Received 8 August 2022, accepted 9 September 2022, date of publication 14 September 2022,
date of current version 23 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206449

SURVEY

A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?

TIAGO GONÇALVES, (Member, IEEE), ISABEL RIO-TORTO, LUÍS F. TEIXEIRA, (Member, IEEE), AND JAIME S. CARDOSO, (Senior Member, IEEE)
Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
University of Porto, 4200-465 Porto, Portugal

Corresponding author: Tiago Gonçalves (tiago.f.goncalves@inesctec.pt)

This work was supported in part by the project ''Transparent Artificial Medical Intelligence (TAMI),'' is co-financed by the European Regional Fund (ERDF) through the Operational Program for Competitiveness and Internationalization-COMPETE2020, the North Portugal Regional Operational Program-NORTE 2020, under Grant NORTE-01-0247-FEDER 045905; and in part by the Portuguese Foundation for Science and Technology (FCT) through the CMU-Portugal International Partnership, the Ph.D. under Grant 2020.06434.BD and Grant 2020.07034.BD, and the project LA/P/0063/2020.

ABSTRACT The increasing popularity of attention mechanisms in deep learning algorithms for computer vision and natural language processing made these models attractive to other research domains. In healthcare, there is a strong need for tools that may improve the routines of the clinicians and the patients. Naturally, the use of attention-based algorithms for medical applications occurred smoothly. However, being healthcare a domain that depends on high-stake decisions, the scientific community must ponder if these high-performing algorithms fit the needs of medical applications. With this motto, this paper extensively reviews the use of attention mechanisms in machine learning methods (including Transformers) for several medical applications based on the types of tasks that may integrate several works pipelines of the medical domain. This work distinguishes itself from its predecessors by proposing a critical analysis of the claims and potentialities of attention mechanisms presented in the literature through an experimental case study on medical image classification with three different use cases. These experiments focus on the integrating process of attention mechanisms into established deep learning architectures, the analysis of their predictive power, and a visual assessment of their saliency maps generated by post-hoc explanation methods. This paper concludes with a critical analysis of the claims and potentialities presented in the literature about attention mechanisms and proposes future research lines in medical applications that may benefit from these frameworks.

INDEX TERMS Artificial intelligence, attention mechanisms, computer vision, deep learning, medical applications, medical image analysis, transformers.

I. INTRODUCTION

The concept of *attention* is not new and has been part of multidisciplinary debates across psychology, physiology, and neuroscience [1]. The inspiration behind the translation of this concept into artificial intelligence (AI) algorithms comes from biological vision systems, which appear to employ a serial computational strategy when inspecting complex scenes [2]. Historically, the pursuit of an *attention mechanism* seems to be deeply related to the problem of *visual search*, where the main objective may be to find a small object in a disorderly environment (e.g., a pen on a desk). A possible solution to this problem is to use images with high resolution and wide field-of-view. However, this results in very high

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

dimensional inputs [3]. Hence, it was essential to devise a strategy capable of selecting the most important parts of the input, deciding where to look next during the processing pipeline, and managing limited computational resources. These aspects motivated the study of *selective visual attention*, which deals with the computational complexity of computer vision tasks and often requires several basic components, such as the selection of a region of interest in the image and the selection of the dimension of features and values of interest [4]. Besides, theoretically, such mechanisms have the potential to guide the research into focusing more on sparse local models [5]. Nevertheless, the scientific community proposed several research lines that certainly influenced the modern methodologies of attention-based mechanisms for deep learning. Bandera *et al.* [6] and Minut *et al.* [3] proposed *reinforcement learning* policies to regularize the training of computer vision systems in the task of target recognition, Koch and Ullman [1] and Tsotsos *et al.* [4] proposed a set of rules that promote visual attention via selective tuning and shifts in the processing focus [1] during the analysis of a visual scene, and Itti and Koch [2] addressed this challenge using a *saliency map* based search mechanism. Recently, the increase of available computational power and the democratized access to *big data* allowed the resurgence of deep learning algorithms and their applications [7], [8]. With this paradigm shift, current methods now seek to integrate *learnable* attention mechanisms on an end-to-end basis [9], with interesting impacts in diverse applications [10].

As the available literature on attention mechanisms continues to grow, the first reviews and surveys on the topic were published and gained visibility. Hence, it is relevant to introduce the reader to the context of such studies and their content. Besides, this overview helps the reader gain intuition about the work developed in this research field while understanding the research questions that this survey aims to answer. Borji and Itti [11] reviewed several models of visual attention based on non-deep learning methods and presented a taxonomy indicating their approaches an capabilities. Cho *et al.* [12] presented an early study on the use of attention mechanisms in convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with a special focus on *soft-attention* for applications such as neural machine translation, image captioning, video description generation or end-to-end neural speech recognition. Wang and Tax [13] published a survey that explored the use of attention mechanisms in RNNs for sequence-to-sequence challenges. Chaudhari *et al.* [14] presented a survey of attention mechanisms for natural language processing, wherein they proposed a taxonomy and reviewed several neural network architectures in which attention has been incorporated while discussing applications in which modeling attention has shown significant impact. Yang [15] performed a brief overview about attention mechanisms for computer vision and introduced a preliminary taxonomy. Lindsay [16] published a review on attention mechanisms that provides intuition about the definition of attention in the neuroscience and psychology literature and covers several use cases of attention in machine learning, indicating their biological counterparts where they exist.

In 2017, Vaswani *et al.* [17] proposed the *Transformer*, a deep neural network architecture composed solely of attention operations that discarded recurrence and convolutions entirely. In the original paper, the authors applied this architecture to the task of neural machine translation and explored the use of attention operations as a means of reducing computational complexity when compared to recurrent or convolution operations. The establishment of a similar architecture for computer vision occurred almost naturally, when Dosovitskiy *et al.* [18] proposed the *Vision Transformer* architecture. Han *et al.* was one of the pioneers in analyzing the advantages and disadvantages associated with the use of Transformer-based architectures in computer vision with a particular interest in the *self-attention* mechanism [19]. Khan *et al.* [20] also provided a comprehensive overview of Transformer-based models from the perspective of their application and architectural design. Guo *et al.* [21] published a comprehensive and systematic review of attention mechanisms for computer vision and created a taxonomy that categorizes attention mechanisms according to their data domain. Xu *et al.* [22] revisited the Transformer architectures with particular emphasis on low-level vision and generation. Recently, Shamshad *et al.* [23] explored the use of Transformer-based architectures for medical image analysis and published a pioneering review that provides insight on the applications of these algorithms in several dimensions of the clinical daily routine. Almost all of the previous papers focus on the current state-of-the-art, performing a comparative analysis based solely on published results, thus, anticipating future and open challenges from a theoretical perspective.

In this survey, we perform an extensive review of attention mechanisms for medical applications (including Transformer-based architectures) with a theoretical discussion and a strong focus on the methodologies and their applications. In other works of the same genre, the authors usually focus on enumerating state-of-the-art strategies with their reported results. Besides, to our knowledge, this is the first work that approaches the topic of attention mechanisms from a broad perspective since we establish the connection between the natural language processing and computer vision domains, given the multi-modality property of medical data. Also, we complement this discussion with the Transformer-based architectures, as they currently are the most popular attention-based models. Hence, we approach this topic with a critical view and ask the following research questions:

1) Will attention mechanisms automatically improve the predictive power of deep learning algorithms for medical image applications?
2) What is the impact of integrating attention mechanisms on model complexity?
3) Can we improve the degree of interpretability of deep learning models solely through attention mechanisms?

4) How practical is it to design and build attention mechanisms for deep learning applications?

To answer these, we develop an experimental protocol with three case studies on medical image classification, using established deep learning architectures with and without state-of-the-art attention blocks and a Transformer-based model. We address the open challenges of interpretability and transparency with the help of *post-hoc* explanation methods and perform a visual analysis of the impact of different attention mechanisms in these outputs. We consider this survey a timely opportunity for the community to start questioning the extent of several claims made in the literature. In addition, our experimental protocol works as a reliable proxy for this analysis.

The main contributions of this paper are:

1) An extensive review of the use of attention mechanisms in machine learning methods (including Transformers) for several medical applications based on the types of tasks that can integrate several work pipelines of the medical domain: medical image classification, medical image segmentation, medical report understanding, and other tasks (medical image detection, medical image reconstruction, medical image retrieval, medical signal processing, and physiology and pharmaceutical research);
2) An experimental case study on medical image classification with three different use cases that focus on the integration of attention mechanisms into established deep learning architectures, the analysis of their predictive power, and a visual assessment of their saliency maps generated by *post-hoc* explanation methods;
3) A critical analysis of the claims and potentialities of attention mechanisms presented in the literature;
4) A discussion on the future challenges of medical applications and how they may benefit from attention mechanisms.

Besides the Introduction, the remainder of this article is organized as follows: Background introduces a historical overview and the fundamental concepts of attention mechanisms in machine learning, Literature Review presents an exhaustive review of the integration of attention mechanisms in algorithms for medical applications, Case Studies details the experimental work that accompanies this paper, and Conclusions and Future Challenges summarizes the main conclusions of this survey and points to future directions towards the study of attention-based algorithms for medical applications. The code is publicly available in a GitHub repository.[1]

II. BACKGROUND

This section presents a historical overview and the fundamental concepts regarding attention mechanisms in machine learning.

Although there is no clear nor unified concept for *attention*, we can refer to it as the ability to flexibly manage limited computational resources. This research topic spans several domains of knowledge and has been studied in conjunction with many other topics in neuroscience and psychology [24], including awareness, vigilance, saliency, executive control, learning, and, more recently, artificial intelligence. Therefore, we have different ways of perceiving the properties of attention mechanisms at biological, psychological or computational levels [16]. From the perspective of the domains of *arousal*, *alertness* or *vigilance*, attention could be described as a general measure of alertness or the ability to engage with the surroundings [25]. From the perspective of *sensorial stimuli*, attention is often deployed as an alert subject to specific sensory input, allowing tight control over both the stimuli and the locus of attention (e.g., selective audio attention, visual attention) [26], [27], [28]. From the perspective of *executive control* [29], the assumption is that there are multiple simultaneous competing tasks and that a central controller is needed to decide which to engage in and when, hence, attention can be thought of as the output of executive control, since the executive control system must select the targets of attention and communicate that to the systems responsible for implementing it [16]. From the perspective of *memory*, attention is understood as the ability of the brain to select the subset of information that is well-matched to the needs of the memory system, thus being a choice regarding the deployment of limited resources [16], [30]. From the perspective of *artificial intelligence*, the first successful implementations were accomplished with encoder-decoder structures in deep neural network architectures, where the main goal is to learn the best attention weights that connect the encoder to the decoder [12].

In the context of deep learning, RNNs were the pioneering attention-based structures. This is mainly due to their ability to learn and process sequential data (e.g., data with a temporal component), which makes RNNs a class of machines with dynamic states (i.e., their state depends on both the input to the system and the current state), such that a signal received at a given moment can alter the behaviour of the network at a much later point in time [31]. These properties allow RNNs to process variable-length structures without further modification, a property that has been extensively explored in several cognitive problems such as speech recognition [32], [33], text generation [34], handwriting generation [35] and machine translation [36]. The work described in [37] is perhaps the most classic example of the design and implementation of computational attention mechanisms with RNNs. In the original paper, the authors focused on the task of neural machine translation and proposed an RNN-based encoder-decoder architecture in which the task of the encoder is to compress the input data (in this case, an English sentence), while the task of the decoder is to receive and decode this data to achieve a meaningful translation (in this case, a French sentence). The principal motivation for introducing an attention mechanism in this learning pipeline is related to the difficulty of achieving a perfect alignment between the input and output sequences, which is an almost

[1] https://github.com/TiagoFilipeSousaGoncalves/survey-attention-medical-imaging

impossible goal in most cases. The role of the attention mechanism is to learn weights that make the decoder focus on the relevant parts of the input sequences to produce meaningful outputs. At the time of publication, this methodology improved the state-of-the-art and paved the way for new research questions and exploration of novel attention-based frameworks.

The study of attention mechanisms comprises two main research lines: *language, text, and speech*; and *computer vision*. Since the clinical practice (the subject of this survey) involves multiple modalities of data, we examine these two main dimensions to help the reader become familiar with the various nomenclatures, taxonomies, and practices. We believe this introductory approach is important to understand the context of the discussion presented in the following sections.

A. ATTENTION MECHANISMS FOR LANGUAGE, TEXT, AND SPEECH

An insightful and comprehensive taxonomy for categorising attention mechanisms for language, text, and speech has been proposed by Chaudhari *et al.* [14]. The authors relate most of the concepts to the encoder-decoder model. Hence, for better understanding, we need to define the following: *input sequences*, the input vectors of the model (see Fig. 1's "Input Sequence"); *output sequences*, the output vectors of the model (see Fig. 1's "Output Sequence"); *candidate states*, the hidden states of the encoder (see Fig. 1's "Encoder Vector"); *query states*, the hidden states of the decoder (see Fig. 1's "Decoder Vector").

Following this rationale, Chaudhari *et al.* [14] organise the different types of attention mechanisms into different (non-mutually exclusive) categories:

- **Number of abstraction levels:** This category takes into account the number of representation/feature levels where the model will learn the attention weights. At this level, we may consider the following types of attention: *single-level attention*, when the attention weights are computed only for the original input sequence; and *multi-level attention*, when we apply the attention mechanism on multiple levels of abstraction of the input sequence, usually in a sequential manner.
- **Number of positions:** This category takes into account the number of positions of the input sequence where the attention weights are learned. At this level, we may consider the following types of attention: *soft attention*, which consists of a weighted average of all the hidden states of the input sequence (i.e., the same as candidate states) to build the context vector (see Fig. 1's "Context Vector"); *global attention*, which is the term used in the machine translation field to describe soft attention; *hard attention*, which consists of a stochastic sampling of the hidden states to build the context vector; *local attention*, also part of the machine translation field, consists of the detection of an attention point and on the use of a window around that point to compute local attention weights.
- **Number of representations:** This category considers the number of different feature representations of the input sequence used in the learning task. At this level, we can consider the following types of attention: *single-representational attention*, which is the most common and consists in learning the attention weights using only a single feature representation; *multi-representational attention*, which consists in learning the attention weights using multiple representations (of the same input sequence) and creating a context vector that is the result of a weighted combination of these multiple representations and their attention weights; *multi-dimensional attention*, which consists in computing the attention weights along several dimensions of the representation of the input sequence, which will result in the computation of the relevance of each dimension to the learning task.
- **Number of sequences:** This category takes into account the number of input and output sequences. At this level, we may consider the following types of attention: *distinctive attention*, when the candidate and query states belong to two different input and output sequences respectively; *co-attention*, when we consider multiple input sequences at the same time and the main goal is to jointly learn their attention weights; *self-attention*, when

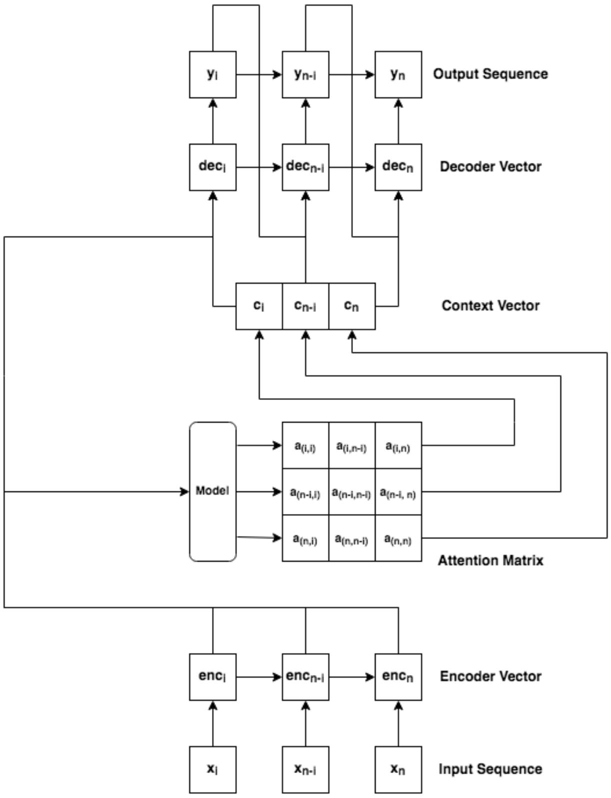


FIGURE 1. Block diagram of the general attention mechanism for language, text and speech, according to [14].

the query and candidate states belong to the same input sequence.

