Politecnico di Torino

1859

ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Electrical, Electronics and Communications Engineering

(35thcycle)

Biological Image Analysis Through Deep Learning Techniques

By

Sergio Cannata

******

Supervisor(s):

Prof. Eros Pasero, Supervisor
Prof. Giansalvo Cirrincione, Co-Supervisor
Prof. Guido Lombardi, Co-Supervisor

Doctoral Examination Committee:

Prof. Francesco Carlo Morabito, Referee, Università Mediterranea di Reggio Calabria
Prof. Luca Grilli, Referee, Università degli Studi di Foggia
Prof. E.F, University of...
Prof. G.H, University of...
Prof. I.J, University of...

Politecnico di Torino
2024

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Sergio Cannata
2024

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

In loving memory of Alice

Acknowledgements

I would like to acknowledge my supervisors, Prof. Eros Pasero and Prof. Giansalvo Cirrincione, who supported and assisted my research with wisdom, knowledge and expertise. My gratitude goes as well to all the professors and colleagues whom I've had the pleasure to share this experience with: your support was fundamental and I probably wouldn't have made it this far without you all. Finally, my last words of acknowledgement are for my family and my friends, who helped me get through thick and thin.

Abstract

The impact of the latest advancements in Artificial Intelligence on everyday life is growing stronger day after day at a frantic pace. Deep Learning systems are increasingly becoming capable of executing complex tasks autonomously, gaining credibility in almost any context and/or application. In this scenario, the use of AI software to assist experts in making fast and reliable assessments, especially in clinical applications, is leading to a revolution in image analysis.

The purpose of the research presented in this document is to show how Deep Learning models and algorithms are reshaping the way biological images are treated, inspected and interpreted, heading towards the next level of AI-enhanced medicine. Specifically, it is shown that leveraging Vision Transformer-based architectures to detect and identify peculiar diseases over different types of clinical images and applications represents a powerful and effective technique to tackle different types of image classification problems, both on large and small datasets. Moreover, it is demonstrated that leveraging the most recent Diffusion-based image generation models can effectively boost performance whenever data lacks quality and/or uniformity, when the images in the database are imbalanced among the different classes, or when data samples fail to represent the most significant category.

Given the inherent peculiarities of any specific applications, researchers commonly tackled each problem by customizing the Transformer architecture and adapting it to properly process the data they dealt with. However, this approach shows the lack of standardization purposes. In this sense, the present document proves the validity of leveraging ViT-based models for a variety of classification problems on biological images, suggesting that they can used almost *out of the box* for a plethora of image detection/classification applications, which can easily be extended beyond the clinical field. To defend this statement, the procedures behind image analysis are observed and compared for Vision Transformers and Convolutional Neural Networks, showing that a better understanding of how attention works in image classification can head towards an increased awareness of what makes features *relevant*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and State-of-the-Art

1.1 Motivation

The importance of image analysis in biological and clinical settings is a remarkably relevant topic, given the amount and the significance of information that a picture can carry. Indeed, a fast, reliable and correct identification procedure can undoubtedly enhance experts decision, leading to a timely intervention that can make a difference in critical cases.

In this scenario, the most common Deep Learning based tools for image classification and analysis, efficient as they can be in detecting local features, might fail to gain awareness about the overall picture context. This has been identified in literature as a typical downside of utilizing Convolutional Neural Networks, a deep neural model category that has dominated the scene until the last few years - and is still considered the go-to choice for image detection on most applications.

On the basis of the above considerations, this section will describe how attention-based deep neural models - namely, the Vision Transformer architecture - can successfully be applied to a variety of clinical use cases, solving disease identification problems with different levels of complexity.

1.2 State-of-the-Art

Soon after its development, Vision Transformer has been deployed for clinical image analysis on the most common tasks, such as classification, segmentation, reconstruction and registration. Research has been mainly focused on medical image categories like X-ray images, Computed Tomography (CT) scans, Magnetic Resonance Images (MRI), Histopathological and Fundus images. Results have shown that ViT-based architectures are able to compete with, and often outperform, Convolutional Neural Networks over the listed application categories. In this section, recent examples of Vision Transformer applications for medical image classification will be reported, in order to show an overview of the current state-of-the-art.

Gai et al. [1] showed that Vision Transformers are capable of outperforming their convolutional counterparts in 3D medical image classification by leveraging self-supervised learning on large datasets. Almalik et al. [2] developed a custom version of the ViT architecture, demonstrating that their model exhibits improved robustness against adversarial attacks when classifying X-ray and Fundus images. Park et al. [3] presented a federated, task-agnostic ViT for COVID-19 diagnosis on X-ray images, proving the suitability of the architecture for collaborative learning in medical imaging. Tummala et al. [4] effectively leveraged an ensemble of basic ViT architectures to classify brain tumor on MRIs. The first transformer for multi-modal image classification on ear MRIs was proposed by Dai et al. [5], efficiently combining CNNs and ViTs. Xia et al. [6] successfully integrated a UNet and a Vision Transformer to create an original architecture for pancreatic cancer detection on CT scans. In their paper, Jang et al. [7] proposed an architecture that incorporates 2D and 3D CNNs together with a transformer encoder to perform Alzheimer detection on MRIs.

Ikromjanov et al. [8] leveraged a ViT-based model to identify prostate cancer after extracting Regions-of-Interest (ROIs) on histopathological images. Sun et al. [9] proposed an architecture in which self-attention blocks are fed with the outputs of 1x1 convolutional layers for feature extraction, with the purpose of classifying and grading diabetic retinopathy on Fundus images. Designing custom modules to extract both global and local information, Chen et al. [10] developed a Vision Transformer model for the identification of normal and abnormal gastric histopathological images. Zheng et al. [11] developed a combination of Graph and Vision Transformer to detect adenocarcinoma and squamous cell carcinoma on histopathological lung images.

To analyze 3D knee MRIs, Wang et al. [12] proposed an integration of 3D convolutional layers and Transformers with successfull results in cartilage defect detection. Chen and Krishnan [13] adopted a combination of self-supervised learning and transfer learning to extract morphological features in skin cell tissue classification on histopathological images.

Zhao et al. [14] efficiently merged three parallel convolutional networks with a Transformer to propose a quantitave measurement of hepatocarcinoma on Magnetic Resonance Images. Zhao, C., et al. [15] exploited transfer learning on an attention-based model to effectively identify cervical cancer classification on imbalanced datasets of histopathological cell images.

The validity and versatility of Transformer-based approaches on a wide plethora of medical image classification applications clearly emerge from the depicted state-of-the-art landscape. To remark the impact of Transformers on the medical computer vision domain, it is worth underlining that the above described scenario was originated and shaped by researchers over the notably short span of about three years. This observation portrays the great ferment in this research area.

Indeed, another relevant aspect plainly comes to fore: the most common approach to solving classification tasks is to modify the basic Transformer architecture in order to adapt it to a specific image identification problem. As reasonable as it is, such methodology inevitably orients research paths in a direction that drifts away from any standardization purposes. Undisputedly, every problem - and every dataset - is different and demands to be treated its own way. Moreover, it is evident that no model can be regarded to as a general, universal classifier that can serve all applications. Nonetheless, the work presented in this document clearly demonstrates that the Vision Transformer in its basic configurations can proficiently handle several clinical image identification tasks which span over different areas of application, often outperforming the state-of-the-art on the investigated datasets.

Chapter 2

Vision Transformer

2.1 Model Architecture

Developed by Dosovitskiy et al. [16] to bring the original Transformer [17] to the computer vision domain, the Vision Transformer was designed to overcome the inherent limitations of Convolutional Neural Networks (CNNs). Indeed, CNNs lack the capability of capturing the spatial relationships among areas of the image that are located far from each other. To tackle this issue, Vision Transformers avoid relying on convolutions to analyze images, leveraging the attention mechanism to gather awareness about the global picture context.

The unbearable computational cost associated with the application of attention on all image pixels would have limited the architecture to small-sized pictures only. For this reason, ViT starts its image processing procedure by resizing the picture to a specific size, which is usually 224x224 or 384x384, depending on the model configuration. Subsequently, the image is divided into equally-sized, non-overlapping patches (typically 16x16 or 32x32). This allows to process pictures of almost any size.

Patches are then embedded through learnable linear projections, and a further learnable positional encoding is added. An additional token, called *CLS* (analogous to the classification token on NLP Transformer applications) is prepended to the resulting vector, which is then fed to the next layers. The CLS token is required for image classification tasks, and can be interpreted as a sort of summarization of all the relevant image features. This procedure embeds image patches into vectors that contain information about the relative patch positions, thanks to positional encoding. The above vectors are then concatenated into a unique sequence, meaning that the next layers of the Vision Transformer will process the entire image at once.

The unique vector is layer normalized [18] and processed by an attention layer, as described in the next section. The input before normalization is added to the attention layer output by means of a skip connection. The resulting output is layer-normalized once more, and fed to a Multi-Layer Perceptron (MLP) with two fully-connected layers equipped with a GELU activation function [19]. Output is again added to the input located before layer normalization with another skip connection. Such sequence constitutes a Vision Transformer encoder block, and is repeated for a variable number of times, depending on the network configuration.

Eventually, the output is fed to another Multi-Layer Perceptron that will produce as many outputs as the required number of categories for the given classification task.

2.1.1 The Attention Mechanism

In analogy with an archive search, attention is implemented by projecting each input sequence to three specific vectors, namely *key*, *query* and *value* - with associated dimensions d_k, d_q and d_v. The *attention score* of each input is evaluated as:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

where Q, K and V are the query, key and value matrices, respectively. Each network element computing attention is named *head*. In the Vision Transformer, similarly to what happens in standard transformer-based architectures, many heads evaluate attention in parallel, thus establishing *multi-head attention*, which is defined as:

$$MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O \tag{2.2}$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$, h is the number of attention heads, and $d_k =$

$d_v = d_{model}/h$. d_{model} is the dimensionality of the representations used as input to the multi-head attention, which is the same as the dimensionality of the output.

2.2 Model Applications on Clinical Datasets

The next chapters will describe the application of the techniques so far discussed and analyzed to three different case studies:

- COVID-19 detection on chest X-ray images;
- skin lesion detection on dermatoscopic images;
- mammographic image analysis for breast cancer detection.

Such example applications prove the remarkable effectiveness of deep learning methods in image analysis and classification, demonstrating their capabilities to outperform the state-of-the-art. In addition, it will be shown that the above techniques are able to display information connected to the decision process that pushes systems into assigning a sample into a given category, thus paving the way for model explainability.

Chapter 3

COVID-19 Detection on Chest X-Ray Images

Part of the work presented in this chapter was previously published on *IEEE Access* [20].

