treatment of breast and gynaecological cancer patients. He is an Active Member of important research groups including the EORTC Group for gynaecological cancer, EORTC for breast cancer and the SOLTI Group, and is a Coinvestigator in 37 clinical trials and the Principal investigator in four of them.