An analysis of the literature shows that, until very recently, most of the works in attention mechanisms for language, text, and speech derived from the framework proposed by Bahdanau *et al.* [37]. In [38], the authors trained a long short-term memory network (LSTM) on the task of neural machine translation (i.e., English to German) and presented an explicit comparative study of local and global attention mechanisms. In [39], the authors considered speech recognition as a sequence generation task (i.e., speech to transcription) and developed an attention mechanism with convolutional features to refine the quality of the output sequences. On the other hand, oppositely to this generalized use of RNN-based architectures, in [40], the authors proposed a fully-convolutional network (FCN) with an attention block that uses the dot-product operation trained to generate text from speech. Lately, the successful introduction of the *Transformer* architecture [17], based solely on attention mechanisms, started a new paradigm for their study. Recent studies invest their efforts in the computational or design improvements to the baseline Transformer architecture [41], [42], [43], or in the proposal of novel Transformer-based architectures [44], [45], [46], [47], [48]. On the other hand, opposite to the trend on the increase of model complexity, Merity [49] revisited the fundamentals of LSTMs and proposed a *single-head attention* that could represent an alternative to the use of Transformer-based architectures.

B. ATTENTION MECHANISMS FOR COMPUTER VISION

Attention mechanisms for computer vision got their inspiration from the human visual system, which can detect whether certain features of an image are relevant or not [50]. In contrast to the literature on attention mechanisms for language, text, and speech, there are already a number of interesting review articles on this topic [15], [19], [20], [21], [22]. However, only the work of Guo *et al.* [21] has presented and discussed a complete taxonomy that organizes the different types of attention mechanisms for computer vision into different categories:

- **Channel attention:** This category follows the assumption that, in deep CNNs, different channels in different feature maps usually represent different objects [51]. Therefore, the job of channel attention is to calibrate the weight of each channel adaptively. This mechanism acts as an object selector, thus determining *what to pay attention to*.
- **Spatial attention:** This category is analogous to channel attention. Here, the job of the attention mechanism is to calibrate the weight of each region of the image adaptively. This mechanism acts as an adaptive spatial region selection process, thus determining *where to pay attention*.
- **Temporal attention:** This category considers that data has a temporal dimension, therefore, in the case of computer vision tasks, this type of attention mechanism is often used for video processing. The job of temporal attention is to calibrate the weight of each time frame adaptively. This mechanism acts as a dynamic time selection process, thus determining *when to pay attention*.
- **Branch attention:** This category considers multi-branched deep learning architectures. The job of branch attention is to calibrate the weight of each branch adaptively. This mechanism acts as a dynamic branch selection process, thus determining *which to pay attention to*.
- **Channel and spatial attention:** This category combines the advantages of both channel and spatial attention. This mechanism acts as a dynamic spatial region and object selection process, thus determining *what and where to pay attention*.
- **Spatial and temporal attention:** This category combines the advantages of both spatial and temporal attention. This mechanism acts as a dynamic spatial region and time-frame selection process, thus determining *where and when to pay attention*.

The application of attention mechanisms in computer vision is almost contemporaneous with the first proposals for attention mechanisms in the language, text, and speech domains. Below we present several representative examples from each category.

With respect to *channel attention*, Hu *et al.* [52] proposed the *squeeze-and-excitation* block. This module adaptively recalibrates channel-wise feature responses by explicitly modelling dependencies between channels. This strengthens the representational power of CNNs by improving the quality of spatial encodings across the feature hierarchy. These properties have been explored as fundamental components of attention mechanisms in [53] and [54]. Still on the scope of channel attention, a different research line, related to the *self-attention* mechanism, was further explored in [55]. In this work, the authors proposed *attention augmentation* mechanisms combining both convolutions and self-attention by concatenating the convolutional feature maps with a set of feature maps generated by self-attention. This approach showed improvements in image classification and object detection while maintaining the number of parameters. In [56], the authors addressed the topic of one-shot learning and proposed a neural network architecture that uses the semantic representation (i.e., the embedding) of the label to obtain attention maps that are used to generate image features. They also developed a *multiple-attention* scheme to extract meaningful information from the input images and used an auxiliary training set to learn the attention weights.

Concerning *spatial attention*, Mnih *et al.* [57] designed a novel RNN capable of extracting information from an image or video by adaptively selecting a sequence of regions or locations and only processing these regions at a high resolution. Hence, this method can contribute to a high degree of translation invariance and the ability to process variable input sizes. Dai *et al.* [58] developed the *deformable* *convolution*, which adds flexibility to the regular (spatial) sampling grid of the standard convolution. The intuition behind this approach is that CNNs with these modules extract better features related to the objects of interest in the images (with particular focus on non-rigid objects), thus increasing their robustness to affine transformations. Wang *et al.* [59] revisited the concept of non-local attention and designed the non-local networks, which compute the response at a given spatial location as a weighted sum of the features at all positions in the features. Hu *et al.* [60] addressed the issue of *context exploitation* in CNNs and proposed a framework consisting of a pair of operators: the *gather* operator, which aggregates contextual information across large neighbourhoods of each feature map, and the *excite* operator (inspired by [52]), which redistributes the pooled information to local features.

Regarding *temporal attention*, Xu *et al.* [61] have proposed the *attentive spatial-temporal pooling network* architecture, which operates on video sequences by learning a similarity metric over the features extracted from recurrent and convolutional layers, and computing the attention weights on spatial (i.e., regions in each frame) and temporal (i.e., frames over sequences) levels. Similarly, Chen *et al.* [62] developed an attention mechanism to optimise the similarity learning among short-video snippets, and Zhang *et al.* [63] proposed the *self-and-collaborative* attention network, which also learns a similarity metric between feature representations of video-pairs.

Regarding *branch attention*, Srivastava *et al.* [64] proposed the concept of *highway networks* which consists of deep neural networks with an LSTM-inspired attention gate that allows for computation paths along which information can flow across many layers (i.e., *information highways*). This work can be considered the precursor of this concept of branch attention. Later, Li *et al.* [65] designed *selective kernel networks*, which contain multiple CNN branches with different kernel sizes that are fused using an attention mechanism, guided by the information in these branches. Zhang *et al.* [66] presented a deep neural architecture that contains several replicated branches (i.e., the *cardinal* branches) that contain more replicated branches (i.e., the *split* branches). In this work, the information fusion is performed with an attention mechanism inspired by [52]. Chen *et al.* [67] proposed the *dynamic convolution*, which employs multiple parallel convolution kernels dynamically and aggregates this information through a non-linear attention mechanism.

With respect to *channel and spatial attention*, Woo *et al.* [68] have designed a convolutional block attention module that contains a channel attention module followed by a spatial attention module. According to the authors, this block can be easily integrated into any architecture and enables the regression of finer features for object detection. Following this research line, Zhao *et al.* [69] have proposed the *point-wise spatial attention network* for the task of image segmentation. This network processes contextual information from all positions in the feature map and uses a *self-adaptive* attention mechanism to connect all of them. Using generative adversarial networks (GANs), Zhang *et al.* [70] explored a *self-attention* mechanism to enable both the generator and the discriminator to efficiently model relationships between widely separated spatial regions.

In terms of *spatial and temporal attention*, Du *et al.* [71] introduced the *recurrent spatial-temporal attention network* in which they used an end-to-end LSTM-based network with a spatio-temporal module to process video data, and developed an attention-based mechanism to fuse this information. Song *et al.* [72] used RNNs and LSTMs together to process video data and added spatial and temporal attention modules, whose information is later fused. Gao *et al.* [73] addressed the problem of *visual captioning* with hierarchical LSTM models using spatio-temporal attention to select specific regions or frames to predict the associated words and with adaptive attention to model the importance of visual information or the language contextual information. Yan *et al.* [74] have also addressed the problem of visual captioning with an architecture based on a CNN-encoder and an LSTM-decoder with spatio-temporal attention to extract spatial and temporal features in a video and guide the decoder to automatically select the most significant regions in the most relevant temporal segments for word prediction.

In addition, several works focus on multi-modal data, thus mixing both the computer vision and the language, text, and speech domains. In [50], the authors revisited the concepts of *hard* and *soft* attention. They implemented an architecture consisting of a convolutional block to extract image features and an LSTM network with attention to generate the image description (i.e., automatic image captioning).

Although CNN-based architectures had achieved high performance, the transition to Transformer-based architectures occurred almost naturally. Wu *et al.* [75] discussed three open challenges that motivated the natural implementation of Transformer-based architectures for computer vision tasks: convolutions uniformly process all image patches regardless of their importance, which can lead to spatial inefficiency (e.g., image classification models should prioritize foreground over background); not all images have all concepts, therefore, applying high-level filters to all images could be computationally inefficient (e.g., features related to an image of a person may not be present in an image of a flower); and, the convolution operation is not good at establishing relationships between spatially-distant concepts, which is an essential property in computer vision. For this reason, Dosovitskiy *et al.* [18] have proposed the *Vision Transformer* architecture. Analogously to the original Transformer, to work with a Vision Transformer, one needs to split an image into patches and provide the sequence of linear embeddings of these patches as the input. Contemporaneously, Wu *et al.* [75] proposed a similar architecture that aims to encode semantic concepts in a smaller number of *tokens* and establish a relationship between spatially-distant concepts through an attention mechanism in this token space. Several modifications of the baseline architecture and their application to different tasks have already been reported [76], [77], [78], [79].

III. LITERATURE REVIEW

This section provides a comprehensive review of the integration of attention mechanisms into algorithms for medical applications. For the sake of clarity, we structure this section into different types of tasks that can integrate different workflows of the medical domain: medical image classification, medical image segmentation, medical report understanding, and other tasks (medical image detection, medical image reconstruction, medical image retrieval, medical signal processing, and physiological and pharmaceutical research). The work presented in this section results from the analysis of existing surveys (referenced in the Introduction) and a search in Google Scholar[2] using the terms ''attention mechanism'', ''deep learning'', and ''medical''.

A. MEDICAL IMAGE CLASSIFICATION

Below, we group the works of the most common approaches in recent works.

1) BREAST LESION CLASSIFICATION

In a multi-database and multi-class classification (i.e., benign, malignant, and normal) approach for breast lesions in ultrasound images, Gheflati and Rivaz [80] studied the impact of using data augmentation, and transfer learning on Transformer- and CNN-based architectures, and reported similar, and sometimes improved, results. In a multiple instance learning approach (MIL), Ilse *et al.* [81] proposed a gated attention mechanism in deep neural networks for the classification of histopathological images (the authors also investigated the application of this methodology to colon cancer).

2) COVID-19 CLASSIFICATION

In the task of Coronavirus disease (COVID-19) classification, Han *et al.* [82] proposed a MIL algorithm for the automated screening of COVID-19 from volumetric chest computed tomography (CT) scans that employs an attention mechanism in the pooling operation to select the instances that may be meaningful for the final classification. Perera *et al.* [83] proposed a lightweight Vision Transformer that uses ultrasound images to discriminate between COVID-19, bacterial pneumonia and healthy specimens. Jiang and Lin [84] proposed an architecture that combines two Transformer-based models to classify chest X-ray images as COVID-19, pneumonia, and healthy specimens. Liu and Yin [85] used X-ray image data and applied a Transformer architecture with transfer learning to train a model capable of discriminating between normal and COVID-19 cases. Park *et al.* [86] explored the potential of Vision Transformer architectures in a federated learning setting for the classification of COVID-19 in chest X-ray images and reported that these architectures are suitable for collaborative learning in medical imaging. Hsu *et al.* [87] approached chest CT scan data with an architecture that uses Transformer-based blocks to learn the context of the CT slices, the information at the pixel level, and the spatial-context features with the aid of a deep Wilcoxon signed-rank test to determine the importance of each slice. Zhang and Wen [88] proposed a framework for 3D CT scans that consists of a U-Net [89] module to segment the lungs and a Transformer-based module to extract features and perform classification (i.e., COVID-19 diagnosis). Mondal *et al.* [90] proposed a Vision Transformer with a multi-stage transfer learning strategy that uses CT and X-ray image modalities to discriminate between COVID-19, pneumonia, and healthy specimens. Gao *et al.* [91] tested a Vision Transformer on 2D and 3D medical image data from CT scans and reported improved results against common CNN architectures for COVID-19 classification (i.e., COVID-19 against non-COVID-19). Similarly, Shome *et al.* [92] also employed a Vision Transformer on 2D chest CT X-ray images and tested its predictive performance in both binary (i.e., COVID-19 against healthy specimens) and multi-class (i.e., COVID-19 against bacterial pneumonia against healthy specimens) settings, reporting improved results against common CNN architectures. Ambita *et al.* [93] used a Vision Transformer to detect COVID-19 in CT scan images and employed a data augmentation strategy that relied on the generation of synthetic images with a self-attention-based GAN. Park *et al.* [94] explored the use of a Vision Transformer architecture for COVID-19 classification that uses low-level features generated using a backbone network. The authors also performed several experiments using chest X-ray databases from different institutions.

3) WHOLE SLIDE IMAGE CLASSIFICATION

Lu *et al.* [95] addressed the problem of glioma subtype classification (i.e., a multi-class task) in whole slide image (WSI) data using a contrastive training framework that employs a CNN backbone to learn relevant patch-level feature representations, and a sparse-attention block to aggregate the features of these multiple patches (i.e., multiple instances). Chen *et al.* [96] employed a Transformer-CNN-based architecture to classify gastric histopathology WSIs in a binary setting (i.e., normal against abnormal). Jiang *et al.* [97] addressed the diagnosis of acute lymphocytic leukemia through the classification of leukemic B-lymphoblast cells (i.e., cancer cells) and B-lymphoid precursors (i.e., normal cells) with a Transformer-CNN ensemble, and a data enhancement method that tackles the problem of class imbalance. Zheng *et al.* [98] proposed a novel graph-based Vision Transformer architecture to classify lung WSIs (i.e., adenocarcinoma, squamous cell carcinoma, and normal histology).

4) RETINAL DISEASE CLASSIFICATION

Bodapati *et al.* [99] proposed an architecture for the automatic diagnosis of diabetic retinopathy that uses multiple pre-trained CNNs with spatial pooling to extract features

[2] https://scholar.google.com/

from color fundus retinal images and integrates gated attention blocks to guide the model to focus more on lesion portions of the retinal images while paying less attention to the non-lesion regions. Yu *et al.* [100] addressed the task of retinal disease classification using fundus image data with a Vision Transformer trained under a MIL setting. Similarly, in diabetic retinopathy classification, Sun *et al.* [101] proposed a lesion-aware Transformer architecture that jointly learns to detect the presence of diabetic retinopathy and the location of lesion discovery, using an encoder-decoder structure. Several authors have addressed the task of diabetic retinopathy recognition in a multi-class setting (i.e., no diabetic retinopathy, mild non-proliferative diabetic retinopathy, moderate diabetic retinopathy, severe non-proliferative diabetic retinopathy, and proliferative diabetic retinopathy) with a Vision Transformer architecture [102], [103]. Yang *et al.* [104] used a hybrid CNN-Transformer architecture to tackle ophthalmic image data in a multi-class setting (i.e., normal, diabetes, glaucoma, cataract, age-related macular degeneration, hypertension, myopia, and other abnormalities) using different data pre-processing strategies.

5) MISCELLANEOUS CLASSIFICATION TASKS

In lung cancer classification in screening CT, Al-Shabi *et al.* [105] have extended the study of non-local attention and proposed an architecture that is initially trained on simple examples and gradually grows to increase its ability to handle the task at hand, as the classification task becomes more difficult. Moranguinho *et al.* [106] applied this methodology for the classification of lung biopsy histopathological images with respect to cancer classification and used post-model interpretability algorithms to assess the regions of interest for the predictions of their trained deep neural network. He *et al.* [107] proposed a deep learning architecture for depression recognition that combines a backbone CNN to extract features, a local attention module that focuses on parts of the input images, a global attention module to learn global patterns from the entire input images, and a weighted spatial pyramid pooling layer to learn the depression patterns after the feature aggregation operation. Dai *et al.* [108] explored the task of multi-modal and multi-class classification in head, neck, and knee magnetic resonance imaging (MRI) data using a hybrid CNN-Transformer model, where the CNN module is used as a low-level feature extractor. Datta *et al.* [109] performed a comparative study of the effect of *soft-attention* combined with different backbone architectures in skin cancer classification and reported several advantages against conventional methodologies. Barhoumi *et al.* [110] proposed a model that aggregates several feature maps extracted using multiple Xception CNNs [111] and uses these features to train a Vision Transformer for the intracranial hemorrhage classification problem, using CT images. Liang and Gu [112] added an attention mechanism to a backbone network based on the ResNet [113] to aid the diagnosis of Alzheimer's disease in brain MRI data.