3.1 Context and Motivation

On December 31st, 2019, Chinese health authorities reported an outbreak of pneumonia cases of unknown aetiology in the city of Wuhan (Hubei Province, China). Shortly thereafter, on January 9th, 2020, the China CDC (the Center for Disease Control and Prevention of China) identified a new coronavirus (tentatively named 2019-nCoV) as the aetiological cause of these diseases. Chinese health authorities have also confirmed the inter-human transmission of the virus. On 11th February, the World Health Organization (WHO) announced that the disease transmitted by 2019-nCoV was renamed COVID-19 (Corona Virus Disease). The Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses has officially classified with the name of SARS-CoV-2 the virus provisionally named by the international health authorities 2019-nCoV and responsible of cases of COVID-19 (Corona Virus Disease). The CSG - responsible for defining the official classification of viruses and the taxonomy of the Corona viridae family, after evaluating the novelty of the human pathogen and on the basis of phylogeny, taxonomy and established practice, has formally associated this virus with the coronavirus causing severe acute respiratory syndrome (SARS-CoVs, Severe acute respiratory syndrome coronaviruses) classifying it as Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) [21]. After assessing the severity levels and global spread of the SARS-CoV-2 infection, WHO declared that the COVID-19 epidemic can be in fact considered a pandemic.

After the World Health Organization (WHO) declared the rapid spread of the aggressive COVID-19 virus, the world of scientific research went to great lengths to propose a solution for the early diagnosis of the virus [22]. Indeed, the rapid detection of COVID-19 can help control the spread of the disease.

Nowadays, the most used and most reliable method of diagnosing infection is the Reverse Transcription-Polymerase Chain Reaction (RTPCR). A sample is taken by nose/mouth and pharyngeal swab and analysed by real-time molecular methods through the amplification of the viral genes mostly expressed during the infection. This analysis can only be carried out in highly specialized laboratories, identified by the health authorities and requires on average from 2 to 6 hours to return a result. Another category of tests that have a lower sensitivity and specificity than the previous molecular tests, are the antigen swabbing. This type of test is based on the search for viral proteins (antigens) in respiratory samples. The sampling methods are the same as for molecular tests (nasal and throat swab) but the response time is shorter (about 15 minutes). Finally, the serological tests highlight the presence of antibodies against the virus and reveal whether there has been exposure to the virus; yet, only in a few cases can they detect that an infection is in progress. In the current state of scientific development, serological tests cannot replace molecular tests based on the identification of viral RNA [23]. In recent times, the attention for the diagnosis of infection has focused on imaging tests. Chest X-ray (CXR) and computed tomography (CT) are the most popular imaging techniques for diagnosing COVID-19 disease. The historical conception of diagnostic imaging systems has been fully explored through several approaches ranging from automation engineering to deep learning [24]. Although some studies [25] show an increase in sensitivity when analyzing CT scans as opposed to CXR, this study focuses on chest X-ray images due to their readiness and wide availability, which is not always the case for CT images [26].

Leveraging Convolutional Neural Networks (CNNs) is one of the most popular and effective approaches in the diagnosis of COVID-19 from digitized images.

Several reviews have been carried out to highlight recent contributions to COVID-19 detection [27]. Pre-trained CNN models were used for feature extraction using SVM classifiers with various kernel functions [28]. Then, several pre-trained CNN models were further trained using chest X-ray images for COVID-19 detection. The accuracy of the classification was used to evaluate the performance of the proposed methods. The pre-trained deep CNN models used in the study were ResNet18, ResNet50, ResNet101, VGG16, and VGG19. Since testing the study, the deep characteristics model (ResNet50) and SVM with linear kernel function produced an accuracy score of 94.7%, which was the highest of all results. Test results for fine-tuning the ResNet50 model and end-to-end training of the developed CNN model were 92.6% and 91.6% respectively. Since the number of COVID-19 X-Ray samples is limited, transfer learning (TL) appears as the reference method for classifying disease data to develop accurate automated diagnosis models. In this context, networks are able to acquire knowledge from pre-trained networks on large-scale image datasets or alternative data-rich sources. The classification algorithm based on transfer learning acquired results with an accuracy of 97.66% and an F1-score of 97.61% [29].

The studies suggested that transfer learning can allow the network to extract significant features associated to the COVID-19 disease diagnosis. In fact, several works have applied this idea in order to rapidly develop a reliable tool to assist medical experts in diagnosing COVID-19. The wide popularity of convolutional neural networks made them the first choice for a number of works, in which said architectures manage to identify COVID-struck lungs on X-ray images. Shazia et al. [30] compared performances of several CNNs, presenting a test accuracy of 99.48% obtained by DenseNet121. However, the classification task only dealt with the goal of distinguishing between COVID and viral pneumonia, with the first (and most relevant) class being represented in the test set with just 157 images, versus more than 4000 images for pneumonia. Many other studies tackle the problem of classifying COVID and non-COVID X-ray images, typically viral/bacterial pneumonia, or they add normal lung CXR as a third category [31] [32].

The advent of the Vision Transformer has led many researchers to perform the same kind of task with such recent neural model, and assess its performance against CNNs. ViT's capability to connect local patches of information on a single image and build up the picture context has led the Vision Transformer to often surpass its convolutional counterparts. Thus, many other works have deployed the Vision Transformer for COVID detection. Krishnan applied the ViT to distinguish between

COVID and non-COVID chest X-ray images, assessing performance against some CNNs, reaching a final accuracy of 97.6% [33]. D. Shome et al. applied the model for both three and two classes, including pneumonia as a third category. The study managed to achieve accuracies of 92% and 98%, respectively [34]. Mondal et al. obtained a 96% test accuracy when using the Vision Transformer for classifying chest X-ray images into the same three categories, namely COVID, pneumonia and normal, healthy lungs [35]. Park et al. added a convolutional backbone for feature extraction, and developed a system to assess COVID severity. However, their work classified X-ray images over three categories – namely COVID, normal and a generic class described as "other infection". Authors present test accuracy separately for the three classes with a confidence interval of 95%, and its value never goes above 94.2% (which is the best result for the normal class) [36]. All of the aforementioned papers present studies that are mainly focused on distinguishing between COVID and non-COVID patients, or they include viral pneumonia as a third class when tackling multi-class classification. Almaki et al. took into consideration both viral and bacterial pneumonia, thus working over four classes, and combined their custom CNN with a few selected machine learning algorithms. Yet, their method only reached a final test accuracy of 97.29% [37].

The work presented in this document tackles classification over four categories – including lung opacity as a fourth class – and proves that the Vision Transformer is capable of distinguishing an additional class of pulmonary diseases with considerably high levels of accuracy and specificity. To the best of the author's knowledge, no other similar methods were able to reach such level of accuracy over four different classes of chest X-ray images in the case of automatic COVID-19 detection.

3.2 Dataset

The dataset used in this work is a collection of chest X-ray images that were (and still are) gathered by researchers from different countries, with the specific purpose of creating a publicly available database for COVID-related research. The version used for this work was collected in October 2021, and it consists of 3616 COVID-19 positive cases, 10,192 normal healthy lung images, 6012 pulmonary opacity images (non-COVID lung infection), and 1345 viral pneumonia images [38][39]. All images were downloaded as png-formatted RGB images with a size of 299x299x3.

The lungs are the two organs responsible for supplying oxygen to the body and for the elimination of carbon dioxide from the blood, or the gaseous exchanges between air and blood (a process known as hematosis). Located in the thoracic cavity, they are surrounded by a serous membrane, the pleura, which is essential for the performance of their functions.

Lungs are separated by a space between the spine and the sternum, the mediastinum, which includes the heart, esophagus, trachea, bronchi, thymus and great vessels. Each of the two lungs has, at its upper end, an apex that extends upwards to the base of the neck and, at its lower end, it rests on the diaphragmatic muscle. Their main purpose is to receive the load of carbon dioxide and waste products from the peripheral circulation and to clean up blood: once cleansed, blood is then sent to the heart, from where it is distributed to organs and tissues. An example of healthy lungs X-ray image is shown in Figure 3.1

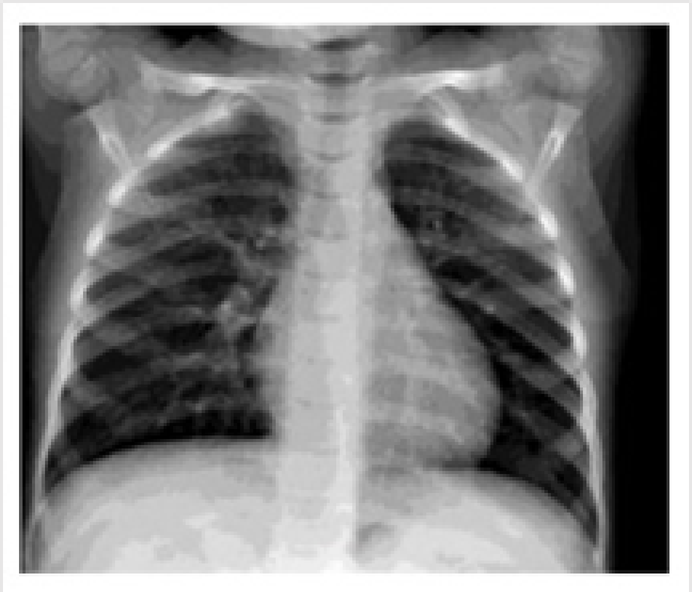


Figure 3.1 Example of healthy lungs on a chest X-ray image

As a general case, when lungs are struck with pneumonia, they fill with fluid and become inflamed, causing difficulty when breathing. For some cases, breathing problems can become severe and require hospitalization with oxygen and ventilator treatments. The kind of pneumonia caused by COVID-19 tends to take hold in both lungs. The air sacs in the lungs fill with fluid, limiting their capability to absorb oxygen and causing shortness of breath, cough, and other symptoms. Even after the disease has passed, lung lesions can cause breathing difficulties that could take months to improve/disappear. An example can be observed in Figure 3.2.

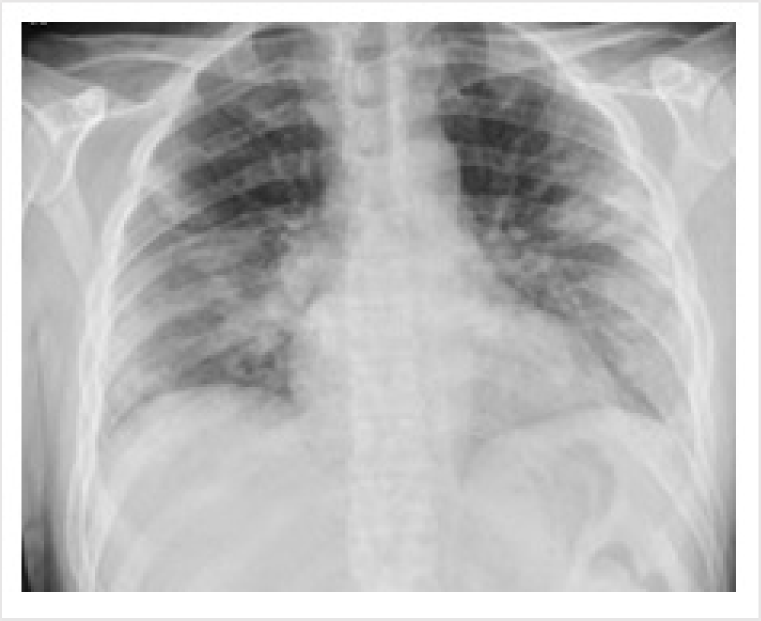


Figure 3.2 Example of COVID-struck lungs on a chest X-ray image

Pulmonary opacity is represented by spots that appear on the lungs and do not usually exceed 3 cm in diameter. In most cases they are benign, meaning they are not cancerous. A pulmonary nodule is usually identified by means of chest X-ray or computed tomography (CT). They may either appear as single nodules or there may be several. A cancerous lung lump is usually larger than 3 cm and can be irregular in shape. Such nodules can be seen in Figure 3.3.