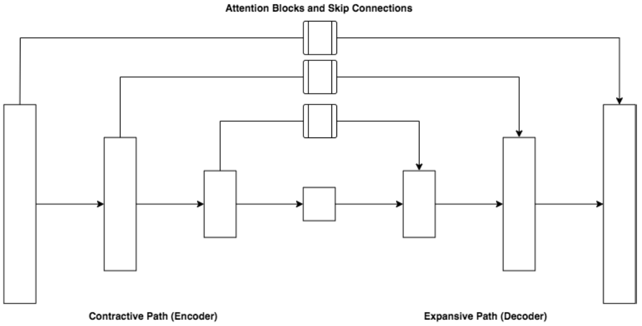

Attention Blocks and Skip Connections
Contractive Path (Encoder)
Expansive Path (Decoder)

FIGURE 2. U-Net architecture with attention blocks, as proposed in [123]. Please note that rectangles represent the feature maps of the encoder and decoder, horizontal arrows represent the regular flow of information, vertical elbow arrows represent the skip connections, and squares represent the attention blocks.

B. MEDICAL IMAGE SEGMENTATION

Using well-known data sets related to different use cases and working on top of [52], Roy *et al.* [114] verified that the addition of a spatial-channel squeeze and excitation block works as an attention mechanism in FCNs and improves the quality of the segmentation maps. Interestingly, in most recent papers, the authors approach the problem of medical image segmentation using encoder-decoder structures that generally comprise a CNN and a Transformer to extract volumetric spatial feature maps, perform global feature modeling and predict refined segmentation maps [115], [116], [117], [118], although there are already some works that replace the CNN-based modules at the encoder or decoder levels and integrate other attention mechanisms to extract features and model long-range dependencies [119]. More recent methodologies on medical image segmentation are taking advantage of a hybrid use of the Vision Transformer and the U-Net with improved results regarding the quality of segmentation maps [120], [121]. Wu *et al.* [122] proposed a *feature adaptive* model that comprises a dual encoder with CNN and Transformer branches to simultaneously capture both local features and global context information. The U-Net architecture [89] has high relevance in biomedical image segmentation. The high flexibility of this network's structure allows the integration of several computing blocks, such as attention mechanisms (see Fig. 2). Hence, as described below, most authors employ a similar structure in their applications.

Below, we group the works of the most common approaches in recent works.

1) BREAST LESION SEGMENTATION

Using breast ultrasound images, Vakanski *et al.* [124] integrated an attention block into the well-known architecture U-Net to learn semantic representations that prioritize spatial regions for the task of breast tumor segmentation. In breast tumor segmentation, using ultrasound imaging data, Zhu *et al.* [125] developed a Transformer-based architecture that incorporates prior information of the region of tumors to obtain accurate segmentation. Following the most recent approaches of Transformer-based architectures for medical image analysis, Liu *et al.* [126] proposed a U-Net-like model that integrates an attention module to focus on the tumor region and combines CNN and Transformer blocks to extract and process features that may lead to improved segmentation maps.

2) COMPUTATIONAL PATHOLOGY

Using microscopic images, Prangemeier *et al.* [127] proposed an architecture based on a CNN-encoder and Transformer-decoder for the *classification* and *instance segmentation* of yeast cells. Using microscopy images of corneal endothelial cells, Zhang *et al.* [128] revisited the problem of cell segmentation with the proposal of the multi-branch hybrid Transformer network, which follows the main structure of its predecessors (i.e., it contains both CNN and Transformer modules) to extract and process local and spatial features that may lead to refined segmentation maps. Shao *et al.* [129] addressed MIL using a Transformer-based architecture that explores morphological and spatial information and considers the correlation between instances. The authors tested their framework in different computational pathology problems. Using high-resolution images, Nguyen *et al.* [130] performed a comparative analysis using architectures based on CNN and Transformer modules and reported improved results with the Transformer-based approaches. In a similar approach, using hyperspectral image data, Yun *et al.* [131] proposed an encoder-decoder architecture with CNN and Transformer blocks to extract and model spatial and spectral features, achieving competitive results against other methodologies.

3) CARDIAC SEGMENTATION

Li *et al.* [132] addressed the task of cardiac image segmentation using MRI data from different domains with a generative adversarial model with an attention loss, proposed to translate the images from existing source domains, and a stack of data augmentation techniques to simulate real-world transformation to boost the segmentation performance for unseen domains. Concurrently, Kong and Shadden [133] tackled the same problem with a GAN with a cycle consistency loss (i.e., CycleGAN [134]) to generate images in different styles, exchanging the low-frequency features of images from different domains, and using an architecture based on an attention-gated U-Net that should learn to focus on cardiac structures of varying shapes and sizes while suppressing irrelevant regions. These strategies have been followed by the community in other contexts as well [135]. In the task of cardiac segmentation, using ultrasound images (i.e., echocardiography), Deng *et al.* [136] developed an architecture that combines a CNN-based encoder-decoder to extract multi-level features, a Transformer structure to fuse these features, and a patch embedding layer created to reduce the number of parameters of the embedding layer and the size of the token sequence. Chen *et al.* [137] proposed a U-Net-like architecture that studied the potential of several settings (e.g., Transformer as encoder, hybrid CNN-Transformer as encoder) that reported increased performances for cardiac segmentation (the authors also reported increased performances on multi-organ segmentation). Huang *et al.* [138] also proposed a Transformer-based model that tackles several tasks: the alignment of features, the enhancement of the long-range dependencies and local context, and the modeling of the long-range dependencies and local context of multi-scale features. They experimented with this architecture on cardiac segmentation databases (also on multi-organ segmentation).

4) MULTI-TASK SEGMENTATION

Oktay *et al.* [123] addressed the task of multi-organ segmentation (i.e., liver, pancreas, spleen) and worked on top of the U-Net architecture by integrating attention modules in the skip connections to promote the learning of relevant features while suppressing feature activations in irrelevant regions. Yao *et al.* [139] proposed a U-Net-based architecture with deep feature concatenation and an attention mechanism branched into several attention gates for the task of scleral blood vessel segmentation. Following this methodology, Chang *et al.* [140] added a Transformer module to the encoder block of the baseline model and reported improved results on multi-organ segmentation tasks. Using CT images, Xie *et al.* [141] proposed a framework that connects a CNN and a Transformer to extract both feature representations and model the long-range dependency on these feature maps. This model also employs a self-attention mechanism that focuses on specific positions of the image, thus reducing the computational complexity. Similarly, for kidney segmentation using CT images, Shen *et al.* [142] also proposed an encoder-decoder architecture that leverages the interaction of CNN and Transformer modules to learn multi-scale features to achieve improved segmentation results. Following this work, Zhang *et al.* [143] developed an architecture for multi-organ segmentation that runs a CNN-based encoder and Transformer-based segmentation network in parallel and fuses the features from these two branches to jointly make predictions. Tang *et al.* [144] proposed a U-Net shaped architecture with a Transformer module for multi-organ segmentation with 3D images and employs a specific pre-training strategy with contrastive learning, masked volume inpainting, and 3D rotation prediction. Using images from different modalities (i.e., MRI and CT) and organs (i.e., brain cortical plate, hippocampus, pancreas), Karimi *et al.* [145] proposed an architecture based entirely on Transformers that uses self-attention between neighboring image patches, without requiring any convolution operations. Cao *et al.* [146] also proposed a U-Net architecture composed solely of Transformer-based modules for multi-organ segmentation that uses skip-connections for local-global semantic feature learning. On colonoscopy images, Sang *et al.* [147] proposed an architecture that takes advantage of a ResNeSt *backbone* [66] connected to the coupled U-Nets [148] architecture and several attention gates to combine multi-level features and yield accurate polyp segmentation. In a multi-task segmentation approach, Zhang *et al.* [149] developed a method that seeks to integrate multi-scale attention and CNN feature extraction using a pyramidal network architecture, by working on multi-resolution images. Lin *et al.* [150] proposed an encoder-decoder architecture comprised of hierarchical Transformer-based modules to model the long-range dependencies and the multi-scale context connections during the process of down-sampling and up-sampling. Li *et al.* [151] proposed a different framework based on Transformers, that uses a squeeze-and-excitation attention module to both regularize the self-attention mechanism of Transformers and learn diversified representations. Ranjbarzadeh *et al.* [152] addressed brain tumor segmentation in multi-modal MRI data using a cascade CNN that processes both local and global features in two different routes, combined with a distance-wise attention mechanism that considers the effect of the location of the center of the tumor and the brain. Ribeiro [153] also addressed the task of brain tumor segmentation and proposed an architecture inspired by the U-Net that integrates an attention mechanism with adaptive-scale context to capture multi-scale information by processing parallel branches with different scales. Petit *et al.* [154] designed a U-shaped architecture for image segmentation with self-attention and cross-attention from Transformers (i.e., U-Transformer). They performed experiments on two abdominal CT-image databases and obtained superior performance when compared to U-Net-based models. Regarding skin lesion segmentation, Wang *et al.* [155] proposed a *boundary-aware* Transformer that aims to take advantage of *boundary-wise priors* through a boundary-wise attention gate that highlights the ambiguous boundaries to generate attention maps to guide the training. For tooth root segmentation, Li *et al.* [156] proposed a Transformer-based architecture with the structure of the U-Net backbone and optimized the training of this model with a Fourier Descriptor loss function that takes advantage of prior knowledge related to shape. Li *et al.* [157] explored the same task through an anatomy-guided multi-branch Transformer with multi-head attention that employs a polynomial curve fitting segmentation strategy (based on keypoint detection) to extract anatomy features. Following the work described in [157], Guo *et al.* [158] implemented a similar approach for retinal vessel segmentation. Regarding this task, several other approaches were proposed. Wang *et al.* [159] introduced a network that focuses on refining segmentation errors by using an attention mechanism that considers the differences between the ground truth and the predicted masks as the (source of) supervision for learning the attention maps. Guo *et al.* [160] proposed a network that builds upon *residual* blocks and spatial attention to promoting the extraction of complex vascular features while alleviating overfitting. Du *et al.* [161] proposed a U-Net-based architecture that contains *pyramid pooling* [162] modules and integrates an attention mechanism in its skip connections to promote the ability to obtain global information and efficient learning of semantic features. Gao *et al.* [163] developed a deep network that combines a feature fusion *residual* module with channel and spatial attention mechanisms to promote the extraction of low and high-dimensional features and effectively explore feature dependencies at spatial and channel levels. Amer *et al.* [164] created a U-Net architecture that comprises *residual* blocks and a multi-scale spatial attention module to capture and select multi-scale context information while ensuring an effective fusion of this information. Zhao *et al.* [165] also proposed a U-Net architecture that focuses on extracting better features at spatial and channel levels. After revisiting the work described in [157], Jin *et al.* [166] employed a 3D U-Net architecture with *residual* blocks and an attention module to extract volumes of interest in liver and brain CT imaging data and perform tumor segmentation. Sobirov *et al.* [167] applied the Transformer-based model proposed by Hatamizadeh *et al.* [118] to the tasks of head and neck tumor segmentation using multi-modal data (i.e., CT and PET images) and compared their results with CNN-based approaches. Yan *et al.* [168] also employed a U-Net-based structure for the task of multi-organ segmentation in 3D medical image data, however, with slightly different changes: in their approach, they used a CNN-encoder and CNN-decoder with a Transformer model in between to fuse contextual information in the neighboring image slices.

C. MEDICAL REPORT UNDERSTANDING

1) MEDICAL REPORT GENERATION

Using two medical databases containing radiological and pathological images, Jing *et al.* [169] proposed a methodology consisting of a multi-task deep learning model that jointly performs the prediction of tags and the generation of paragraphs, a *co-attention* mechanism for locating regions containing abnormalities and generating narrations for them, and a hierarchical LSTM model for generating long paragraphs. One of the first mentions of using Transformer architectures to generate medical image reports is described in [170]. In this work, the authors used X-ray data and proposed a hierarchical framework trained with reinforcement learning. This framework consists of an encoder that extracts visual features through a bottom-up attention mechanism aimed at identifying regions of interest and extracting top-down visual features, and a Transformer-based non-recurrent captioning decoder aimed at generating a coherent paragraph of medical image reports. Lovelace and Mortazavi [171] developed a chest X-ray report generation algorithm that consists of two training stages: the training of the report generation model (based on a Transformer architecture) using a standard language generation objective, and, after, the fine-tuning of this model on clinical observations extracted from reports sampled from the same model to regularize the degree of coherence between the observations from the generated and ground truth reports. In a similar task, Srinivasan *et al.* [172] proposed a deep framework that uses a CNN to extract features of the images and generate tag embeddings for each image, and uses Transformers to encode the image and tag features with self-attention to get a finer representation. Alternatively, Miura *et al.* [173] applied a Transformer-based architecture trained with reinforcement learning which comprises two different rewards, namely, one that encourages the system to generate radiology domain entities consistent with the reference, and one that uses natural language inference to encourage these entities to be described in inferentially consistent ways. On the other hand, Chen *et al.* [174] proposed to generate radiology reports with a memory-driven Transformer, where a relational memory is designed to record key information of the generation process and a memory-driven conditional layer normalization is applied to incorporate the memory into the decoder of Transformer. Also in automatic chest X-ray report generation, Liu *et al.* [175] created a *contrastive-attention* methodology that compares a given input image with normal images to distill contrastive information that better represents the visual features of abnormal regions, thus providing more accurate descriptions for an interpretable diagnosis and improving the quality of the generated reports. Najdenkoska *et al.* [176] tackled the same problem with *variational topic inference* that consists of modeling a set of topics as latent variables to guide sentence generation by aligning image and language modalities in a latent space. Each topic contributes to the generation of a sentence in the report. The entire pipeline is refined with a visual attention module that enables the model to attend to different locations in the image and generate more informative descriptions. Alfarghaly *et al.* [177] proposed a Transformer-based pipeline that involves the use of a fine-tuned CheXNet model [178] to predict specific tags from the image, the computation of weighted semantic features from the predicted tag's pre-trained embeddings, and the conditioning of a pre-trained distilled GPT2 [179] model on the visual and semantic features to generate the full medical reports. Amjoud *et al.* [180] proposed an encoder-decoder framework that combines the benefits of transfer learning for feature extraction and the Transformer architecture for medical report generation. Tipirneni *et al.* [181] proposed a Transformer-based architecture to address multivariate clinical time series with missing values. A different approach was discussed in [182], where the authors propose an encoder-decoder architecture that comprises both CNN and Transformer modules and aims to explicitly quantify the inherent visual-textual uncertainties existing in the multi-modal task of radiology report generation both at the report and sentence levels, and explore a novel method to measure the semantic similarity of radiology reports, which seems to better capture the characteristics of diagnosis information. A different research line focuses on imitating the working patterns of radiologists, which typically consist of examining the abnormal regions and assigning the disease topic tags to these regions and then relying on the years of prior medical knowledge and prior working experience to write reports [183]. Hence, in [183], the authors proposed a *posterior-and-prior knowledge exploring-and-distilling* approach that consists of three modules: the *posterior knowledge explorer*, which aims to provide explicit abnormal visual regions to alleviate visual data bias; the *prior knowledge explorer*, which aims to explore the prior knowledge from the prior medical knowledge graph (i.e., medical knowledge) and prior radiology reports (i.e., working experience) to alleviate textual data bias; and the *multi-domain knowledge distiller* that aggregates this processed knowledge and generates the final reports. A very close framework was described by Liu *et al.* [184], in which the authors proposed an *unsupervised model knowledge graph auto-encoder* which accepts independent sets of images and reports during training, and consists of three modules: the *pre-constructed knowledge graph*, that works as the shared latent space and aims to bridge the visual and textual domains; the *knowledge-driven encoder* which projects medical images and reports to the corresponding coordinates in that latent space; and the *knowledge-driven decoder* that generates a medical report given a coordinate in that latent space. This modular structure is also employed by Nguyen *et al.* [185], which added three complementary modules: a CNN-based *classification* module that produces an internal checklist of disease-related topics (i.e., *the enriched disease embedding*); a Transformer-based *generator* that generates the medical reports from the enriched disease embedding and produces a *weighted embedding representation*; and an *interpreter* that uses the weighted embedding representation to ensure consistency concerning disease-related topics. Similarly, You *et al.* [186] proposed a framework, which includes two different attention-based modules: the *align hierarchical attention* module that first predicts the disease tags from the input image and then learns the multi-grained visual features by hierarchically aligning the visual regions and disease tags; and the *multi-grained Transformer* module that uses the multi-grained features to generate the medical reports. Still, on this line, Hou *et al.* [187] proposed a CNN-RNN-based medical Transformer trained end-to-end, capable of extracting image features and generating text reports that aim to fit seamlessly into clinical workflows.