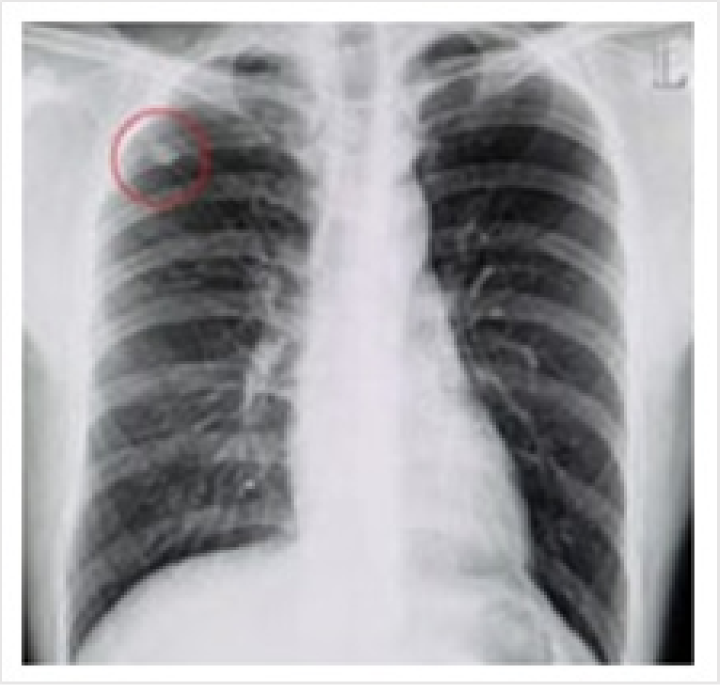


Figure 3.3 Example of lung opacity on a chest X-ray image

Viral pneumonia is defined as a pathological entity in which there is a viral cause of abnormal oxygen and carbon dioxide gas exchanges in the alveoli, secondary to virus-mediated inflammation and/or immune response [40]. In X-ray images, areas of the chest are generally visible as lighter, whitish spots in the regions affected by pneumonia, as shown in Figure 3.4.

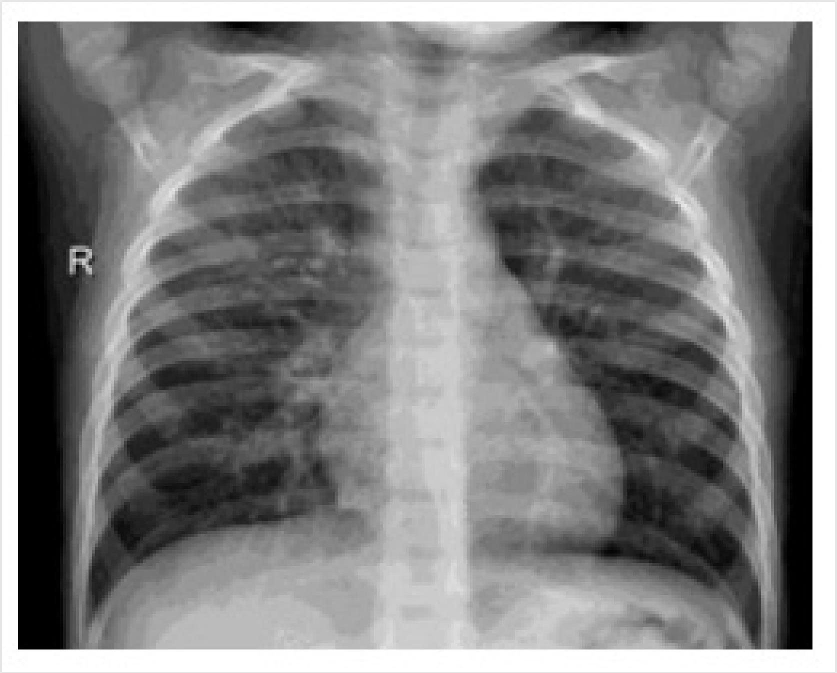


Figure 3.4 Example of viral pneumonia on a chest X-ray image

3.3 Methodology

Ever since their pioneering use [41], Convolutional Neural Networks (CNNs) have proved themselves to be extremely powerful tools when it comes to image classification. The basic principle is the application of convolutional layers, which are able to extract significant features from images by means of a sequence of operations over a selected area of the image itself.

A typical Convolutional Neural Network shall appear as depicted in Figure 3.5.

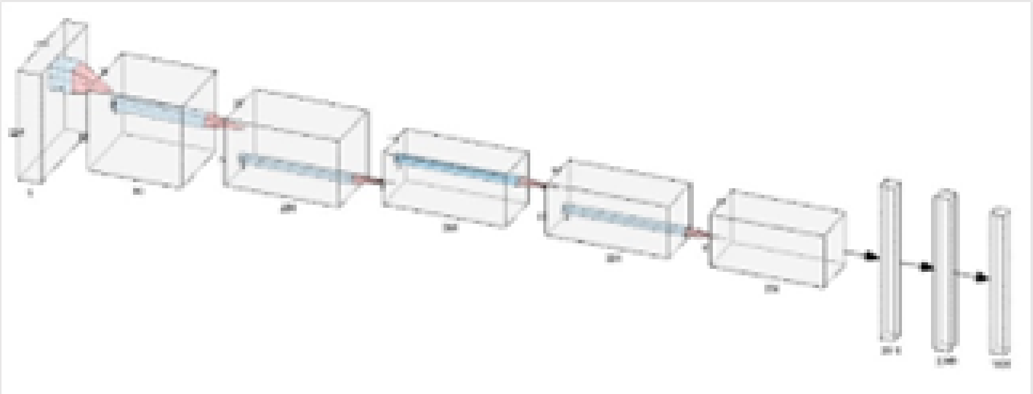


Figure 3.5 Example of a convolutional neural network

Each convolutional layer is generally followed by a pooling layer, which basically modifies the size of the input in order to make it suitable for the next stage. At the very top, a fully-connected or dense layer is present, with the purpose of classifying the input image into one of the given categories, generally applying a softmax function to the input. As powerful as they are, CNNs do exhibit some issues, such as the inability to retain information about the composition and position of specific elements within an image, and to pass such information on to subsequent layers. For this reason, several architectures were developed and presented in recent years. Specifically, Transformers have aroused great interest, especially in NLP applications [17]. In this work the focus is centered on what is probably the most popular version of the Transformer architecture for image classification, the Vision Transformer, or ViT [16]. The peculiar structure and basic working principles of this network are represented in Figure 3.6.

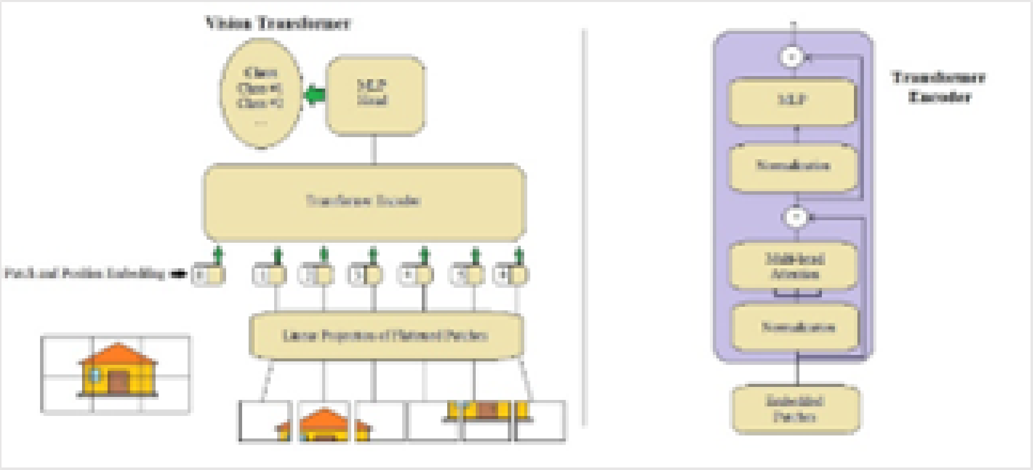


Figure 3.6 Vision Transformer architecture

Input images are divided into patches, and the linear projection of flattened patches is embedded, in order to preserve positional information. Embedded patches are arranged into a sequence and then fed to the Encoder, which exploits the multi-head attention technique to extract information, patterns and relationships among image patches. Eventually, outputs are fed to a Multi-Layer Perceptron to perform classification.

In order to better clarify the similarities and differences between CNNs and ViT, a brief of the architectures of three among the most relevant and popular Convolutional Networks will follow. InceptionV3 originated as a module for GoogLeNet [42], with the purpose of allowing for deeper networks without increasing too much the number of parameters.

1x1 convolution blocks were introduced to reduce dimensionality. Such layers act as rectified linear activators as well, so their purpose is two-fold. The next CNN chosen for analysis and comparison is Xception [43], which was described as an "extreme" version of InceptionV3 with the exploitation of the so-called *depthwise separable convolution*, and a subsequent redefinition of the Inception module. The basic underlying concept is the assumption that cross-channel correlation and spatial correlation can be mapped separately. This leads to the idea of using a 1x1 convolution to map cross-channel correlations at first, and apply 3x3 convolutions to map spatial correlations later on. This has been proved to slightly outperform InceptionV3. Remarkably, the "middle flow" section of the network presents a skip connection, which is indeed the key element of the next CNN to be described, ResNet50. First introduced by He et al. [44], ResNet had the peculiarity of using skip connections to tackle the problem of vanishing gradients. This approach was successful and resulted in a number of variations of the original topology.

For a better understanding of what the Vision Transformer performs on images, it is worth to delve deeper into the concept of *self-attention*. This idea is derived from the field of Natural Language Processing. In fact, when it comes to translating a text into another language, it is necessary to be aware of the position of each word with respect to each other, and to the context of the sentence in order to give the proper meaning to each word. In this sense, self-attention tries to mimic the thoughts and procedures behind a language translation process.

The whole mechanism starts by assigning to each input three vectors to represent it, namely *key*, *query* and *value*. All of them are obtained by multiplying the input vector by a set of weights (which need to be initialized). Subsequently, each query is multiplied – through a dot product – by each key, and the resulting output is scaled by a factor equal to the square root of the dimension of the key vector. The result goes through a softmax operator, and is later multiplied by the value vector. The outcome of this sequence of operations is the *attention score* of the given input. Each block performing such sequence is called *head*.