2) MISCELLANEOUS TASKS

Recently, Zhou *et al.* [188] proposed a cross-supervised methodology that acquires free supervision signals from original radiology reports accompanying the radiography images, and employs a Vision Transformer, designed to learn joint representations from multiple views within every patient study. Regarding automatic surgical instruction generation, Zhang *et al.* [189] introduced a Transformer-based encoder-decoder architecture trained with *self-critical* reinforcement learning to generate instructions from surgical images, and reported improvements against the existing baselines over several caption evaluation metrics. Wang *et al.* [190] proposed a Transformer-base training framework designed for learning the representation of both the image and text data, either paired (e.g., images and texts from the same source) or unpaired (e.g., images from one source coupled with texts from another source). They applied their methodology to different chest X-ray databases and performed experiments in three different applications, i.e., classification, retrieval, and image regeneration.

D. OTHER TASKS

This subsection includes medical tasks whose interest in the field is currently increasing: medical image detection, medical image reconstruction, medical image retrieval, medical signal processing, and physiology and pharmaceutical research.

1) MEDICAL IMAGE DETECTION

Regarding colorectal polyp detection, Shen *et al.* [191] approached several colonoscopy databases and proposed an architecture constituted by a CNN for feature extraction, Transformer encoder layers interleaved with convolutional layers for feature encoding and recalibration, Transformer decoder layers for object querying, and a feed-forward network for detection prediction. Similarly, Liu *et al.* [192] addressed the same task using an encoder-decoder architecture with a ResNet50-based CNN backbone as the encoder, and a Transformer-based module as the decoder and Mathai *et al.* [193] addressed the task of lymph node detection using MRI data employing analogous strategies.

2) MEDICAL IMAGE RECONSTRUCTION

Gong *et al.* [194] addressed the reconstruction of linear parametric images from dynamic positron emission tomography (PET) and implemented a 3D U-Net [195] model with an attention layer that focuses on prior anatomical information during training to improve the quality of the reconstruction of dynamic PET images of the human brain. Liang and Gu [112] also explored the reconstruction of brain MRI data as a regularization strategy to aid the diagnosis of Alzheimer's disease.

3) MEDICAL IMAGE RETRIEVAL

Grundmann *et al.* [196] and Kim *et al.* [197] developed new methods to perform an automatic retrieval and analysis of clinical notes. Li *et al.* [198] and Ma *et al.* [199] developed new models for the automated evaluation of root canal therapy and significant stenosis detection in coronary CT angiography of coronary arteries, respectively.

4) MEDICAL SIGNAL PROCESSING

Nauta *et al.* [200] explored the combination of attention mechanisms and causal discovery techniques. Using simulated functional MRI data containing blood-oxygen-level dependent data sets for 28 different underlying brain networks, this work proposed a deep learning framework based on attention that learns a causal graph structure by discovering causal relationships in observational time series data.

5) PHYSIOLOGY AND PHARMACEUTICAL RESEARCH

Lately, deep learning has also revealed itself to be a popular framework for physiology and pharmaceutical research (e.g., drug combination prediction, protein residue-residue contact prediction). Naturally, since this field also involves high-stake decisions, the increase in the predictive performance of deep learning architectures in several problems of this topic [201], [202], [203], [204] motivated the need to explain the internal mechanisms of these algorithms and improve their explainability. Interestingly, the strategy employed by machine learning practitioners is based on attention mechanisms. Therefore, we consider it relevant to include works on this matter in this survey, even though they may not be considered direct medical applications. Filipavicius *et al.* [205] studied the use of Transformer-based language models for protein classification using long sequences. The authors also performed several experiments on the compression of the protein sequences and prepared new pre-training and protein-protein binding prediction databases. Chen *et al.* [206] presented an attention-based (i.e., regional and sequence attention) CNN for protein contact prediction and for the identification of interpretable patterns that contain useful insights into the key fold-determining residues in proteins. Wang *et al.* [207] recently proposed a deep learning framework based on graph neural networks with an attention mechanism to identify drug combinations that can effectively inhibit the viability of specific cancer cells. Khan *et al.* [208] used a Transformer-based architecture for the analysis of high-dimensional gene expression data and reported improved results on the task of lung cancer sub-types classification.

IV. CASE STUDIES

Most of the works of the previous section are often based on predictive performance (i.e., accuracy) rather than showing that these algorithms are moving towards the increase of transparency or interpretability. In a high-stake decision area such as healthcare, the authors must care about framing their proposals to the context of the application. Therefore, we decided to design an experimental set of case studies, involving three different use cases (i.e., breast lesion, chest pathology, and skin lesion), wherein the main goal is to provide insight into the impact of using attention mechanisms in deep learning algorithms for medical image applications. On the one hand, we assess this impact from the perspective of predictive performance (i.e., accuracy). On the other hand, we report post-model attribution maps obtained with different interpretability frameworks to visualize the image regions that contributed most to the final predictions. We acknowledge that the visual assessment of these attribution maps is subject to a high degree of subjectivity. However, we argue that this analysis can serve as an indirect evaluation of the degree of interpretability of these models.

A. DATA

To ensure that our conclusions are independent of the data, we selected medical image classification databases related to three different use-cases: diabetic retinopathy, chest pathologies and skin lesion classification. Table 1 presents the number of training, validation and test examples for each data set,

TABLE 1. Data splits and meaning of labels for each data set.

Data set	Splits			Labels	
	Train	Validation	Test	0	1
APTOS2019	2334	778	550	Normal	Diabetic Retinopathy
ISIC2020	19593	6568	6965	Benign	Malign
MIMIC-CXR	61203	534	1072	Normal	Pleural Efusion

and the meaning of the labels we used in this classification case study.

1) APTOS2019

The APTOS2019 data set contains retinography images taken under a variety of imaging conditions (mydriatic vs. non-mydriatic) and from different camera models and ethnicities. The data set was generated by Aravind Eye Hospital in India in cooperation with Asia Pacific Tele-Ophthalmology Society (APTOS). Each retinal image has been graded independently by two ophthalmologists, with any disagreements adjudicated by a third ophthalmologist. A retinopathy severity score was assigned to each image according to the NHS diabetic eye screening guidelines[3] as R0 (no diabetic retinopathy), R1 (background), R2 (pre-proliferative), R3 (proliferative), M0 (no visible maculopathy), M1 (maculopathy). A curated version of this data set was used in the *APTOS 2019 Blindness Detection* hosted on Kaggle.[4] For the binary classification case, we grouped all the non-R0 classes into the "diabetic retinopathy" class (see Table 1).

2) ISIC2020

The ISIC2020 [209] data set contains dermoscopic training images of unique benign and malignant skin lesions from over 2000 patients. Each image is associated with one of these individuals using a unique patient identifier. All malignant diagnoses have been confirmed via histopathology, and benign diagnoses have been confirmed using either expert agreement, longitudinal follow-up, or histopathology. The data set was generated by the International Skin Imaging Collaboration (ISIC) and images were collected by the following sources: Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and the University of Athens Medical School. The data set was curated for the *SIIM-ISIC Melanoma Classification Challenge* hosted on Kaggle.[5]

3) MIMIC-CXR

The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 [210] is a large publicly available data set of chest radiographs in the Digital Imaging and Communications in Medicine (DICOM) format with free-text radiology reports. The data set contains 227835 imaging studies for 65379 patients of the Beth Israel Deaconess Medical Centre Emergency Department (Boston, MA) between 2011–2016. Each imaging study can contain one or more images, usually a frontal and a lateral view. A total of 377,110 images are available in the data set. Studies are made available with a semi-structured free-text radiology report that describes the radiological findings of the images, written by a practicing radiologist during routine clinical care. The data set is de-identified to satisfy the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements. Protected health information (PHI) has been removed. For the binary classification case we used a subset of images of the anterior-posterior view related to the presence or absence of pleural efusion (see Table 1).

B. METHODOLOGY

We divided the experimental part of this survey into several phases, described below. We conducted all the experiments using the PyTorch library [211] for Python.

1) BACKBONE MODELS

In the first phase of this study, we trained two deep learning backbones (i.e., DenseNet-121 [212], ResNet-50 [113]) on each data set. DenseNet-121's architecture [212] allows connecting all layers (with matching feature-map sizes) directly with each other, thus improving the flow of information and gradients throughout the network, and facilitating their training. ResNet50's architecture introduced the *deep residual learning framework*, which consists of adding *skip connections* that perform *identity mapping*, and adding their outputs to the outputs of the stacked layers. We refer the reader to the original papers for more details. We used these backbone models in our experiments because they are well known and widely used in the deep learning community. In this work, we refer to these models as *DenseNet-121* and *ResNet-50*.

2) SQUEEZE-AND-EXCITATION ATTENTION BLOCK

In the second phase, we adapted the *Squeeze-and-Excitation* (SE) attention block [52] (see Fig. 3) to each of the backbones and trained these new architectures on the three data sets. According to the authors of the original paper [52], this attention block was designed to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. Since one of the original implementations uses the ResNet-50, we used the author's implementation in our study, which consists of integrating the SE block between the *residual layers* and the activation function (in this case, the *rectified linear unit* (ReLU)). Following this rationale, in the case of DenseNet-121, we integrated the SE block between the *dense* and the *transition* blocks. We decided to use the SE block in our experiments because it was one of the first proposals of channel attention for computer vision, and it is widely used by the community as a comparison reference. Besides, since we are building our case studies in medical image classification using RGB images, our interests also rely on studying the effects of attention on

[3] https://www.gov.uk/government/collections/diabetic-eye-screening-commission-and-provide

[4] https://www.kaggle.com/c/aptos2019-blindness-detection

[5] https://www.kaggle.com/c/siim-isic-melanoma-classification/overview

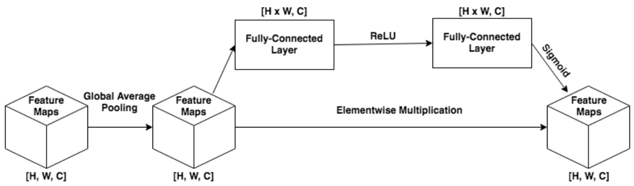


FIGURE 3. Block diagram of the *Squeeze-and-Excitation* (SE) attention block, according to [52].

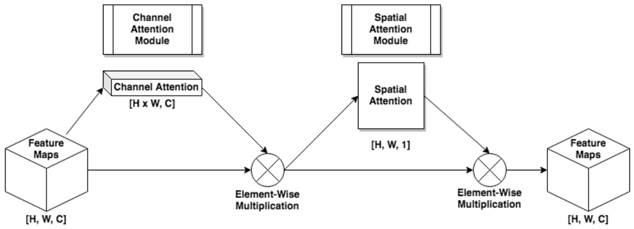


FIGURE 4. Block diagram of the *Convolutional Block Attention Module* (CBAM), according to [68].

the channel dimension. In this work, we refer to these models as *SEDenseNet-121* and *SEResNet-50*.

3) CONVOLUTIONAL BLOCK ATTENTION MODULE

In the third phase, we adapted the *Convolutional Block Attention Module* (CBAM) [68] (see Fig. 4) to each of the backbones and trained these new architectures on the three data sets. According to the authors of the original paper, the CBAM mechanism integrates two specific attention blocks: the *channel attention module*, which aims to produce a channel attention map by exploiting the inter-channel relationship of features and is considered as a feature detector (see Fig. 5); and the *spatial attention module*, which aims to generate a spatial attention map by utilizing the inter-spatial relationship of features, thus being complementary to the channel attention (see Fig. 6). Like the SE attention block, we integrated the CBAM mechanism at the same topographical location of the architectures of the backbone models. Following the decision on the SE block, we considered it important to study the spatial dimension of images. Therefore, since CBAM integrates both the attention and spatial dimensions in its pipeline, we can perceive this module as an upgrade to the SE block. Besides, it is important to note that this module is also well-known in the community, and, therefore, it is validated as well. In this work, we refer to these models as *CBAMDenseNet-121* and *CBAMResNet-50*.

4) DATA-EFFICIENT IMAGE TRANSFORMER

In the fourth phase, we tested a Transformer-based architecture composed solely of attention mechanisms. The Data-efficient image Transformer (DeiT) [79] is an architecture inspired by the Vision Transformer [18], and trained with fewer parameters. In this case, we used the *DeiT-Ti* variation [79], which has a comparable number of parameters against the chosen CNN backbones. Besides being validated by the community (as in previous examples), this Vision Transformer architecture has a proportional number of parameters (see Table 4) to the backbone models, thus allowing us to mitigate eventual model complexity effects on the predictive performance of this model. In this work, we refer to this model as *DeiT*.

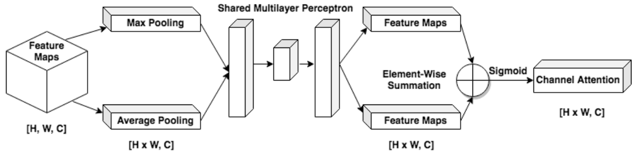


FIGURE 5. Block diagram of the *Channel Attention Module* of the *Convolutional Block Attention Module* (CBAM), according to [68].

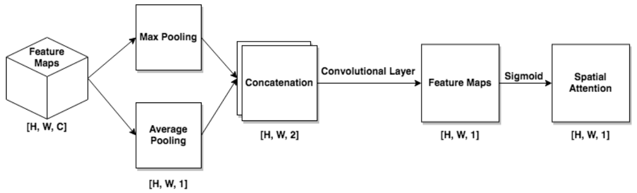


FIGURE 6. Block diagram of the *Spatial Attention Module* of the *Convolutional Block Attention Module* (CBAM), according to [68].

5) POST-MODEL INTERPRETABILITY

In the fifth phase, we generated visual explanation maps (i.e., saliency maps) using the *Deep Learning Important FeaTures* (DeepLIFT) [213] and the *Layer-wise Relevance Propagation* (LRP) [214] *post-hoc* methods. DeepLIFT [213] compares the activation of each neuron to its related *reference activation*, and assigns contribution scores according to the difference. LRP [214] is a methodology that aims to create visualizations of the contributions of single pixels to predictions. We decided to employ this strategy as a proxy for the degree of interpretability of models, thus allowing our work to be comparable to current research. We used the Captum [215] library for Python to generate these *post-hoc* saliency maps for the backbones and their attention-based versions. Regarding the DeiT, we refer the reader to [216], which proposed a framework to generate LRP attributions for Transformer-based architectures.

6) DATA PROCESSING

Regarding the data processing pipeline, all the images were resized to the dimensions of 224 × 224, and a z-normalization was applied to each RGB channel. The *mean* and *standard deviation* parameters used in this operation depend on the initialization of the weights of the model. In the case of DenseNet-121, ResNet-50, SEDenseNet-121, SEResNet-50, CBAMDenseNet-121 and CBAMResNet-50, the weights were initialized from a pre-training on the ImageNet database [217] with mean $= [0.485, 0.456, 0.406]$, and std $= [0.229, 0.224, 0.225]$. In the case of DeiT, the

TABLE 2. Data augmentation parameters used during the training step of all the phases.

Parameter	Value
Angle of rotation in degrees	$[-10, 10]$
Horizontal translation shift	0.05
Vertical translation shift	0.1
Scaling factor	$[0.95, 1.05]$
Horizontal flip probability	0.5

TABLE 3. Set of hyper-parameters used during the training step of all the phases. Please note that we present the batch size as a closed interval since we had to adapt its value during some of the training runs due to computational constraints (e.g., availability of GPU RAM).

Hyper-parameter	Value
Number of epochs	300
Loss function	Cross-entropy
Optimiser	Adam
Learning rate	1×10^{-6}
Batch size	$[2, 32]$

weights were initialized from a pre-training on the ImageNet database with mean $= [0.500, 0.500, 0.500]$, and std $= [0.500, 0.500, 0.500]$.

7) DATA AUGMENTATION

Regarding the data augmentation strategy, we decided to create a pipeline composed of several random affine transformations, i.e., random rotations, random translations, random scaling and random horizontal flip. Table 2 presents the parameters of data augmentation used in our experiments.

8) TRAINING

We trained all the models in the different databases using the same optimization hyper-parameters. During training, we saved the best model weights according to a decrease in the validation loss. Table 3 presents the optimization hyper-parameters used in our experiments. We also conducted a *low data regime* experiment, hence, all the models were trained using 1%, 10%, 50% and 100% of the training data. In imbalanced data sets (i.e., ISIC2020, MIMIC-CXR) we applied class-weights in the loss function.

C. RESULTS

1) PREDICTIVE PERFORMANCE

Figs. 7, 8 and 9 respectively present the predictive performance (i.e., accuracy) results of the different models on the APTOS2019, ISIC2020 and MIMIC-CXR data sets, using different percentages of training data.

2) MODEL COMPLEXITY

Table 4 presents the model complexity (i.e., the number of parameters) information for the models used in our experiments.

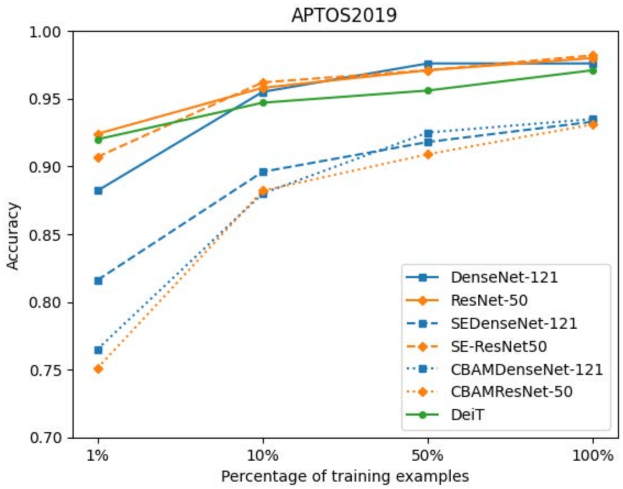


FIGURE 7. Predictive performance (i.e., accuracy) results of the different models on the APTOS2019 data set, using different percentages of training data.

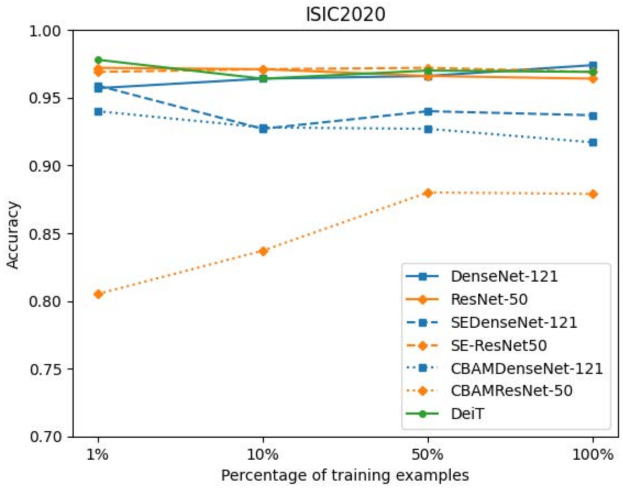


FIGURE 8. Predictive performance (i.e., accuracy) results of the different models on the ISIC2020 data set, using different percentages of training data.

TABLE 4. Model complexity (i.e., the number of parameters) information for the models used in our experiments.

Model	Number of parameters
DenseNet-121	7,054,210
ResNet-50	23,512,130
SEDenseNet-121	7,357,314
SEResNet-50	26,027,074
CBAMDenseNet-121	7,360,706
CBAMResNet-50	26,044,722
DeiT	5,486,786

3) POST-MODEL INTERPRETABILITY

Tables 5, 6 and 7 present examples of LRP and DeepLIFT *post-hoc* saliency maps obtained for images of the APTOS2019 data set with labels 0 and 1 correctly classified,

TABLE 5. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 0 correctly classified as 0 by all models.

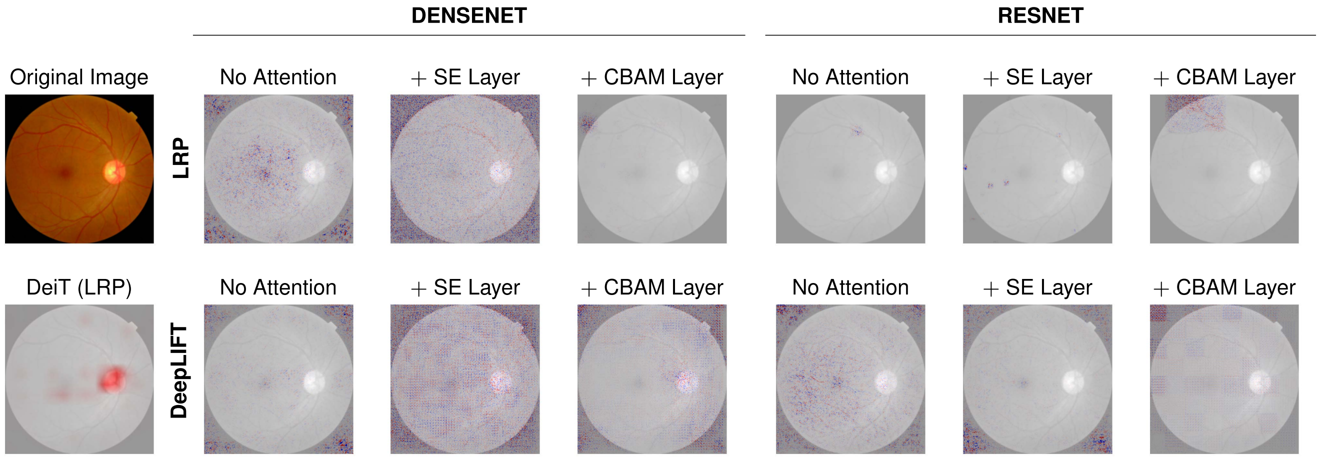
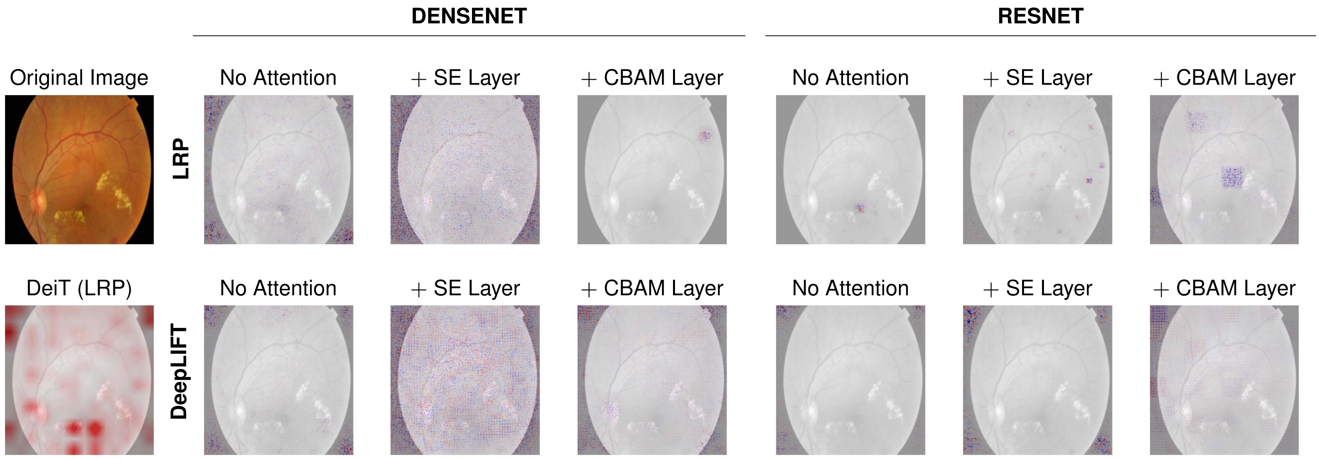


TABLE 6. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 1 correctly classified as 1 by all models.



and with label 0 incorrectly classified by all models, respectively. The same structure is followed for examples of the ISIC2020 (Tables 8, 9, 10 and 11) and MIMIC-CXR (Tables 12, 13, 14 and 15) data sets, with the addition of examples from the label 1 incorrectly classified by all models. We used the Matplotlib [218] library for Python to generate these saliency maps visualizations, using the "bwr" color map, which represents negative values in blue, zero values in white and positive values in red.

D. DISCUSSION

1) PREDICTIVE PERFORMANCE

All the experiments related to the predictive performance of deep learning models on the different data sets suggest that it is not clear that one should expect improvements in their accuracy when using attention mechanisms. Several attention-based architectures tend to lower their predictive performances against their baseline architectures. Besides, even when directly comparing a Transformer-based architecture (i.e., DeiT), the same baseline architectures seem to achieve comparable results. Given that the intuition behind attention mechanisms is that these end up learning the most relevant features, one might expect that attention-based architectures would perform better when trained in low data regimes. However, results obtained in all data sets suggest that this might not be the case. In fact, we can observe that, apart from the Transformer-based architecture, the baseline backbone models often perform better. Hence, it is reasonable to conclude that the use of attention mechanisms will not *always* bring benefits to the training of deep learning algorithms, an argument that contrasts with a trending narrative in most recent papers. On the other hand, the results reported in the literature often relate to marginal or residual improvements in the state-of-the-art backbone networks. Given that the training of deep learning algorithms is generally a stochastic process, there is a need to assess these reported improvements with a more critical view and with robust statistical tests.

TABLE 7. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 0 incorrectly classified as 1 by all models.

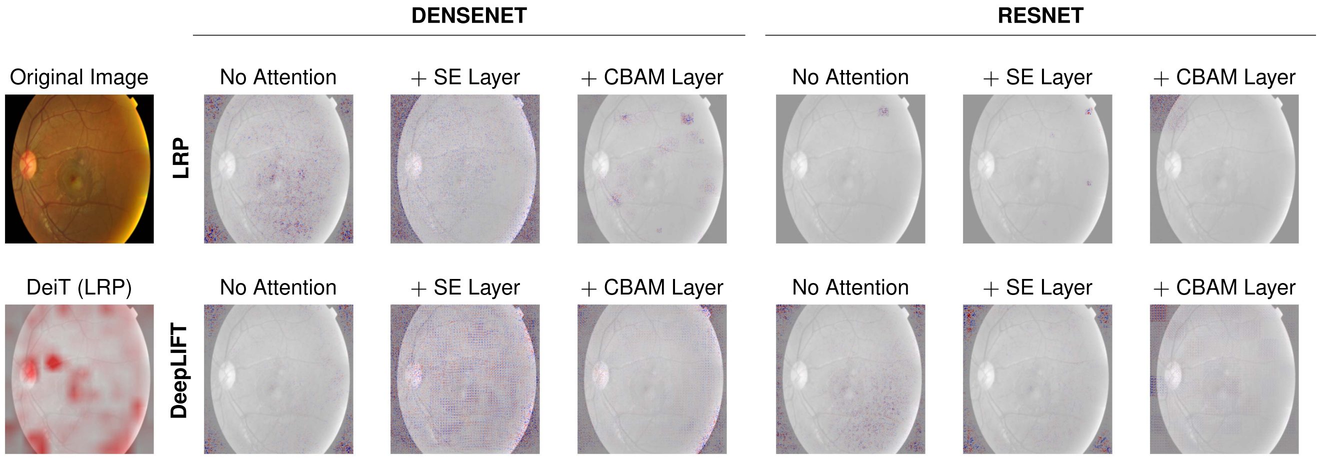


TABLE 8. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 0 correctly classified as 0 by all models.

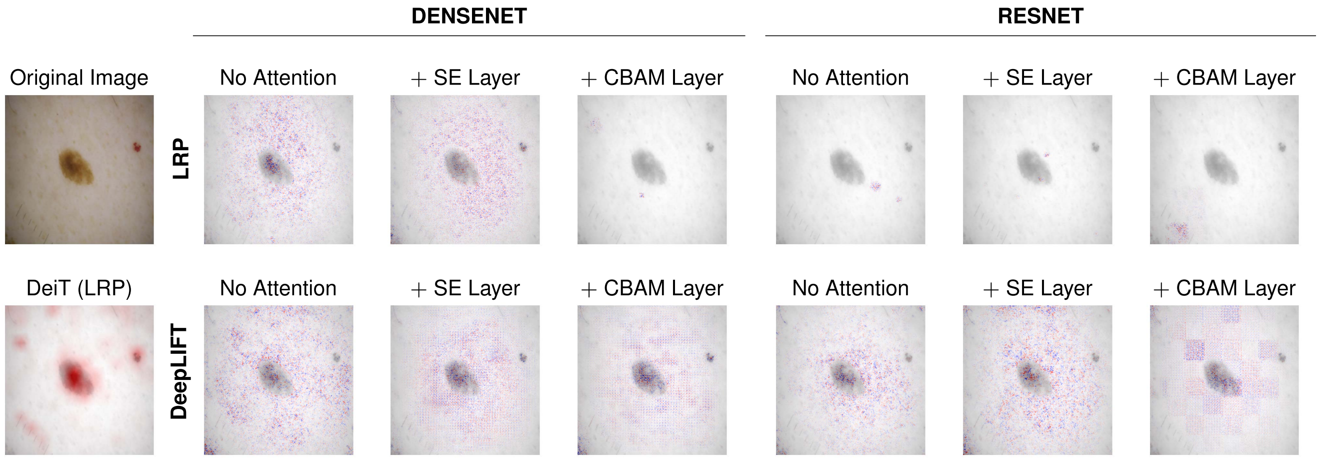


TABLE 9. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 1 correctly classified as 1 by all models.

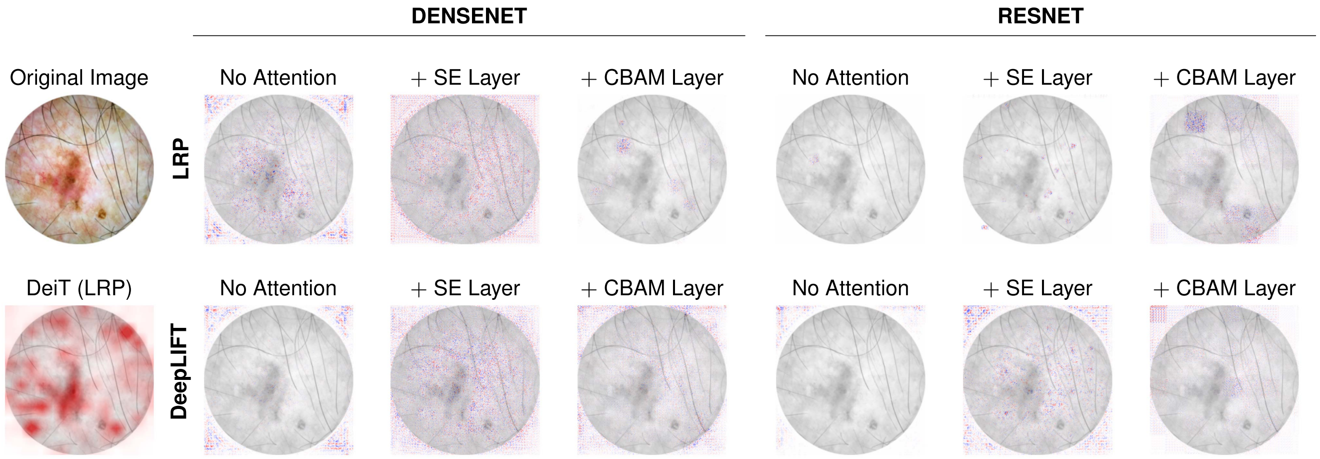


TABLE 10. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 0 incorrectly classified as 1 by all models.

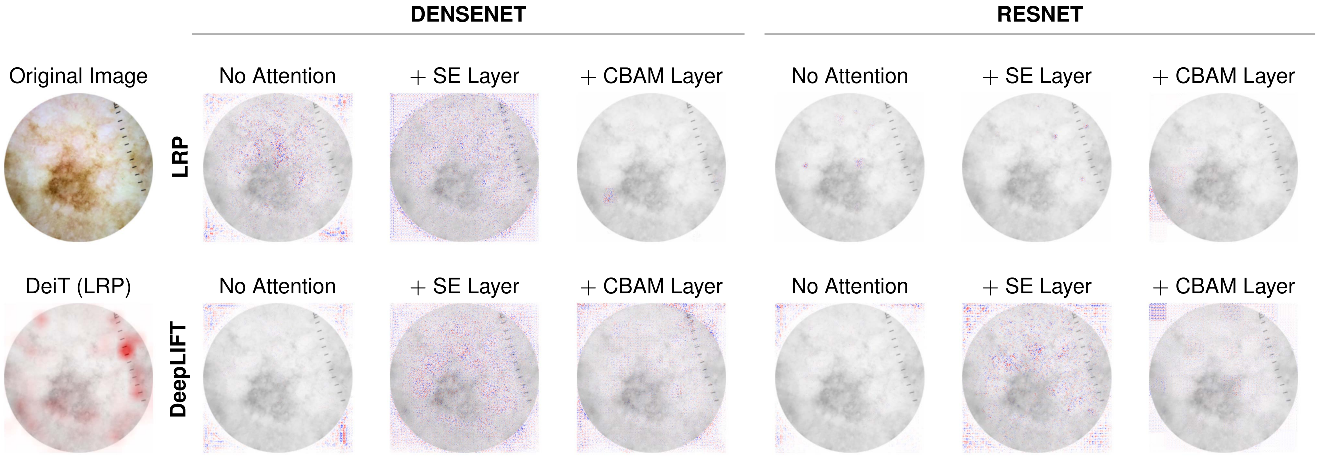


TABLE 11. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 1 incorrectly classified as 0 by all models.