Keys, queries and values can be linearly projected to d_k, d_q and d_v dimensions, where d_k, d_q and d_v are the dimensions of key, query and value vectors, respectively. Thus, attention can be evaluated N times in parallel on each projected version of keys, queries and values, in what is called *multi-head attention*.

After creating image patches, the Vision Transformer embeds the linear projection of flattened patches and feeds the embedding to the Encoder Transformer block, allowing for multi-head attention to capture feature and relationships among patches.

Although the peculiar architecture of the Vision Transformer revolves around heads and self-attention, it does not stop at that. In fact, the inputs to each head are embedded both linearly and then again by using sine and cosine functions at different frequencies. This allows to capture information about the position of the single patch with respect to the entire image. A learnable classification token (indicated with "0" on the left side of Figure 3.6) is prepended to the sequence of embedded patches so that the network can perform the classification task. In fact, the state of this token at the output of the Multi-layer Perceptron shall represent the input image in its entirety, retaining information that is relevant for classification. Subsequently, embedded inputs go through a normalization layer before actually being fed to the Encoder block, and are combined to the Encoder output via a residual connection. The result is once again normalized and then fed to a Multi-Layer Perceptron, which consists of two fully-connected layers with a GELU activation function, defined as follows:

$$GELU(x) = xP(x \leq X) = x\Phi(x) = x \cdot \frac{1}{2}[erf(x/\sqrt{2}] \tag{3.1}$$

This section of the network is responsible for performing the actual image classification on the basis of all the information that was extracted and processed by the Transformer heads.

3.4 Experimental Setup

The first step of the process consisted in training and testing the three aforementioned CNNs over the chest X-ray dataset. All of the networks were trained by exploiting the transfer learning technique, which allows to retain and freeze network weights derived from previous training sessions over specific datasets. In this case, weights obtained over the ImageNet21k database [45] were used. Subsequently, only some of the top layer weights were set to be trainable. Indeed, the differences among the chosen architectures caused the number of trainable layers to change from one neural network to another. The reduced number of unfrozen layer weights were trained and tested over the chest X-Ray database.

As far as the Vision Transformer is concerned, it is worth noting that the fine tuning process for this architecture is rather different with respect to CNNs: in fact, all of ViT's weights are subtly modified during this process, and no layers are actually frozen.

The fine-tuning method allowed to significantly reduce the overall amount of time and computational resources dedicated to the training and testing phases for all convolutional networks.

All networks were trained and tested on a PC with a CPU@3.70GHz with TensorFlow 2.5.0 and Keras. Hyperparameters were configured in the very same fashion for all architectures: initial learning rate was set to 0.0001; fine-tuning learning rate to 0.00001; the chosen optimizer was Adam; batch size was set to 32;

dropout coefficient equal to 0.5; loss function of choice was the Sparse Categorical Cross Entropy function. This choice for the loss function is motivated by the fact that the classes of our expression are mutually exclusive, that is, each image belongs to exactly one class. The same number of 60 epochs was set both for the initial training epochs and for fine-tuning epochs, for a total of 120 epochs.

A difference was set in the layer selected to start the fine-tuning process, just as previously mentioned, as follows: the fine-tuning threshold was set at layer 308 (over 311 total layers) for InceptionV3; at layer 128 (over 132 total layers) for Xception; at layer 172 (over 175 total layers) for ResNet50. The next step was the deployment of the Vision Transformer architecture, or ViT, which can be seen as some kind of equivalent of the BERT Transformer [46] applied to vision and image classification. Once again, transfer learning was exploited in order to reduce the total amount of time spent on training, validating and testing.

The following hyperparameters were used: the base architecture is ViT-B_32, which is based on the "base" version of BERT (12 layers, a hidden size set to 768, 12 heads and a patch size of 32x32 for the input); batch size was set to 32; learning rate was set to 0.00001; the selected optimizer was Rectified Adam; loss function of choice was the Categorical Cross Entropy function; label smoothing was set to 0.2; overall number of epochs was 30. All settings and hyperparameters were chosen in order to make a fair comparison against CNNs, all the while taking into account the significant architectural differences between CNNs and the Vision Transformer.

In order to contrast data imbalance among classes, data augmentation with random horizontal flipping and random rotation (set to 0.2), as well as Mitchell-Netravali [47] filtering were applied to the database images before feeding them to the CNNs. On the other hand, no operation of any kind was performed before feeding the images to the Vision Transformer, with the exception of resizing them from an initial resolution of 299x299x3 to 224x224x3 in order to fit the ViT input layer.

The dataset was split using 70% of data for training, 10% for validation and 20% for testing. The split was kept identical for each model.

3.5 Results

Table 3.1 shows a comparison of the results of the Vision Transformer versus the convolutional networks.

Table 3.1 Performance comparison between Vision Transformer and Convolutional Neural Network architectures

Network Architecture	Test Accuracy
InceptionV3	0.7936
Xception	0.8362
ResNet50	0.8558
ViT	**0.9930**

ResNet50 exhibits the best performance among the convolutional neural networks of choice, with a test accuracy of about 86%. However, the Vision Transformer architecture clearly outperforms the selected CNNs on this specific image classification task with an outstanding accuracy of 99.3%. A few more indicators are shown in Table 3.2 and Figures 3.7, 3.8, 3.9 and 3.10 to further describe the performance of the ViT architecture over the four classes of the database: precision, recall, F1-score, a visualization of the attention map and the confusion matrix, all of which are metrics and parameters commonly used to assess the performance of deep learning architectures over given tasks – like image classification. The Support column in Table 3.2 refers to the number of images per category that were used to test the Vision Transformer ability to assign images to each category.

Table 3.2 Precision, recall and F1-score for Vision Transformer

Class	Precision	Recall	F1-score	Support
COVID (class: 0)	0.97	0.94	0.96	353
Lung Opacity (class: 1)	0.87	0.93	0.90	602
Normal (class: 2	0.95	0.92	0.94	1019
Viral Pneumonia (class: 3	0.96	0.98	0.97	135

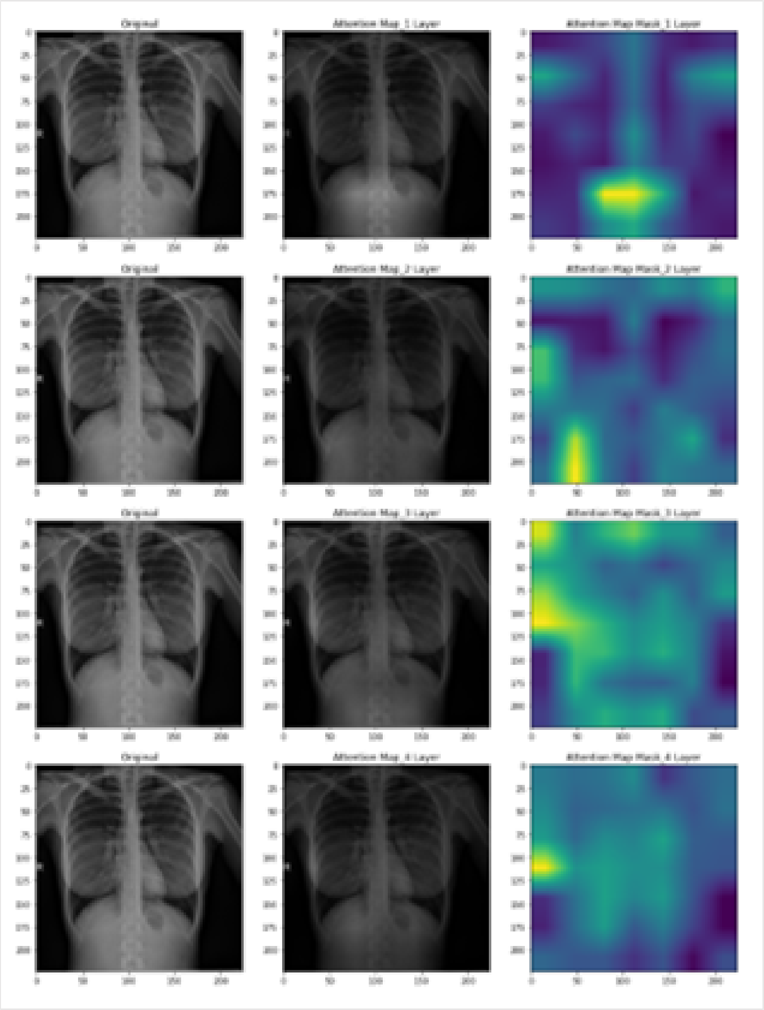


Figure 3.7 Example of Vision Transformer attention maps on a COVID-19 chest X-ray image. Attention maps for layers 1 to 4 are shown

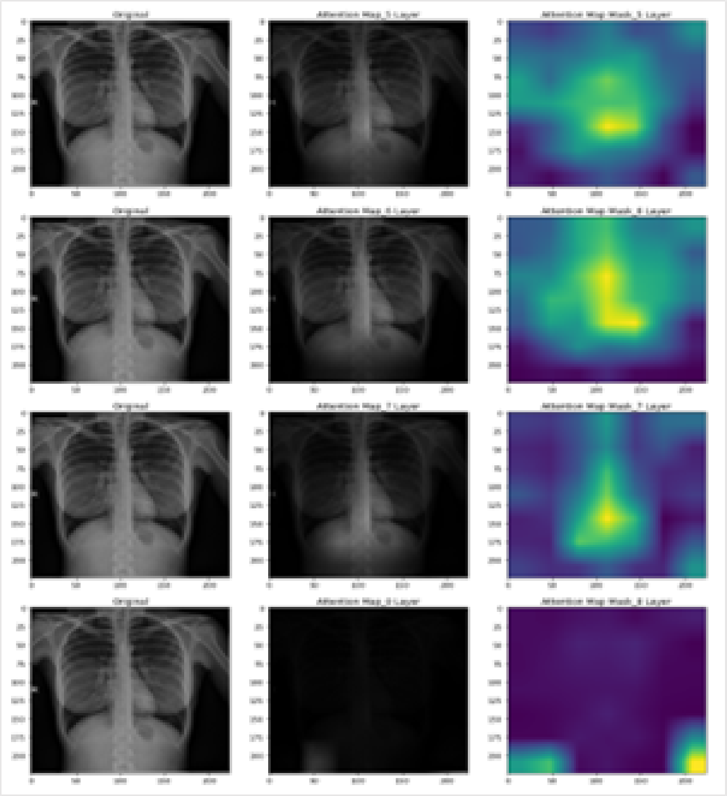


Figure 3.8 Example of Vision Transformer attention maps on a COVID-19 chest X-ray image. Attention maps for layers 5 to 8 are shown

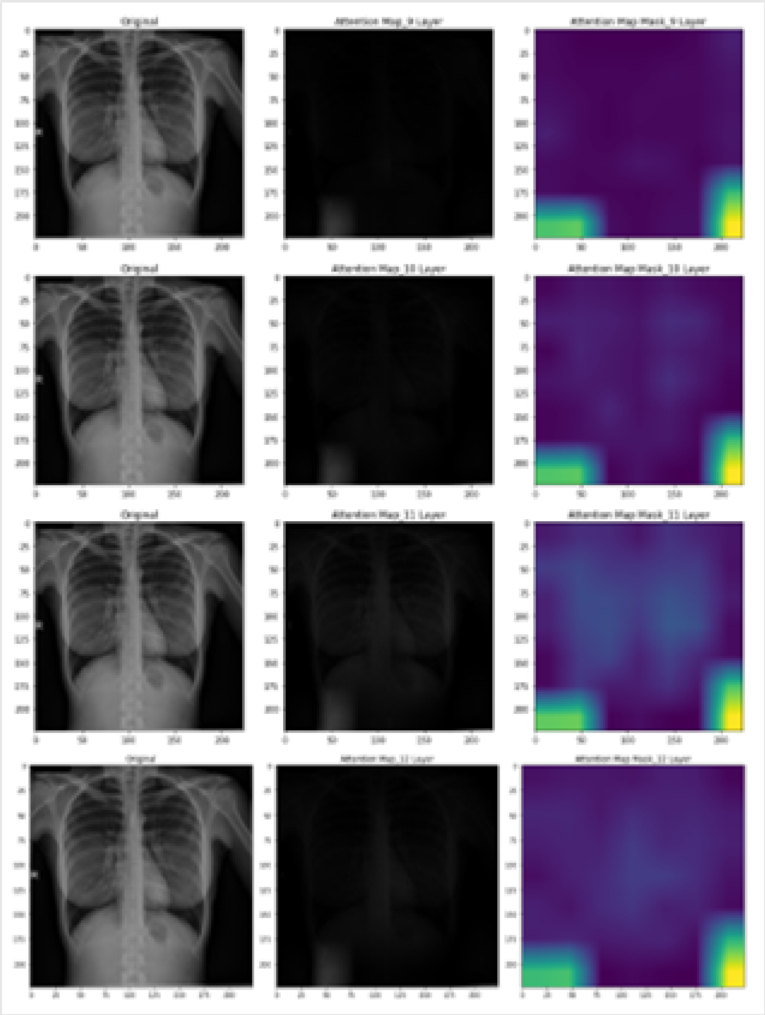


Figure 3.9 Example of Vision Transformer attention maps on a COVID-19 chest X-ray image. Attention maps for layers 9 to 12 are shown

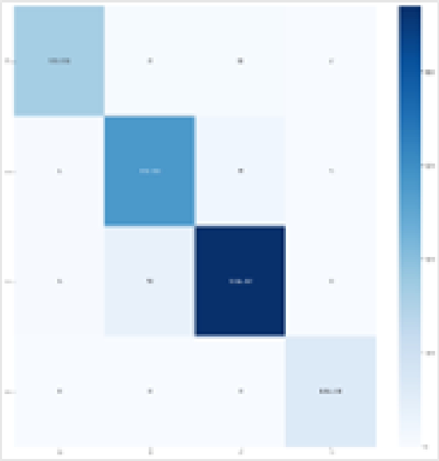


Figure 3.10 Confusion Matrix on the test set

These results show that the Vision Transformer is highly capable to correctly classify images in each category. This is further verified when evaluating metrics like precision, recall and F1-score, which allow for a better insight of the Vision Transformer performance. In fact, given the overall number of images classified by the network into a given category, precision expresses the ratio of how many of those images have been correctly included by the Transformer into that category. On the other hand, considering the number of images that actually belong to a specific category, recall indicates the proportion of those images that were correctly associated to that class by the neural network. Typically, a high precision score implies a poor recall value, and viceversa. For this reason, F1-score is also taken into account, since it represents a sort of combination of precision and recall into a single metric. As it is possible to observe, all three parameters are remarkably high for the Vision Transformer architecture over all four classes.

The attention map, which can be defined as a matrix that represents the relative importance of layer activations at different spatial locations with respect to the given task, can be observed in Figures 3.7, 3.8 and 3.9 for all 12 layers of the Vision Transformer. Even though it might not convey any particular information to the human eye, the attention map can help to analyze the behaviour of a network, since it visualizes some aspects of the input image that are interpreted as relevant features by the architecture, thus leading it to assign such image into a specific class.

As a closing remark, it is worth highlighting that the Vision Transformer is able to reach a significantly higher accuracy with respect to Convolutional Neural Networks after iterating for only 30 epochs, as opposed to 60 + 60 = 120 overall epochs to train and fine tune the other architectures.

3.6 Discussion

Results presented so far have proved that a specific architecture, the Vision Transformer, is able to achieve a significantly better performance with respect to other network configurations on this peculiar application. This paves the way to the exploitation of the Vision Transformer, and attention-based networks in general, for the purpose of assisting, accelerating and automatizing clinical diagnosis. Indeed, a fast, accurate and reliable tool to promptly identify lung infections can assume crucial relevance when the disease of interest is the cause of a pandemic situation. However, one of the main characteristics of deep learning, and of neural networks in general, is the lack of transparency, meaning that the mechanisms that lead such algorithms when making decisions are often obscure. This often leads to situations in which the network excels at performing its tasks on a given dataset but is unable to generalize over different scenarios. This becomes particularly significant when the primary purpose of the algorithm is to provide a fast and reliable response when assisting physicians in clinical diagnosis. For this reasons, future work in this direction should be dedicated to shed some light on what might lead a deep neural network into making a specific choice when facing alternatives, trying to bring clarity into what has traditionally been perceived as a black box. For instance, an interesting path to follow could be represented by the kind of ideas described by Hassani et al. [48], who tried to exploit the convolutional networks' capabilities to extract significant features from images and feed those information to a Transformer, as opposed to using the patching and embedding method. Another way could be the analysis of how initial data is divided into clusters, and a subsequent comparison against the outcome of the classification task performed by the transformer, in order to investigate the factors that push the network to put a given image into a certain category.

Chapter 4

Skin Lesion Detection on Dermatoscopic Images

Part of the work presented in this chapter was previously published on published on *Sensors* [49] and accepted for presentation at the 18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics (CIBB 2023) hosted by University of Padova.

4.1 Context and Motivation

The World Health Organization (WHO) reported that skin cancer is one of the most prevalent cancers globally with over a million new cases yearly, making up one-third of all cancer diagnoses [50]. WHO has identified UV radiation as the leading cause, suggesting limiting sun exposure and checking the skin regularly for unusual lesions. In such contexts, Melanoma, a form of skin cancer, has seen a significant increase over the past 30 years [51], which arises from pigment-producing cells (the melanocytes). For such reasons, early diagnosis and treatment are fundamental to greatly improve a patient's prognosis, as well as stopping the spread of the disorder. In the literature, computer-aided diagnosis (CAD) technologies have been used to assist in the detection of skin cancer [52]. CAD systems, incorporating Machine Learning (ML) and Deep Learning (DL) models, are able to examine dermatoscopic skin images, including melanoma, identifying unusual tissue patterns and categorizing them into either cancerous or non-cancerous groups [53]. Specifically, many works have focused on using ML and data mining techniques to classify skin melanoma disease [54, 55]. Among DL methods, both convolutional neural networks (CNNs) and Vision Transformers (ViTs) have been used to detect and classify melanoma and skin cancer by showing promising results [56–58].

While these models already display outstanding results in detecting and classifying skin diseases, in this study it is demonstrated how the utilization of synthetic images - generated using a new Diffuser Data Augmentation method - may enhance the performance of a ViT-based classifier for the classification of dermatoscopic melanoma images. The experimental outcomes, derived from a public melanoma skin cancer dataset, not only affirm the practicality of this technique but also showcase a superior level of accuracy when compared with the traditional ViT-based classification approach.

4.2 Dataset

The considered dataset is a collection of 2000 dermoscopic images (in JPEG format) of skin lesions, publicly available at the ISIC Archive [59] and used for the ISIC 2017 Challenge - Skin Lesion Analysis Towards Melanoma Detection [60]. The objective was the classification of lesions, with images classified into three categories: melanoma (MM), nevus (NCN) and seborrheic keratosis (SK). The training set included 374 cases of MM, 254 cases of SK and 1372 of NCN. The validation and test sets contained 150 and 600 images, respectively, with various sizes, angles and illumination conditions (see Table 4.1).

Table 4.1 Details of the ISIC dataset

Class	**Training Set**	**Validation Set**	**Test Set**
MM	374	30	118
NCN	1372	78	392
SK	254	42	90
Total	2000	150	600

4.3 Methodology

The first approach consisted in fine-tuning the ViT-Large model (by using 32×32 patches) on the given dataset, with the purpose of developing a system able to distinguish between the different categories of skin lesions with a high degree of accuracy and specificity. To this aim, a combination of standard data augmentation techniques was applied to the training set: random horizontal flip followed by a sequence of transformations as implemented by the RandAugment library [61].

However, the intrinsically high severity of melanoma with respect to the other categories made it necessary to attempt and increase detection capabilities for such skin lesion class. Therefore, a DDPM was trained using melanoma images from the original training set, with the purpose of generating high-fidelity artificial melanoma pictures. Such model was later used to create 998 synthetic images that were included in the original training set, in order to match Nevus cardinality - the highest one in the database (see Table 4.2).

Table 4.2 Details of the ISIC dataset after the Diffuser-based augmentation

Class	Training Set	Validation Set	Test Set
MM	1372 (374+998)	30	118
NCN	1372	78	392
SK	254	42	90

Examples of true and generated melanoma images are displayed in Figure 4.1. It has been observed a greater variability in the generated image distribution. The resulting dataset was employed to train the fine-tuned model from scratch, effectively increasing sensitivity for melanoma cases - and overall accuracy.