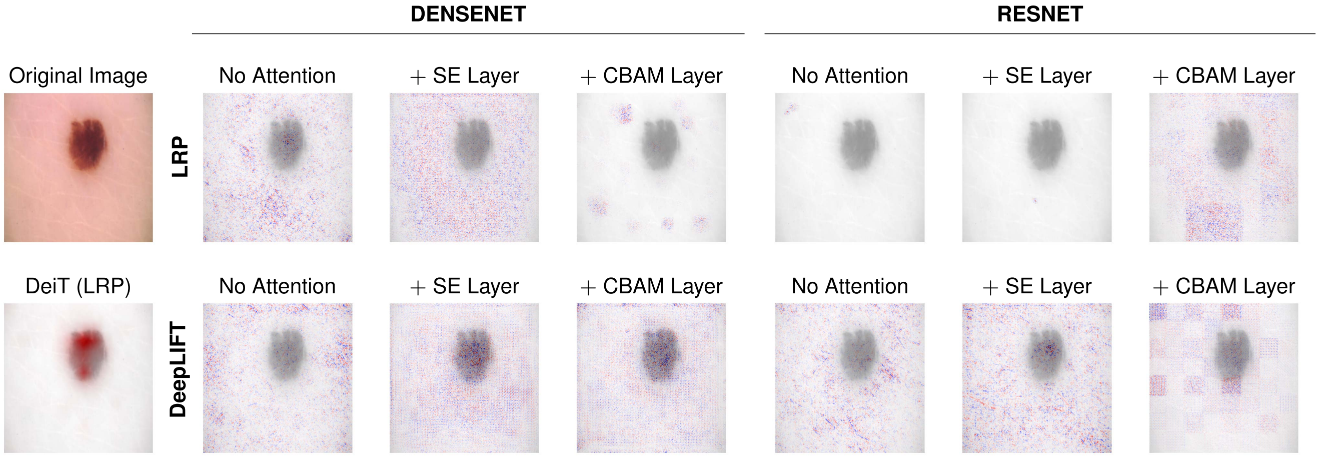


TABLE 12. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 0 correctly classified as 0 by all models.

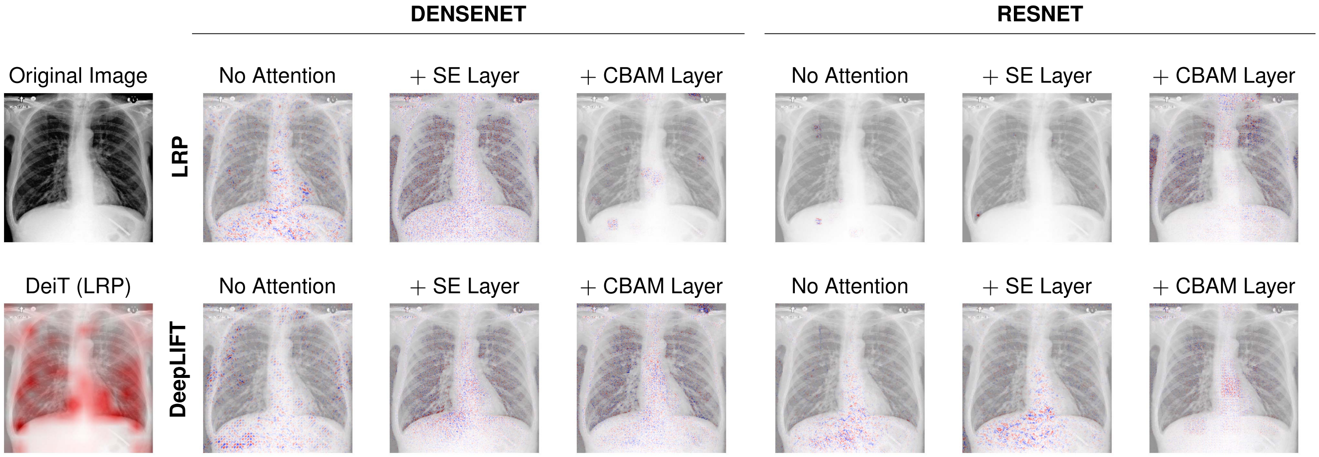


TABLE 13. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 1 correctly classified as 1 by all models.

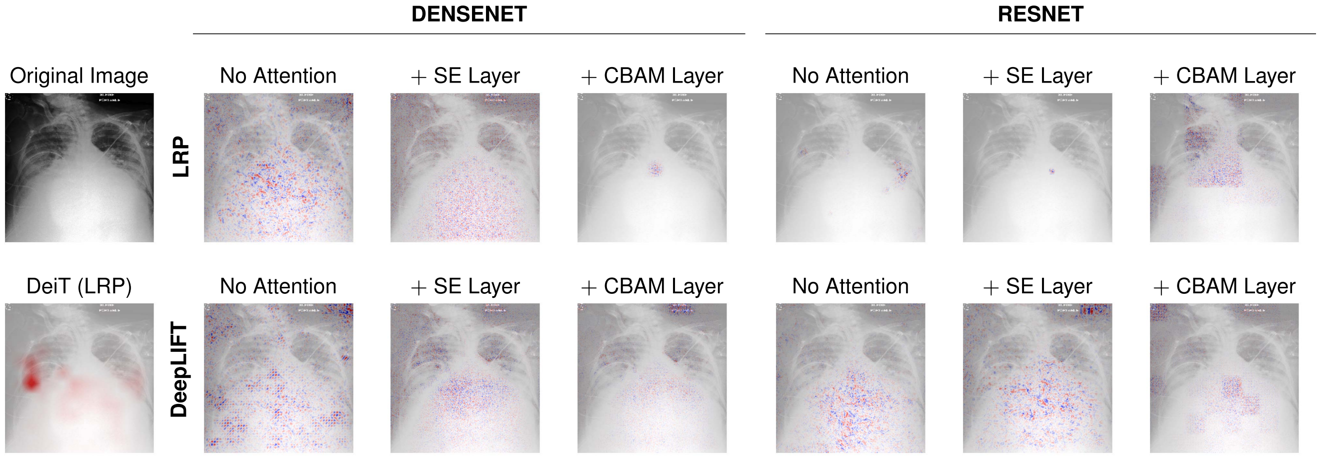


TABLE 14. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 0 incorrectly classified as 1 by all models.

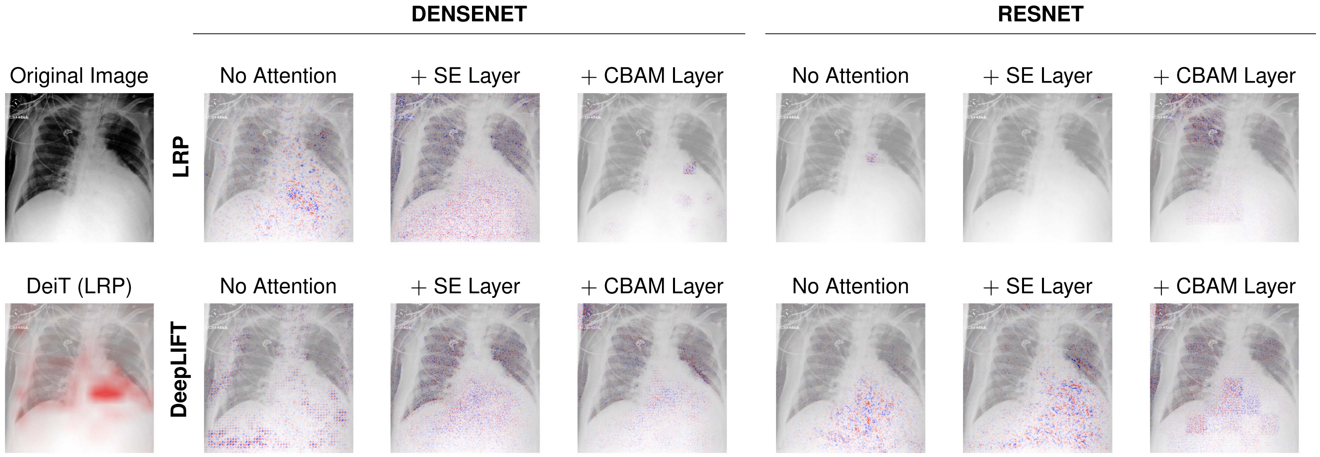
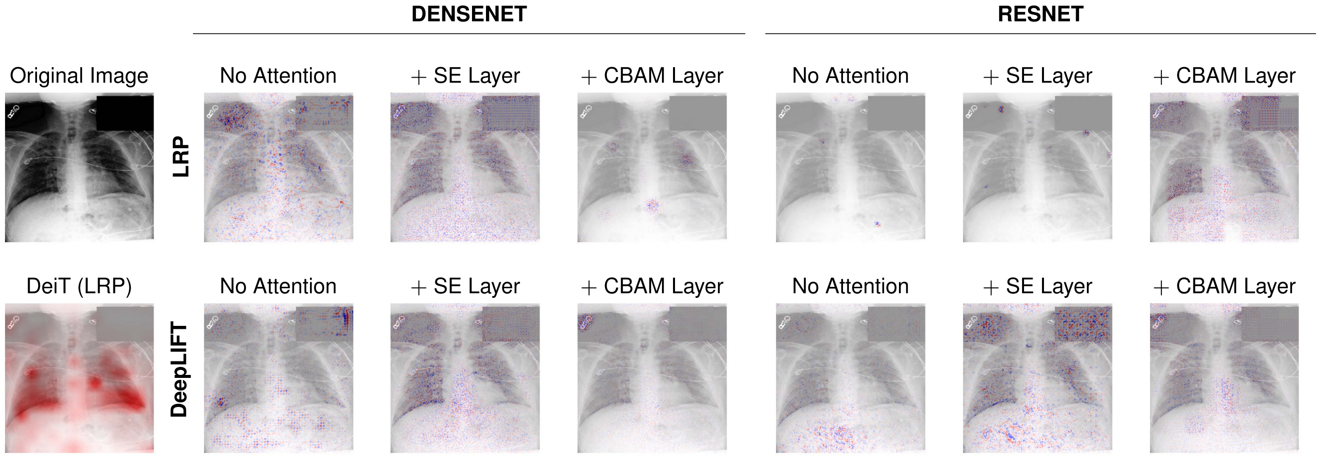


TABLE 15. Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 1 incorrectly classified as 0 by all models.



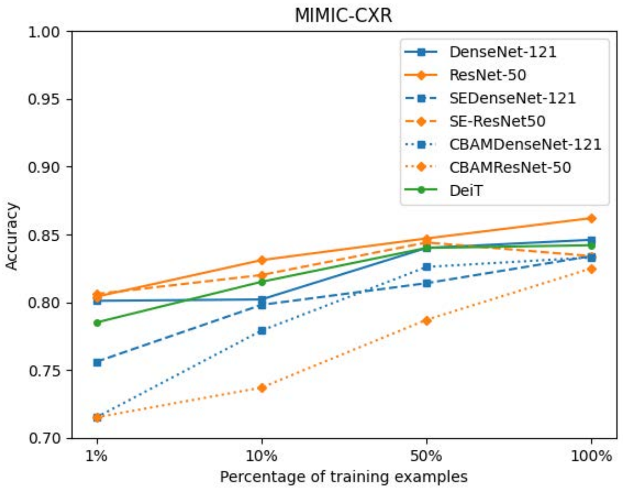


FIGURE 9. **Predictive performance (i.e., accuracy) results of the different models on the MIMIC-CXR data set, using different percentages of training data.**

2) MODEL COMPLEXITY

The integration of attention mechanisms increases the number of parameters of the deep learning models, thus increasing their complexity. This information allows us to conclude that, at least for computer vision applications, it is not necessarily true that the use of attention mechanisms contributes to the decrease of model complexity. While this may be true in several natural language processing applications, it is not correct to generalize such a rule to other types of applications. On the other hand, since these attention mechanisms often rely on simple operations (e.g., matrix multiplications), such as convolutions, we acknowledge that their use may reduce the training time of deep learning algorithms (similarly to what happened when the community started using CNNs). This aspect, though, was not objectively assessed in our experiments. However, since they are particularly efficient at modeling long-range dependencies (e.g., Transformer-based architectures), we expect these mechanisms to keep increasing their popularity and, hence, become widely used in medical image applications. Besides, Cordonnier *et al.* [219] reported that these mechanisms may operate similarly to CNNs. Another question arises from these results: are these attention-based algorithms allegedly performing well because of the inner-functioning of their attention mechanisms themselves or just because we are increasing the number of model parameters? While one may report that this issue is nonsense, we point out that some Transformer-based architectures have a considerably high number of parameters [79].

3) POST-MODEL INTERPRETABILITY

Although there are already several works that criticize the subjectivity of *post-hoc* saliency maps [220], [221], [222], [223], [224], [225], it is still common to use these methods to justify improvements related to the degree of interpretability of models. In this sense, when analyzing the results obtained with the DenseNet-121 and ResNet-50 models, we would expect that, with the increase in complexity of the attention mechanism, the distribution of the important pixels around the image would also be more focused (i.e., would have less variance). However, that does not seem to happen in our use cases. Interestingly, besides the inherent properties of the data and the task (i.e., images of the retina are different from skin images and chest X-ray images), the results drastically change when we use a different backbone. In general, ResNet-based models attain less noisy DeepLIFT and LRP saliency maps, with some exceptions in the ISIC2020 data set (e.g., Table 11). On the other hand, the LRP saliency maps obtained for the DeiT seem to highlight somewhat clear regions of the images. Besides, it is also important to remind the reader that these frameworks allow us to generate explanations even for the cases where the model miss-classifies. We also stress that one of the limitations of such analysis is that there does not exist an objective ground truth of what a high-quality visual explanation is. Therefore, it is not trivial, in computer vision tasks such as this, to conclude with complete confidence that, even for the cases where the model succeeds, it learned the right correlations. Hence, can we believe the narrative that attention mechanisms are learning the most relevant features of the image? Or is this type of analysis a result of luck in most cases? Truth is, our results still suggest that there is a high degree of subjectivity (i.e., the interpretation of these saliency maps deeply depends on the human that is interpreting them) and that there is no apparent correlation between the use of attention mechanisms and the visual aspect of the *attributions*. Besides, we stress that these visualizations shown in the literature depend on several parameters (e.g., type of color map, overlay parameter). Moreover, the method we use to generate such visualizations may often change what we expect users to observe. This is also related to another open challenge in *post-hoc* explanation methods that is the sign of the attributions. What should we expect? Positive attributions for the positive class and negative attributions for the negative class, or the other way around? Should we always normalize the sign of the attributions so it is always positive? This motivates our statement on the need for more objective methods to assess the degree of interpretability of models.

V. CONCLUSION AND FUTURE CHALLENGES

This section summarizes the main conclusions of this survey and points to future directions toward the study of attention-based algorithms for medical applications.

A. CONCLUSION

This survey presented an exhaustive overview of the use of attention mechanisms for medical applications and provided an experimental study on medical image classification that approached three different use cases. We found that backbone models can attain equivalent predictive performances to

Transformer-based architectures with equivalent model complexity (i.e., number of parameters). Moreover, when using a *post-hoc* framework to visually assess what type of features these models can extract, we can conclude that there is still a high degree of subjectivity in such analysis. The results are very noisy, even for the cases of attention mechanisms, which is counter-intuitive (i.e., with attention mechanisms, these saliency maps should have less noise). The community is moving toward using attention mechanisms (specially Transformer-based ones) and arguing that these frameworks increase the quality, transparency, and interpretability of deep learning architectures. However, we state that this is not true and that there are several open challenges one needs to address to achieve this.

B. FUTURE CHALLENGES

1) ATTENTION MECHANISMS: PAST OR FUTURE?

The research on attention-based models is in constant evolution and requires a validation and maturation step. Besides, as these models keep rising in popularity and keep convincing the community of their benefits, we acknowledge that they will appear as a common block in future architectures. However, the scientific community must keep in mind that high-stake decision areas (e.g., finance, healthcare, justice) require algorithms to be fair and transparent [224]. Therefore, even if attention mechanisms are pushing deep learning algorithms towards the limits of their predictive power, we must start thinking about creating interpretable frameworks that allow us to audit and assess these algorithms concerning the specific conditions of their domains (i.e., in the case of this work, healthcare).