4.4 Experimental setup

In order to determine the optimal configuration, the early experimental phase was subdivided into two main steps: i) a search for the best learning rate, and ii) an ablation study, with the dual purpose of investigating the effect of network depth on performance, and determining the proper trade-off between model complexity and efficiency. Specifically, to determine the best learning rate, a systematic search was

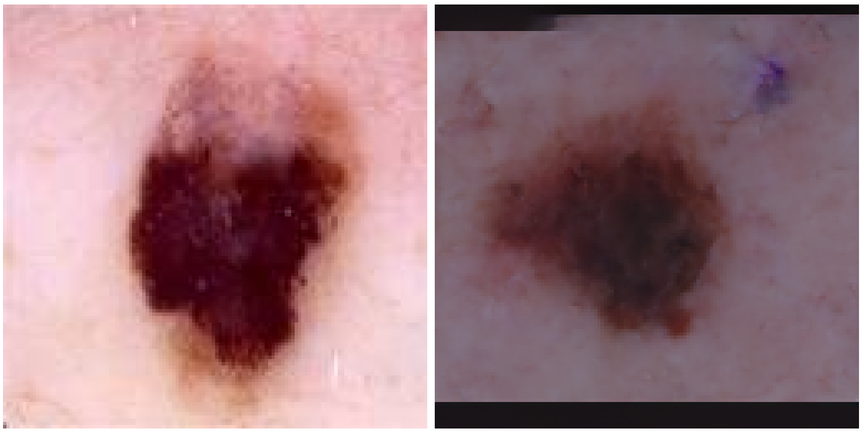


Figure 4.1 Examples of a real melanoma image (left) and a synthetic one (right) generated via the Diffuser.

conducted by training the model with different learning rate values. The procedure started with a relatively large learning rate ($1.5e^{-5}$) which was gradually reduced (down to $0.5e^{-6}$), monitoring the model's performance at each iteration. The model's performance metrics were evaluated considering the accuracy on a validation set to identify the learning rate that resulted in the best performance. The ablation study served two main purposes: a) examining the impact of network depth on performance to understand how it influenced the model's effectiveness, and b) striking a balance between model complexity and efficiency. Through these two steps in the experimental phase, the main objective was to refine the model and identify the optimal configuration that maximized performance while taking into account computational efficiency.

Subsequently, an additional experiment was carried, in order to evaluate the impact of the data augmentation procedure. The performance of the two training approaches (i.e. without and with data augmentation) was assessed as described in the next section.

Experiments were conducted in the Google Colab environment, deploying the models available on Hugging Face. The best-selected model was ViT-Large (307 million parameters), with an input resolution of 384 × 384 and a patch size of 32 × 32.

4.5 Results

The proposed model, pre-trained on ImageNet [62], was fine-tuned on the ISIC 2017 dataset for 40 epochs, using the hyperparameters listed in Table 4.3. The ViT-based model was trained using several configurations, listed in Table 4.4. In particular, the performance was evaluated by varying the learning rate of the Adam optimizer and changing the number of hidden layers. The scheduler, used for the training phase, applies a linear learning rate variation with a warm-up: when the set-point for the learning rate (maximum value to be reached) is chosen, the scheduler starts from a lower value until the set-point is reached, and decreases with a gentler slope in the last epochs. In particular, having set the number of layers to 24, different set-points were set (i.e. $0.5e^{-6}$, $0.5e^{-5}$, $1.0e^{-5}$, $1.5e^{-5}$). Such a scheduling method guarantees that the experiments explore the influence of the learning rate over a sufficiently wide dynamic range. In general, the effect of high learning rate values is reflected in a learning curve trend that does not decrease effectively, as a consequence of the model not learning how to classify properly. On the other hand, low learning rate values can lead the system to learn rather efficiently (aside from the risk of getting stuck in local minima). However, this usually requires a large number of epochs to reach convergence. Considering this, since with the values $0.5e^{-5}$, $1.0e^{-5}$, and $1.5e^{-5}$ the system obtains the same accuracy value (94.83%), it was decided to choose the configuration with an intermediate learning rate. The proposed intermediate learning rate value requires a reasonable number of epochs for training.

Table 4.3 Model hyperparameters

Hyperparameter	**Value**
number of layers	24
number of heads	16
hidden size	1024
optimizer	Adam with β_1=0.99, β_2=0.999, $\varepsilon = 1e^{-8}$
learning rate	$1.0e^{-5}$
learning rate scheduler	linear with a warm-up ratio of 0.1
batch size	16
gradient accumulation steps	4

A large difference in performance, on the other hand, was found as the number of layers varied. Figure 4.2 shows how accuracy varies as the number of layers

Table 4.4 Tested configurations.

Configuration	Batch Size	Learning rate	Layers	Accuracy
#1	16	$1.5e^{-5}$	24	0.948
#2 (Best Model)	**16**	**1.0e^{-5}**	**24**	**0.948**
#3	16	$0.5e^{-5}$	24	0.948
#4	16	$0.5e^{-6}$	24	0.945
#5	16	$1.0e^{-5}$	20	0.925
#6	16	$1.0e^{-5}$	16	0.902
#7	16	$1.0e^{-5}$	12	0.850
#8	16	$1.0e^{-5}$	10	0.817
#9	16	$1.0e^{-5}$	8	0.788
#10	16	$1.0e^{-5}$	4	0.707
#11	16	$1.0e^{-5}$	2	0.655

changes. In particular, different values in the range [2 - 28] have been tried. As shown in Figure 4.2, accuracy has an increasing trend as the number of layers increases (as expected): the increase is greater as the first few layers increase, whereas subsequently, the growth becomes slower. In the range [24 - 26] the accuracy presents the maximum, where the curve has a plateau, and the accuracy stays constant at 94.83%, while for higher values the accuracy begins to decrease. Indeed, this is a consequence of overfitting, since the model complexity is increasing. Considering these findings and following *Occam's razor*, a value of 24 was chosen for the number of hidden layers (which is also the default value for the ViT-Large architecture), minimizing the architecture complexity and optimizing classification performance.

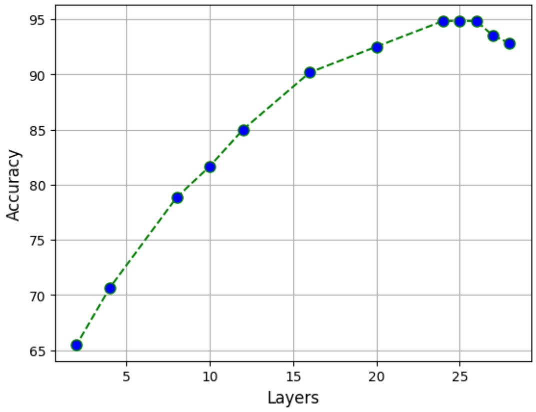


Figure 4.2 Variation in accuracy as a function of the number of layers.

Regarding the best configuration, the model reached a final test accuracy of 0.948. Figure 4.3 shows the training and validation losses. It is possible to observe a drop in both curves at epoch #5, which is likely due to the warmup ratio of the learning rate scheduler. Moreover, the curves' peaks that can be seen on the plot might be associated with the selected batch size. The test confusion matrix for configuration #2 is shown in Figure 4.4.

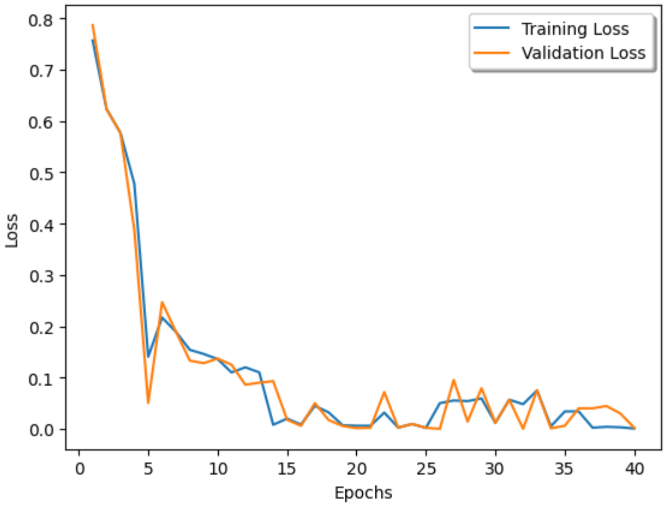


Figure 4.3 Training and Validation Loss of configuration #2 (**best model**).

The purpose of the additional experimental tests was to evaluate the improvement brought about by the data augmentation procedure. The performance of the two training approaches (i.e. without and with data augmentation) are assessed and com-

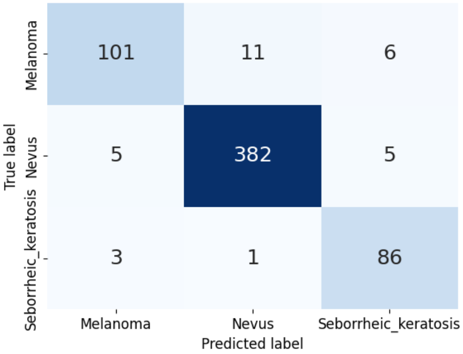


Figure 4.4 Confusion Matrix for configuration #2 (**best model**).

pared by observing the confusion matrices and the loss curves shown in Figure 4.5 and Figure 4.6, respectively.

Specifically, Figure 4.5, reporting the confusion matrices calculated on the test set, shows a decrease of false negatives (from 17 to 10), thanks to the fact that the Diffuser images help enhancing the melanoma information, while Figure 4.6 shows a slower convergence for the enhanced Transformer, because of the wider distribution of synthetic images.

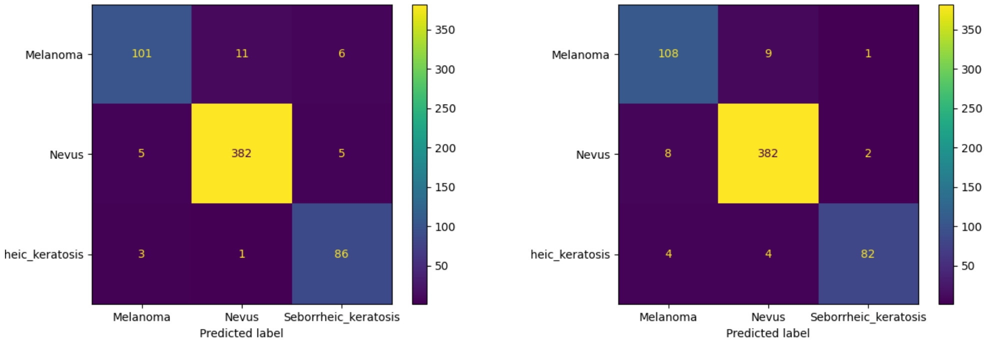


Figure 4.5 Confusion matrices obtained by the ViT-based classifier without (left) and with (right) the use of the Diffuser for data augmentation.

Table 4.5 compares the classification performance obtained by the ViT-based classifier without and with the use of the Diffuser data augmentation.

The Nevus class exhibits no relevant changes either on the confusion matrix or on any other metric, with the exception of a small specificity decrease. Seborrheic

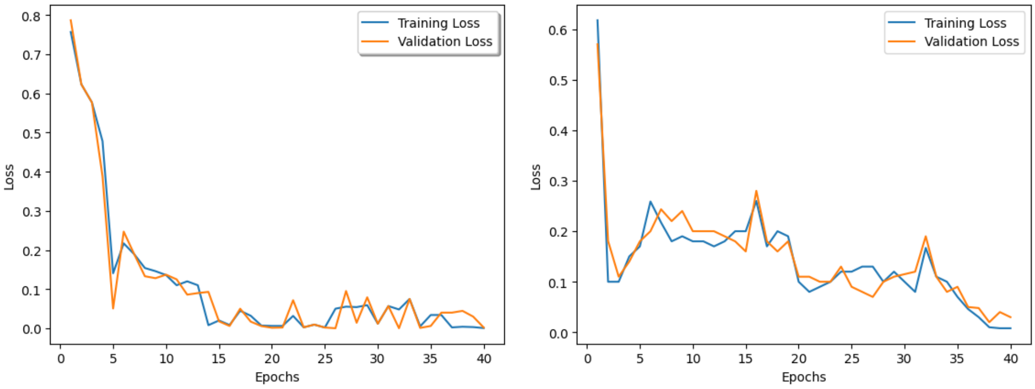


Figure 4.6 Train and validation losses *vs.* epochs without (left) and with (right) the use of the Diffuser for data augmentation.

keratosis experienced slight increases in accuracy and specificity, at the expense of reduced sensitivity and AUROC scores. Melanoma, on the other hand, suffered a marginal specificity decrease, whilst showing increased scores on all other metrics - namely, AUROC and sensitivity. These results clearly demonstrate the benefits of leveraging Diffusers to effectively contrast class imbalance by artificially enhancing the training set.