2) DESIGN AND INTEGRATION OF ATTENTION MECHANISMS

A potential disadvantage of the investment in attention mechanisms is related to the design of novel strategies of attention and their integration into well-studied backbone models. Several works on attention often rely on a specific backbone architecture for their experiments. Besides, if we look at the topographies of these deep learning algorithms, it is not always clear for the users where they should place these modules, and why it makes sense to put them in a specific place. Once again, another question arises: are these attention modules dependent on the backbone into which they are integrated to? Can we consider that the reported improvements related to the predictive performance of deep learning models are solely due to the addition of an attention module? Can we come up with an objective strategy that guides future users in building and integrating attention mechanisms into their models?

3) THE RISE OF TRANSFORMERS

Apart from the current deep algorithms with attention mechanisms, it is crucial to talk about Transformer-based architectures. The attentive reader might have noticed that most of the recent approaches to the use of attention-based models in medical applications rely on Transformer-based modules. While there is hype on the use of these structures, it is not clear whether they are more interpretable or not, or if their generalization power is superior to the other deep models.

4) INTERPRETABILITY IS THE PATH TO BETTER ALGORITHMS

Even if we acknowledge that the Transformer-based algorithms are provoking a shift in the paradigm, it is not clear that they are improving the transparency of algorithms. For instance, in [79], the authors show that when training a DeiT using *knowledge distillation* techniques, the Transformer ends up learning better when the *teacher model* is a CNN. Hence, if the Transformer is learning with a non-interpretable model by design, how can we trust that the Transformer is inherently interpretable? On the other hand, given the capacity of these attention-based models to learn long-range dependencies, we must create objective techniques to assess the quality of the features learned by these frameworks, while moving towards the design of intrinsically interpretable attention-based architectures. Even if we intend to keep using visual saliency maps to explain our models, in a high stake decision field such as healthcare, we must achieve a clear standard, validated by the clinical community, of what these maps should look like and what is their effective meaning.

REFERENCES

[1] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*. Dordrecht, The Netherlands: Springer, 1987, pp. 115–141.

[2] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, nos. 10–12, pp. 1489–1506, Jun. 2000.

[3] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention," in *Proc. 5th Int. Conf. Auton. Agents*, 2001, pp. 457–464.

[4] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, nos. 1–2, pp. 507–545, Oct. 1995.

[5] D. H. Ballard, "Animate vision," *Artif. Intell.*, vol. 48, no. 1, pp. 57–86, 1991.

[6] C. Bandera, F. J. Vico, J. M. Bravo, M. E. Harmon, and L. C. Baird, "Residual Q-learning applied to visual attention," in *Proc. ICML*, 1996, pp. 20–27.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Feb. 2015. [Online]. Available: http://www.nature.com/articles/nature14539

[8] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Oct. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0893608014002135

[9] S. Jetley, N. A. Lord, N. Lee, and P. Torr, "Learn to pay attention," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=HyzbhfWRW

[10] I. Makarov, M. Bakhanova, S. Nikolenko, and O. Gerasimova, "Self-supervised recurrent depth estimation with attention mechanisms," *PeerJ Comput. Sci.*, vol. 8, p. e865, Jan. 2022.

[11] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2012.

[12] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder–decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.

[13] F. Wang and D. M. J. Tax, "Survey on the attention based RNN model and its applications in computer vision," 2016, *arXiv:1601.06823*.

[14] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021, doi: 10.1145/3465055.

[15] X. Yang, "An overview of the attention mechanisms in computer vision," in *Proc. J. Phys., Conf.*, 2020, vol. 1693, no. 1, Art. no. 012173.

[16] G. W. Lindsay, "Attention in psychology, neuroscience, and machine learning," *Frontiers in Comput. Neurosci.*, vol. 14, p. 29, Apr. 2020.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, 2021, pp. 1–22. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[19] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," 2020, *arXiv:2012.12556*.

[20] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10, doi: 10.1145/3505244.

[21] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, pp. 331–368, Mar. 2022, doi: 10.1007/s41095-022-0271-y.

[22] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, and C. Xu, "Transformers in computational visual media: A survey," *Comput. Vis. Media*, vol. 8, no. 1, pp. 33–62, 2022.

[23] F. Shamshad, S. Khan, S. Waqas Zamir, M. Haris Khan, M. Hayat, F. Shahbaz Khan, and H. Fu, "Transformers in medical imaging: A survey," 2022, *arXiv:2201.09873*.

[24] M. M. Chun, J. D. Golomb, and N. B. Turk-Browne, "A taxonomy of external and internal attention," *Annu. Rev. Psychol.*, vol. 62, pp. 73–101, 2011.

[25] B. S. Oken, M. C. Salinsky, and S. Elsas, "Vigilance, alertness, or sustained attention: Physiological basis and measurement," *Clin. Neurophysiol.*, vol. 117, no. 9, pp. 1885–1901, Sep. 2006.

[26] N. Kanwisher and E. Wojciulik, "Visual attention: Insights from brain imaging," *Nature Rev. Neurosci.*, vol. 1, no. 2, pp. 91–100, Nov. 2000.

[27] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, Psychophys.*, vol. 77, no. 5, pp. 1465–1487, 2015.

[28] F. Hutmacher, "Why is there so much more research on vision than on any other sensory modality?" *Frontiers Psychol.*, vol. 10, p. 2246, Oct. 2019.

[29] E. K. Miller and T. J. Buschman, "Neural mechanisms for the executive control of attention," in *The Oxford Handbook of Attention*. Oxford, U.K.: Academic, 2014.

[30] M. Aly and N. B. Turk-Browne, "How hippocampal memory shapes, and is shaped by, attention," in *The Hippocampus from Cells to Systems*. Cham, Switzerland: Springer, 2017, pp. 369–403.

[31] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing machines," 2014, *arXiv:1410.5401*.

[32] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[33] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.

[34] I. Sutskever, J. Martens, and G. Hinton, "Generating text with recurrent neural networks," in *Proc. 28th Int. Conf. Int. Conf. Mach. Learn.* Madison, WI, USA: Omnipress, 2011, pp. 1017–1024.

[35] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:1308.0850*.

[36] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds. San Diego, CA, USA: OpenReview, 2015, pp. 1–15.

[38] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1412–1421. [Online]. Available: https://aclanthology.org/D15-1166

[39] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1. Cambridge, MA, USA: MIT Press, 2015, pp. 577–585.

[40] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=HJtEm4p6Z

[41] S. Sukhbaatar, E. Grave, G. Lample, H. Jegou, and A. Joulin, "Augmenting self-attention with persistent memory," 2019, *arXiv:1907.01470*.

[42] S. Sukhbaatar, E. Grave, P. Bojanowski, and A. Joulin, "Adaptive attention span in transformers," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 331–335. [Online]. Available: https://aclanthology.org/P19-1032

[43] A. Fan, T. Lavril, E. Grave, A. Joulin, and S. Sukhbaatar, "Addressing some limitations of transformers with feedback memory," 2020, *arXiv:2002.09402*.

[44] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "Delight: Deep and light-weight transformer," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–19. [Online]. Available: https://openreview.net/forum?id=ujmgfuxSLrO

[45] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," 2020, *arXiv:2006.04768*.

[46] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[47] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=rkgNKkHtvB

[48] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng, "Synthesizer: Rethinking self-attention for transformer models," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10183–10192.

[49] S. Merity, "Single headed attention RNN: Stop thinking with your head," 2019, *arXiv:1911.11423*.

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[51] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5659–5667.

[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[53] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Oct. 2019, pp. 1971–1980.

[54] N. Quader, M. M. I. Bhuiyan, J. Lu, P. Dai, and W. Li, "Weight excitation: Built-in attention mechanisms in convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 87–103.

[55] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3286–3295.

[56] P. Wang, L. Liu, C. Shen, Z. Huang, A. Van Den Hengel, and H. Tao Shen, "Multi-attention network for one shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2721–2729.

[57] V. Mnih, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–19.

[58] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.

[59] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[60] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[61] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial–temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4733–4742.

[62] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1169–1178.

[63] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, and L. Lin, "SCAN: Self-and-collaborative attention network for video person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4870–4882, Oct. 2019.

[64] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[65] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 510–519.

[66] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.

[67] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11030–11039.

[68] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[69] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.

[70] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[71] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial–temporal attention network for action recognition in videos," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, Mar. 2018.

[72] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–8.

[73] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, May 2020.

[74] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "STAT: Spatial–temporal attention mechanism for video captioning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, Jan. 2020.

[75] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.

[76] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Wiesbaden, Germany: Springer, 2020, pp. 213–229.

[77] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," 2021, *arXiv:2103.10697*.

[78] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[79] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[80] B. Gheflati and H. Rivaz, "Vision transformer for classification of breast ultrasound images," 2021, *arXiv:2110.14731*.

[81] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

[82] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, and W. Zhang, "Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2584–2594, Aug. 2020.

[83] S. Perera, S. Adhikari, and A. Yilmaz, "Pocformer: A lightweight transformer architecture for detection of COVID-19 using point of care ultrasound," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 195–199.

[84] J. Jiang and S. Lin, "COVID-19 detection in chest X-ray images using swin-transformer and transformer in transformer," 2021, *arXiv:2110.08427*.

[85] C. Liu and Q. Yin, "Automatic diagnosis of covid-19 using a tailored transformer-like network," in *Proc. J. Phys., Conf.*, 2021, vol. 2010, no. 1, Art. no. 012175.

[86] S. Park, G. Kim, J. Kim, B. Kim, and J. C. Ye, "Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis," in *Proc. Neural Inf. Process. Syst. Found. (NIPS)*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. CA, USA, 2021, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=Ggikq6Tdxch

[87] C.-C. Hsu, G.-L. Chen, and M.-H. Wu, "Visual transformer with statistical test for COVID-19 classification," 2021, *arXiv:2107.05334*.

[88] L. Zhang and Y. Wen, "A transformer-based framework for automatic COVID19 diagnosis in chest CTs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 513–518.

[89] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[90] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. Prathosh, "xViTCOS: Explainable vision transformer based COVID-19 screening using radiography," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–10, 2021.

[91] X. Gao, Y. Qian, and A. Gao, "COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models," 2021, *arXiv:2107.01682*.

[92] D. Shome, T. Kar, S. N. Mohanty, P. Tiwari, K. Muhammad, A. AlTameem, Y. Zhang, and A. K. J. Saudagar, "COVID-transformer: Interpretable COVID-19 detection using vision transformer for healthcare," *Int. J. Environ. Res. Public Health*, vol. 18, no. 21, p. 11086, 2021.

[93] A. A. E. Ambita, E. N. V. Boquio, and P. C. Naval, "COViT-GAN: Vision transformer forCOVID-19 detection in ct scan imageswith self-attention GAN fordataaugmentation," in *Proc. Int. Conf. Artif. Neural Netw.* New York, NY, USA: Springer, 2021, pp. 587–598.

[94] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J.-K. Lim, and J. C. Ye, "Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102299. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521003443

[95] M. Lu, Y. Pan, D. Nie, F. Liu, F. Shi, Y. Xia, and D. Shen, "SMILE: Sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images," in *Proc. MICCAI Workshop Comput. Pathol.*, 2021, pp. 159–169.

[96] H. Chen, C. Li, G. Wang, X. Li, M. Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun, S. Ai, and M. Grzegorzek, "GasHis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection," 2021, *arXiv:2104.14528*.

[97] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model," *Comput. Intell. Neurosci.*, vol. 2021, Aug. 2021, Art. no. 7529893.

[98] Y. Zheng, R. Gindra, M. Betke, J. Beane, and V. B. Kolachalama, "A deep learning based graph-transformer for whole slide image classification," *medRxiv*, to be published. [Online]. Available: https://ieeexplore.ieee.org/document/9779215

[99] J. D. Bodapati, N. S. Shaik, and V. Naralasetti, "Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification," *J. Ambient Intell. Humanized Comput.*, vol. 12, pp. 9825–9839, Jan. 2021.

[100] S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, and Y. Zheng, "MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Wiesbaden, Germany: Springer, 2021, pp. 45–54.

[101] R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, and Y. Zhang, "Lesion-aware transformers for diabetic retinopathy grading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10938–10947.

[102] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision Transformer-based recognition of diabetic retinopathy grade," *Med. Phys.*, vol. 48, no. 12, pp. 7850–7863, 2021.

[103] N. AlDahoul, H. Abdul Karim, M. Joshua Toledo Tan, M. A. Momo, and J. Ledesma Fermin, "Encoding retina image to words using ensemble of vision transformers for diabetic retinopathy grading," *F1000Research*, vol. 10, p. 948, Sep. 2021.

[104] H. Yang, J. Chen, and M. Xu, "Fundus disease image classification based on improved transformer," in *Proc. Int. Conf. Neuromorphic Comput. (ICNC)*, Oct. 2021, pp. 207–214.

[105] M. Al-Shabi, K. Shak, and M. Tan, "ProCAN: Progressive growing channel attentive non-local network for lung nodule classification," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108309. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321004891

[106] J. Moranguinho, T. Pereira, B. Ramos, J. Morgado, J. L. Costa, and H. P. Oliveira, "Attention based deep multiple instance learning approach for lung cancer prediction using histopathological images," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2852–2855.

[107] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, Jan. 2021.

[108] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.

[109] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*, M. Reyes, P. Henriques Abreu, J. Cardoso, M. Hajij, G. Zamzmi, P. Rahul, and L. Thakur, Eds. Cham, Switzerland: Springer, 2021, pp. 13–23.

[110] Y. Barhoumi and R. Ghulam, "Scopeformer: N-CNN-ViT hybrid model for intracranial hemorrhage classification," 2021, *arXiv:2107.04575*.

[111] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1251–1258.

[112] S. Liang and Y. Gu, "Computer-aided diagnosis of Alzheimer's disease through weak supervision deep learning framework with attention mechanism," *Sensors*, vol. 21, no. 1, p. 220, 2021.

[113] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[114] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2018.

[115] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 109–119.

[116] Q. Jia and H. Shu, "BiTr-unet: A CNN-transformer combined network for MRI brain tumor segmentation," 2021, *arXiv:2109.12271*.

[117] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3D tumor segmentation," 2021, *arXiv:2111.13300*.

[118] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," 2022, *arXiv:2201.01266*.

[119] Y. Li, W. Cai, Y. Gao, and X. Hu, "More than encoder: Introducing transformer decoder to upsample," 2021, *arXiv:2106.10637*.

[120] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. Wai Kin Kong, "TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation," 2021, *arXiv:2107.05274*.

[121] G. Xu, X. Wu, X. Zhang, and X. He, "LeViT-UNet: Make faster encoders with transformer for medical image segmentation," 2021, *arXiv:2107.08623*.

[122] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102327.

[123] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[124] A. Vakanski, M. Xian, and P. E. Freer, "Attention-enriched deep learning model for breast tumor segmentation in ultrasound images," *Ultrasound Med. Biol.*, vol. 46, no. 10, pp. 2819–2833, 2020.

[125] X. Zhu, H. Hu, H. Wang, J. Yao, W. Li, D. Ou, and D. Xu, "Region aware transformer for automatic breast ultrasound tumor segmentation," in *Medical Imaging With Deep Learning*. Proceedings of Machine Learning Research, 2022. [Online]. Available: https://openreview.net/forum?id=2bVDHzy7xwV

[126] Y. Liu, Y. Yang, W. Jiang, T. Wang, and B. Lei, "3D deep attentive U-Net with transformer for breast tumor segmentation from automated breast volume scanner," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 4011–4014.

[127] T. Prangemeier, C. Reich, and H. Koeppl, "Attention-based transformers for instance segmentation of cells in microstructures," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 700–707.

[128] Y. Zhang, R. Higashita, H. Fu, Y. Xu, Y. Zhang, H. Liu, J. Zhang, and J. Liu, "A multi-branch hybrid transformer network for corneal endothelial cell segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Wiesbaden, Germany: Springer, 2021, pp. 99–108.

[129] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and Y. Zhang, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds. Red Hook, NY, USA: Curran Associates, 2021, pp. 2136–2147. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf

[130] C. Nguyen, Z. Asad, R. Deng, and Y. Huo, "Evaluating transformer-based semantic segmentation networks for pathological image segmentation," in *Medical Imaging 2022: Image Processing*, vol. 12032, O. Colliot and I. Išgum, Eds. Bellingham, WA, USA: SPIE, 2022, pp. 942–947, doi: 10.1117/12.2611177.

[131] B. Yun, Y. Wang, J. Chen, H. Wang, W. Shen, and Q. Li, "SpecTr: Spectral transformer for hyperspectral pathology image segmentation," 2021, *arXiv:2103.03604*.

[132] H. Li, J. Zhang, and B. Menze, "Generalisable cardiac structure segmentation via attentional and stacked image adaptation," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*. Cham, Switzerland: Springer, 2020, pp. 297–304.

[133] F. Kong and S. C. Shadden, "A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention U-Net," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*. Cham, Switzerland: Springer, 2020, pp. 287–296.

[134] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[135] V. M. Campello, P. Gkontra, and C. Izquierdo, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3543–3554, Dec. 2021.

[136] K. Deng, Y. Meng, D. Gao, J. Bridge, Y. Shen, G. Lip, Y. Zhao, and Y. Zheng, "TransBridge: A lightweight transformer for left ventricle segmentation in echocardiography," in *Proc. Int. Workshop Adv. Simplifying Med. Ultrasound*. Cham, Switzerland: Springer, 2021, pp. 63–72.

[137] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[138] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer," 2021, *arXiv:2109.07162*.

[139] C. Yao, J. Tang, M. Hu, Y. Wu, W. Guo, Q. Li, and X.-P. Zhang, "Claw U-Net: A UNet variant network with deep feature concatenation for scleral blood vessel segmentation," in *Proc. 1st CAAI Int. Conf. Artif. Intell. (CICAI)* Berlin, Germany: Springer-Verlag, 2021, pp. 67–78, doi: 10.1007/978-3-030-93049-3_6.

[140] Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, "TransClaw U-Net: Claw U-Net with transformers for medical image segmentation," 2021, *arXiv:2107.05188*.

[141] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 171–180.

[142] Z. Shen, H. Yang, Z. Zhang, and S. Zheng, "Automated kidney tumor segmentation with convolution and transformer network," in *Kidney and Kidney Tumor Segmentation*, N. Heller, F. Isensee, D. Trofimova, R. Tejpaul, N. Papanikolopoulos, and C. Weight, Eds. Cham, Switzerland: Springer, 2022, pp. 1–12.

[143] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 14–24.

[144] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3D medical image analysis," 2021, *arXiv:2111.14791*.

[145] D. Karimi, S. D. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 78–88.

[146] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[147] D. Viet Sang, T. Quang Chung, P. Ngoc Lan, D. Viet Hang, D. Van Long, and N. Thi Thuy, "AG-CUResNeSt: A novel method for colon polyp segmentation," 2021, *arXiv:2105.00402*.

[148] Z. Tang, X. Peng, S. Geng, Y. Zhu, and D. N. Metaxas, "CU-Net: Coupled U-Nets," in *Proc. Brit. Mach. Vis. Conf.* Newcastle, U.K.: BMVA Press, Aug. 2018, p. 305. [Online]. Available: http://bmvc2018.org/contents/papers/0338.pdf

[149] Z. Zhang and W. Zhang, "Pyramid medical transformer for medical image segmentation," 2021, *arXiv:2104.14702*.

[150] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," 2021, *arXiv:2106.06716*.

[151] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion transformers," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Z.-H. Zhou, Ed. 2021, pp. 807–815, doi: 10.24963/IJCAI.2021/112.

[152] R. Ranjbarzadeh, A. B. Kasgari, S. J. Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–17, 2021.

[153] A. Ribeiro, "Study of attention mechanisms and ensemble methods for medical image semantic segmentation," Univ. Minho, Braga, Portugal, Tech. Rep., 2019.

[154] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-Net transformer: Self and cross attention for medical image segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2021, pp. 267–276.

[155] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 206–216.

[156] Y. Li, S. Wang, J. Wang, G. Zeng, W. Liu, Q. Zhang, Q. Jin, and Y. Wang, "GT U-Net: A U-Net like group transformer network for tooth root segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2021, pp. 386–395.

[157] Y. Li, G. Zeng, Y. Zhang, J. Wang, and Q. Jin, "AGMB-transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 4, pp. 1684–1695, Apr. 2022.

[158] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "SA-UNet: Spatial attention U-Net for retinal vessel segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1236–1242.

[159] J. Wang, Y. Zhao, L. Qian, X. Yu, and Y. Gao, "EAR-NET: Error attention refining network for retinal vessel segmentation," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2021, pp. 1–7.

[160] C. Guo, M. Szemenyei, Y. Yi, W. Zhou, and H. Bian, "Residual spatial attention network for retinal vessel segmentation," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2020, pp. 509–519.

[161] X.-F. Du, J.-S. Wang, and W.-Z. Sun, "UNet retinal blood vessel segmentation algorithm based on improved pyramid pooling method and attention mechanism," *Phys. Med. Biol.*, vol. 66, no. 17, 2021, Art. no. 175013.

[162] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[163] J. Gao, Q. Huang, Z. Gao, and S. Chen, "Image segmentation of retinal blood vessels based on dual-attention multiscale feature fusion," *Comput. Math. Methods Med.*, vol. 2022, Jul. 2022, Art. no. 8111883.

[164] A. Amer, T. Lambrou, and X. Ye, "MDA-Unet: A multi-scale dilated attention U-Net for medical image segmentation," *Appl. Sci.*, vol. 12, no. 7, p. 3676, 2022.

[165] P. Zhao, J. Zhang, W. Fang, and S. Deng, "SCAU-Net: Spatial-channel attention U-Net for gland segmentation," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 670, Jul. 2020.

[166] Q. Jin, Z. Meng, C. Sun, H. Cui, and R. Su, "RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 1471, Jan. 2020.

[167] I. Sobirov, O. Nazarov, H. Alasmawi, and M. Yaqub, "Automatic segmentation of head and neck tumor: How powerful transformers are?" in *Medical Image With Deep Learning*. Proceedings of Machine Learning Research, 2022. [Online]. Available: https://openreview.net/forum?id=reIO5WfgbLd

[168] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "After-UNet: Axial fusion transformer UNet for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 3971–3981.

[169] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2577–2586. [Online]. Available: https://aclanthology.org/P18-1240

[170] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 673–680.

[171] J. Lovelace and B. Mortazavi, "Learning to generate clinically coherent chest X-ray reports," in *Proc. Conf. Empirical Methods Natural Lang. Process., Findings*, 2020, pp. 1235–1243.

[172] P. Srinivasan, D. Thapar, A. Bhavsar, and A. Nigam, "Hierarchical X-ray report generation via pathology tags and multi head attention," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 600–616.

[173] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.* Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2021, pp. 5288–5304. [Online]. Available: https://aclanthology.org/2021.naacl-main.416

[174] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 1439–1449. [Online]. Available: https://aclanthology.org/2020.emnlp-main.112

[175] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive attention for automatic chest X-ray report generation," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2021, pp. 269–280. [Online]. Available: https://aclanthology.org/2021.findings-acl.23

[176] I. Najdenkoska, X. Zhen, M. Worring, and L. Shao, "Variational topic inference for chest X-ray report generation," in *Medical Image Computing and Computer Assisted Intervention–(MICCAI)*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham, Switzerland: Springer, 2021, pp. 625–635.

[177] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Inform. Med. Unlocked*, vol. 24, Jan. 2021, Art. no. 100557.

[178] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[179] A. Radford, J. Wu, R. Child, D. Luan, and D. Amodei, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[180] A. B. Amjoud and M. Amrouch, "Automatic generation of chest X-ray reports using a transformer-based deep learning model," in *Proc. 5th Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2021, pp. 1–5.

[181] S. Tipirneni and C. K. Reddy, "Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series," 2021, *arXiv:2107.14293*.

[182] Y. Wang, Z. Lin, Z. Xu, H. Dong, J. Tian, J. Luo, Z. Shi, Y. Zhang, J. Fan, and Z. He, "Trust it or not: Confidence-guided automatic radiology report generation," 2021, *arXiv:2106.10887*.

[183] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13753–13762.

[184] F. Liu, C. You, X. Wu, S. Ge, S. Wang, and X. Sun, "Auto-encoding knowledge graph for unsupervised medical report generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

[185] H. Nguyen, D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, and L. Cheng, "Automated generation of accurate & fluent medical X-ray reports," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3552–3569. [Online]. Available: https://aclanthology.org/2021.emnlp-main.288

[186] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 72–82.

[187] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "RATCHET: Medical transformer for chest X-ray diagnosis and reporting," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 293–303.

[188] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nature Mach. Intell.*, vol. 4, pp. 32–40, Jan. 2022.

[189] J. Zhang, Y. Nie, J. Chang, and J. J. Zhang, "Surgical instruction generation with transformers," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2021, pp. 290–299.

[190] X. Wang, Z. Xu, L. Tam, D. Yang, and D. Xu, "Self-supervised image-text pre-training with mixed data in chest X-rays," 2021, *arXiv:2103.16022*.

[191] Z. Shen, R. Fu, C. Lin, and S. Zheng, "COTR: Convolution in transformer network for end to end polyp detection," in *Proc. 7th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2021, pp. 1757–1761.

[192] S. Liu, H. Zhou, X. Shi, and J. Pan, "Transformer for polyp detection," 2021, *arXiv:2111.07918*.

[193] T. S. Mathai, S. Lee, D. C. Elton, T. C. Shen, Y. Peng, Z. Lu, and R. M. Summers, "Lymph node detection in T2 MRI with transformers," in *Medical Imaging 2022: Computer-Aided Diagnosis*, vol. 12033, K. Drukker, K. M. Iftekharuddin, H. Lu, M. A. Mazurowski, C. Muramatsu, and R. K. Samala, Eds. Bellingham, WA, USA: SPIE, 2022, pp. 855–859, doi: 10.1117/12.2613273.

[194] K. Gong, C. Catana, J. Qi, and Q. Li, "Direct reconstruction of linear parametric images from dynamic pet using nonlocal deep image prior," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 680–689, Mar. 2022.

[195] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.

[196] P. Grundmann, S. Arnold, and A. Löser, "Self-supervised answer retrieval on clinical notes," 2021, *arXiv:2108.00775*.

[197] B.-H. Kim and V. Ganapathi, "Read, attend, and code: Pushing the limits of medical codes prediction from clinical notes by machines," in *Proc. 6th Mach. Learn. Healthcare Conf.* (Proceedings of Machine Learning Research), vol. 149, K. Jung, S. Yeung, M. Sendak, M. Sjoding, and R. Ranganath, Eds. Proceedings of Machine Learning Research, 2021, pp. 196–208. [Online]. Available: https://proceedings.mlr.press/v149/kim21a.html

[198] Y. Li, G. Zeng, Y. Zhang, J. Wang, Q. Zhang, Q. Jin, L. Sun, Q. Lian, N. Xia, R. Peng, K. Tang, Y. Wang, and S. Wang, "AGMB-transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy," 2021, *arXiv:2105.00381*.

[199] X. Ma, G. Luo, W. Wang, and K. Wang, "Transformer network for significant stenosis detection in CCTA of coronary arteries," in *Medical Image Computing and Computer Assisted Intervention–(MICCAI)*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham, Switzerland: Springer, 2021, pp. 516–525.

[200] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 1, pp. 312–340, 2019.

[201] R. Shrestha, E. Fajardo, N. Gil, K. Fidelis, A. Kryshtafovych, B. Monastyrskyy, and A. Fiser, "Assessing the accuracy of contact predictions in CASP13," *Proteins: Struct., Function, Bioinf.*, vol. 87, no. 12, pp. 1058–1068, 2019.

[202] W. Zheng, Y. Li, C. Zhang, R. Pearce, S. Mortuza, and Y. Zhang, "Deep-learning contact-map guided protein structure prediction in CASP13," *Proteins: Struct., Function, Bioinf.*, vol. 87, no. 12, pp. 1149–1164, 2019.

[203] J. Hou, T. Wu, R. Cao, and J. Cheng, "Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13," *Proteins: Struct., Function, Bioinf.*, vol. 87, no. 12, pp. 1165–1178, 2019.

[204] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, and C. Qin, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.

[205] M. Filipavicius, M. Manica, J. Cadow, and M. R. Martinez, "Pre-training protein language models with label-agnostic binding pairs enhances performance in downstream tasks," in *Proc. Mach. Learn. Struct. Biol. Workshop (NeurIPS)*, 2020, pp. 1–20.

[206] C. Chen, T. Wu, Z. Guo, and J. Cheng, "Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction," *Proteins: Struct., Function, Bioinf.*, vol. 89, no. 6, pp. 697–707, 2021.

[207] J. Wang, X. Liu, S. Shen, L. Deng, and H. Liu, "DeepDDS: Deep graph neural network with attention mechanism to predict synergistic drug combinations," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab390. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab390/6375262, doi: 10.1093/bib/bbab390.

[208] A. Khan and B. Lee, "Gene transformer: Transformers for the gene expression-based classification of lung cancer subtypes," 2021, *arXiv:2108.11833*.

[209] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, and L. Caffery, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, no. 1, pp. 1–8, 2021.

[210] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, pp. 1–8, 2019.

[211] A. Paszke, S. Gross, F. Massa, and A. Lerer, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[212] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[213] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[214] 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, 2015, Art. no. e0130140.

[215] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for PyTorch," 2020, *arXiv:2009.07896*.

[216] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 782–791.

[217] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[218] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May/Jun. 2007.

[219] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–18. [Online]. Available: https://openreview.net/forum?id=HJlnC1rKPB

[220] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[221] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[222] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (Un) reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Wiesbaden, Germany: Springer, 2019, pp. 267–280.

[223] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[224] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[225] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–14.

TIAGO GONÇALVES (Member, IEEE) received the M.S. degree in bioengineering (biomedical engineering) from the Faculty of Engineering of the University of Porto (FEUP), in 2019, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. He is also a Research Assistant at the Centre for Telecommunications and Multimedia, Visual Computing and Machine Intelligence (VCMI) Research Group, INESC TEC. His research interests include machine learning, explainable artificial intelligence (in-model approaches), computer vision, medical decision support systems, and machine learning deployment.

ISABEL RIO-TORTO is currently pursuing the Ph.D. degree in computer science with the Faculty of Sciences of the University of Porto (FCUP). She is also a Research Assistant at the Centre for Telecommunications and Multimedia, VCMI Research Group, INESC TEC. She is an Invited Teaching Assistant at the Faculty of Engineering of the University of Porto (FEUP). Her research interests include explainable artificial intelligence, computer vision, natural language processing, and the connection between vision and language. She previously studied electrical and computers engineering at FEUP, where she developed her master's thesis, which won her the First place at 2020 Fraunhofer Portugal Challenge (M.Sc. category) and the Best 2020 Master Thesis Award by the Associação Portuguesa de Reconhecimento de Padrões (APRP).

LUÍS F. TEIXEIRA (Member, IEEE) received the Licenciatura (five-year) degree in electrical and computer engineering, the M.S. degree in computer networks and services, and the Ph.D. degree in electrical and computer engineering from the Universidade do Porto, in 2001, 2004, and 2009, respectively. He is currently an Assistant Professor at the Department of Informatics Engineering, Faculdade de Engenharia da Universidade do Porto (FEUP), and a Researcher at the Center for Telecommunications and Multimedia, INESC TEC. Previously, he was a Researcher at INESC Porto, from 2001 to 2008, a Visiting Researcher at the University of Victoria, in 2006, a Senior Scientist at Fraunhofer AICOS, from 2008 to 2013, and the President of the Associação Portuguesa de Reconhecimento de Padrões (APRP), from 2019 to 2021. He has also been teaching courses in computer vision, machine learning, and other areas for graduate studies. He (co-) supervised or is (co-) supervising more than 90 M.Sc. students and six Ph.D. students. He has participated in 12 concluded or ongoing projects (EU-funded, national, and contracts) as a Researcher and a WP Leader. His main research interests include computer vision, machine learning, and human–centered computing.

JAIME S. CARDOSO (Senior Member, IEEE) was the President of the Portuguese Association for Pattern Recognition (APRP), affiliated to the IAPR, from 2012 to 2015. He is currently a Full Professor at the Faculty of Engineering of the University of Porto (FEUP). Image and video processing focuses on medicine and biometrics. The work on machine learning cares mostly with the adaptation of learning to the challenging conditions presented by visual data, with a focus on deep learning and explainable machine learning. The particular emphasis of the work in decision support systems goes to medical applications, always anchored on the automatic analysis of visual data. He has coauthored more than 300 papers, more than 100 of which in international journals, which attracted more than 6900 citations, according to Google Scholar. His research interests include computer vision, machine learning, and decision support systems.