Table 4.5 Results in comparison between the training with and without data augmentation via the Diffuser. Performance was computed in the one-*vs.*-rest strategy.

Class	**Evaluation metric**	**ViT-based**	**Vit-based + Diffusers**
Melanoma	accuracy$_{MM}$	0.958	0.963
	sensitivity$_{MM}$	0.856	0.915
	specificity$_{MM}$	0.983	0.975
	AUROC$_{MM}$	0.919	0.945
Nevus	accuracy$_{NVN}$	0.963	0.962
	sensitivity$_{NVN}$	0.974	0.974
	specificity$_{NVN}$	0.942	0.938
	AUROC$_{NVN}$	0.958	0.956
Seborrheic Keratosis	accuracy$_{SK}$	0.975	0.982
	sensitivity$_{SK}$	0.955	0.911
	specificity$_{SK}$	0.978	0.994
	AUROC$_{SK}$	0.967	0.953

4.6 Discussion

In the literature, various methods have been proposed to improve the decision-making process in the diagnosis of skin lesions, particularly in cases of melanoma. Authors in [63] proposed a comprehensive comparative study of U-Net and attention-based skin lesion image segmentation methods. The results indicate that the hybrid TransUNet outperforms other benchmarking methods on segmentation tasks, achieving an accuracy of 0.921 and a Dice coefficient of 0.898. Authors in [64] proposed the use of Fully Transformer Networks (FTN) for skin lesion analysis, a hierarchical Transformer that uses a Spatial Pyramid Transformer (SPT) to capture long-range contextual information from skin lesion images. Their conducted experiments on the public International Skin Imaging Collaboration (ISIC) skin lesion segmentation and classification datasets have demonstrated the effectiveness and efficiency of FTN. In [65], the authors presented a novel deep-learning-based framework, named Attention Deeplabv3+, for skin lesion segmentation. Their proposed method uses attention mechanisms in two stages to capture the relationships between channels and to emphasize more on the relevant field of view. Their experimental results on public skin cancer datasets showed high state-of-the-art performance. Authors in [66] proposed a novel self-attention-based network for diagnosing melanocytic lesions from digital whole slide images. The method outperforms other state-of-the-art methods and achieves results comparable to 187 practising U.S. pathologists. In [67], the authors proposed a framework that employs methods for data augmentation and a Medical Vision Transformer-based classification model for classifying skin cancer. The results achieved on the HAM10000 datasets showed that the proposed model outperforms state-of-the-art techniques for skin cancer classification, proving that early detection can increase survival rates by up to 70%, leading to improved outcomes for patients with this deadly disease. However, the potential of ViT has not yet been exploited to its fullest in the classification of melanoma skin cancers.

Melanoma detection is a critical task in improving cancer diagnosis and prognosis. To this aim, this work proposed a ViT-based architecture for efficient MM recognition compared to NCN and SK.

The proposed model (including training and testing procedures) was built-up using publicly available skin cancer data from the ISIC challenge. This choice enables a fair comparison with other competitors who have employed similar methods to address the same clinical issue by using the same dataset. The performance in classifying skin lesion images of the introduced architecture, based on ViTs, outperforms the current state-of-the-art models. This improvement can be attributed to the model's ability to capture and model long-range spatial relationships within the images effectively.

Table 4.6 shows the performance of the proposed ViT-based approach and a comparison with the other state-of-the-art solutions that used the ISIC 2017 dataset. It should be noted that in the specific application scenario approached by this study, there is no possibility of changing (increasing) the number of classes (i.e. MM, NCN, and SK). Indeed, the number of classes is established in the ISIC 2017 skin lesion classification challenge, to which the dataset refers. The optimal configuration is able to accurately identify images from all three categories, showing remarkable classification capabilities on the given dataset. The accuracy of 0.948, the sensitivity of 0.928 and the specificity of 0.967 are significantly higher compared with other models. Only the AUROC results are a bit lower when compared to the other works. Reducing the number of layers has led to a decrease in accuracy. Yet, it is interesting to observe that employing a ViT-based architecture with 20 layers instead of 24 still yielded a final accuracy above the state of the art.

Table 4.6 Model Test Performance on the ISIC 2017 dataset.

Model Name	**Accuracy**	**Sensitivity**	**Specificity**	**AUROC**
IRv2+Soft Attention [68]	0.904	0.916	0.833	**0.959**
ARL-CNN50 [69]	0.868	0.878	0.867	0.958
SEnet50 [70]	0.863	0.856	0.865	0.952
RAN50 [71]	0.862	0.878	0.859	0.942
ResNet50 [44]	0.842	0.867	0.837	0.948
Proposed ViT-based approach	**0.948**	**0.928**	**0.967**	0.948

An in-depth analysis was given in order to assess the influence of hyperparameters and, consequently, to select the best model configuration, in terms of batch size, learning rate, and number of layers. First of all, the influence of the learning rate during the training phase was evaluated. After selecting the best learning rate value, further investigations were conducted to explore model performance as the number of layers within the architecture. This process was motivated by the necessity to balance efficacy and complexity. Indeed, despite the apparent direct proportionality between performance and network depth, the availability of a tool that can guarantee comparable results when deployed on devices with limited computational resources (e.g. mobiles, tablets, smartwatches, increasingly used in e-Health and m-Health) is paramount. Furthermore, extensive ablation studies were conducted to explore the impact of individual components of the ViT model, which highlight their respective contributions to overall performance.

The proposed architectural solution represents a trade-off between computational cost and classification performance: the study performed in the tuning phase of the hyperparameters allows to obtain a predictive model with very high sensitivity/specificity, corresponding to very low values of *falses* (positive as well as negative). This is due to the self-attention mechanism, which, by taking into account the correspondences between patches, can better understand the image's content. The Transformer, in general, works in overfitting. The described ablation analysis also confirms this, which shows that not all layers are needed. Considering that the last layers on the Transformer represent the most abstract relations in the image, it can be deduced that the keys to a better classification require only low-level features of the image. However, the high computational cost of the Transformer must not be neglected, implying that the proposed classification is offline.

Despite the better results of the proposed approach at convergence, the dynamics of training (see Figure 4.3) can represent an issue because of the presence of high peaks. This phenomenon is probably due to the small size and the internal covariate shift of the mini-batches.

As for the data augmentation techniques - and the related experiments - this study shows the advantage of leveraging Diffusers for synthetic image generation to increase the number of training samples, despite a longer convergence time. In the evaluations made (considering the one-*vs.*-rest strategy), all metrics show values well above 90% (the only exception being the sensitivity associated with melanoma). It has been noticed that the use of synthetic images for melanoma class reduced by 41% the false negatives. This demonstrates that the classifier performs very well with all three classes (i.e. MM, NVN, and SK). Model built-up by exploiting also synthetic images allows to obtain a more balanced performance, in terms of sensitivity and specificity.

Chapter 5

Mammographic Image Analysis for Breast Cancer Detection

Part of the work presented in this chapter was previously presented at the 31st edition of the Italian Workshop on Neural Networks (WIRN 2023) hosted by the Italian Society of Neural Networks (SIREN).

5.1 Context and Motivation

Breast cancer represents the primary health concern for women, and the utilization of mammographic screening has proven to be highly beneficial in reducing mortality risk by 58.2% in 2021 [72]. However, the presence of factors such as human perception, breast density, and the unique nature of cancer itself, contribute to significant rates of both false positives and false negatives [73]. Convolutional Neural Networks (CNNs) have demonstrated promising outcomes in addressing medical imaging challenges among various deep learning methods [74]. However, CNNs have limitations in comprehending long-range spatial connections within images. To tackle this issue, Vision Transformers (ViTs) have emerged as a recent and effective proposal, adapting the power of Transformers to the domain of medical computer vision [75]. By using self-attention mechanisms, ViTs enable the model to capture global relationships and dependencies between image patches, allowing for a more comprehensive understanding of the image as a whole. For this purpose, a ViT-based model was implemented in this work for breast cancer classification.

In addition, considering the huge amount of data needed for training, geometric and Diffuser-based data augmentation (DA) techniques were implemented. The Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) was used to train the ViT-based architecture and quantify the impact of DA techniques on classification performance.

The ViT-based model trained using the geometric DA achieved 69.25% accuracy, compared with 74.42% using the diffusion model. Eventually, original training images were combined with all generated images (both using geometric and Diffuser-based DA), yielding a large training set containing 13304 samples. Results suggest that this approach further improved the model performance reaching 77.01% accuracy, which is above the state-of-the-art.

5.2 Dataset

The CBIS-DDSM dataset is a collection of digital mammograms in *.png* format. This dataset is a subset of the larger Digital Database for Screening Mammography (DDSM), which also contains normal as well as benign and malignant cases with verified pathological information [76]. In this study, the choice was made to work with malignant and benign breast lesions. The CBIS-DDSM dataset was divided into three distinct subsets: training, testing and validation. The training subset, before applying DA, comprises a total of 930 mammography images, with an equal distribution between malignant and benign cases. For the test subset, a total of 348 mammography images are available, with 144 representing malignant cases and 204 representing benign cases. The purpose of this subset is to assess the performance and generalization of breast cancer detection algorithms on unseen data. To ensure reliable model validation, a portion of the training subset was set aside for validation purposes. Finally, a total of 233 images, 116 malignant samples and 117 benign samples, were assigned for validation, allowing for an evaluation of the algorithm's performance during the training process.

5.3 Methodology

5.3.1 Image preprocessing

This section describes pre-processing steps, which play a key role in optimising analysis and interpretation. The methodology involves manual cropping to eliminate artifacts, automatic thresholding to emphasise tumour regions, and subsequent automatic cropping to ensure image size uniformity for the CBIS-DDSM dataset. In Figure 5.1 the overall flow diagram of the implemented processing chain is shown.

First of all, to ensure optimal quality and eliminate unwanted artifacts, the images underwent a meticulous manual cropping process. As many artifacts (e.g. radiopaque markers) as possible were removed manually, allowing to focus exclusively on the regions of interest.

Successively, the Otsu thresholding technique was employed to mask only the whole breast, allowing a reduction of the influence of non-uniform black regions of the background. The automatic thresholding process significantly enhances the visibility and contrast of the breast area, enabling more accurate localization and subsequent analyses.

Finally, to minimise non-uniformity in image sizes and facilitate consistent analysis, the pre-processed images underwent automatic cropping to achieve uniform dimensions (640×640). Standardization of image size is crucial to eliminate bias. By standardizing the image size, potential variations and distortions are minimized, enabling fair and objective comparisons among different image samples. The automatic cropping process enhances the reliability and reproducibility of subsequent analysis techniques.

5.3.2 Geometric Data Augmentation

Classical Data Augmentation techniques were applied to enhance the final dataset containing mammograms. These techniques aimed to increase the diversity and variability of the dataset, improving the robustness and generalization capability of machine learning models. The following techniques were employed: (horizontal, vertical and diagonal) flips, exposure changes, noise addition, contrast enhancement, saturation enhancement, CLAHE, and blurring.

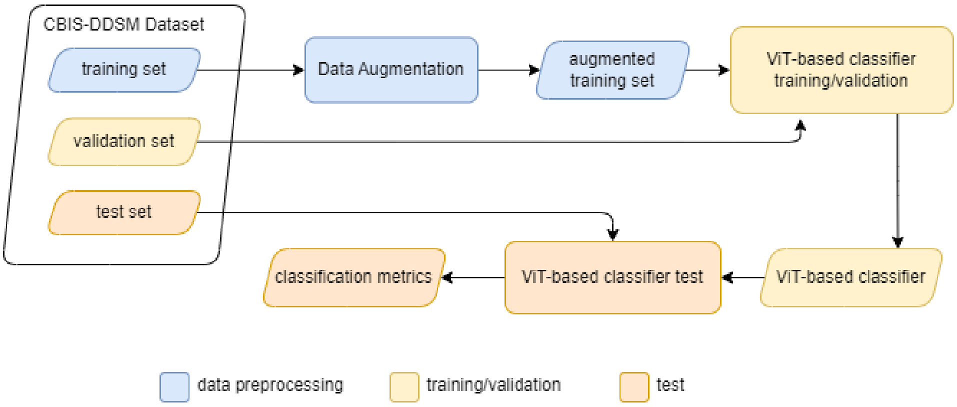


Figure 5.1 Overall flow diagram of the implemented processing chain

Horizontal and vertical flips were introduced to incorporate variations in breast orientation and position. These flips mirrored the images, enabling the models to learn from different perspectives of the breast. Moreover, diagonal flips ranging from $-30°$ to $+30°$ were introduced to further augment the dataset. Random diagonal flips added variations in the angular orientation, preparing the models to handle rotated breast orientations commonly encountered in clinical settings.

Exposure changes were introduced to simulate different acquisition conditions during mammogram acquisition. By modifying the brightness and contrast, the augmented dataset captured a wider range of exposure levels, preparing the models for variations encountered in real-world scenarios.

Noise addition enabled the models to learn to distinguish between relevant structures and noise artifacts, leading to improved performance in noisy environments. Noise was generated from a normal distribution with a mean of 0 and a standard deviation of 0.1, a noise pattern commonly observed in medical imaging devices.

Contrast enhancement techniques were employed to amplify the differences between structures and enhance their visibility. By adjusting the pixel intensities, contrast enhancement facilitated the detection of subtle patterns and abnormalities, contributing to improved accuracy in breast cancer detection. Moreovor, CLAHE (Contrast Limited Adaptive Histogram Equalization) was applied as a local contrast enhancement technique. CLAHE improved the visibility of both global and local structures, enabling the models to extract relevant information from different regions.

Saturation enhancement modified the colour saturation of the mammogram images. This technique captured variations in colour distribution, which can be informative for certain classification tasks where colour plays a significant role in characterizing different tissue types.

Blurring techniques, such as Gaussian filtering, were employed to reduce high-frequency noise and details. Blurring helped to remove noise artefacts and irrelevant features, allowing the models to focus on the more important structures and patterns.

By applying these classical DA techniques, the final dataset containing mammograms was enriched with 8373 images (4188 for malignant lesions and 4185 for benign lesions). This augmentation enhanced the dataset's diversity, enabling more effective training and improving robustness and generalization capability of machine learning models: this, consequently, brings better performance in breast cancer detection and classification tasks.

5.3.3 Diffuser Data Augmentation

In recent years, the use of Diffusion models has rapidly increased in the field of synthetic image generation, leading them to compete - and often outperform - other methods [77], like Variational Auto-encoders (VAEs) [78] and Generative Adversarial Networks (GANs) [79].

For this reason, a Denoising Diffusion Probabilistic Model (DDPM) [80] was trained, with the purpose of enhancing the training set with a collection of artificial images that could capture and represent the most relevant features of both benign and malignant lesions. The basic concept behind the training of such a model is a step-by-step procedure that gradually corrupts an image with noise through Markov chains, until the image distribution is equivalent to an isotropic Gaussian. Subsequently, a denoising procedure is executed with the aim of reverting the image back to its original appearance. The purpose is to gather information about the original image distribution, and leverage such knowledge to try and generate artificial samples that closely match original distribution.

To achieve this goal, DDPMs leverage a U-Net architecture [81] enhanced with skip connections and self-attention, so that the model is able to handle different conditioning information for image generation. In fact, Diffusion models can also create samples from a text prompt, or a text-and-image input combination. For this work, no text prompt was used and the Diffusion model only relied on images as input for training.

In order to generate both benign and malignant images, two DDPMs were trained using either all of the benign or the malignant samples of the original training set (200 epochs and a learning rate of $1e^{-4}$ for both models). At the end of the training phase, each model was used to generate 2000 samples for each class, leading to a total of 4000 synthetic images.

5.4 Experimental setup

To quantify the impact of the used DA techniques and assess their validity with respect to the state-of-the-art, four different experiments were conducted:

- a Vision Transformer model pretrained on the ImageNet database [62] was fine-tuned on the CBIS-DDSM dataset for 100 epochs, and its performance was set as a baseline for the next experiments;
- the very same procedure was executed again for 50 epochs, using the geometric DA method for the training set;
- the ImageNet-pretrained ViT model was fine-tuned on CBIS-DDSM for 50 epochs leveraging diffuser-generated images to enrich the original training set;
- fine-tuning of the ImageNet-pretrained ViT was executed once again for 50 epochs, this time using a training dataset containing original, geometric-augmented and diffuser-augmented images.

The model version used for this work is ViT-base, with the structure and hyperparameters listed in Table 5.1.

5.5 Results

Experimental trials aimed, first of all, to evaluate the impact of different DA techniques, and, successively to compare the obtained results with the state-of-the-art. Concerning the DA impact, confusion matrices for the CBIS-DDSM test set are displayed in Figure 5.2. It is possible to notice:

Table 5.1 ViT-base model hyperparameters.

Hyperparameter	**Value**
Number of layers	12
Number of heads	12
Hidden size	768
Input image size	224×224
Patch size	16×16
Optimizer	Adam with β_1=0.99, β_2=0.999 and $\varepsilon = 1.0e^{-8}$
Learning rate	$1.0e^{-5}$
Learning rate scheduler	Linear with a warm-up ratio of 0.1
Batch size	16
Gradient accumulation steps	4

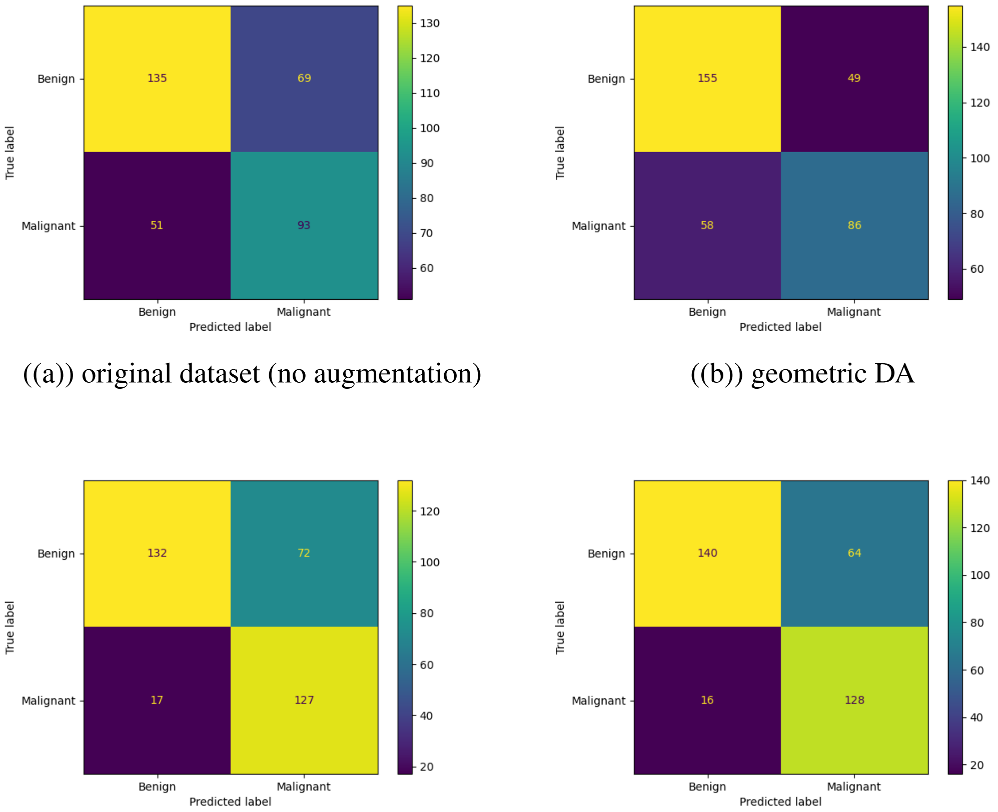
- an increased capability to detect benign images (with a slight decrease for the malignant class) when applying the geometric augmentation approach;
- a different trend with the Diffuser-based technique, which appears to boost the overall accuracy of the model;
- this tendency is slightly more evident when examining the combination of the two augmentation approaches.

The above cited aspects are reflected in the results reported in Table 5.2, in which the combination of geometric and diffuser-based augmented images leads to an increase in accuracy, sensitivity and specificity.

In comparison with the state-of-the-art, reported in Table 5.3, it is worth highlighting that the proposed approach, leveraging classic together with Diffuser-based DA, makes it possible to enhance classification accuracy, thanks to a greater ability of the model to generalize on unknown samples.

Figure 5.3 shows the training and validation losses for all experiments. It can be noted that the gap between the two curves tends to disappear when diffuser-generated images are inserted into the training set. The addition of said images to the geometric-generated samples shows a slight smoothing effect on the plots.

Finally, Figure 5.4 shows an example of the ViT attention maps on a malignant lesion image. Such maps are the visual representation of the attention scores computed by the self-attention mechanism within the Transformer architecture, showing how



((a)) original dataset (no augmentation)

((b)) geometric DA

((c)) diffuser-based DA

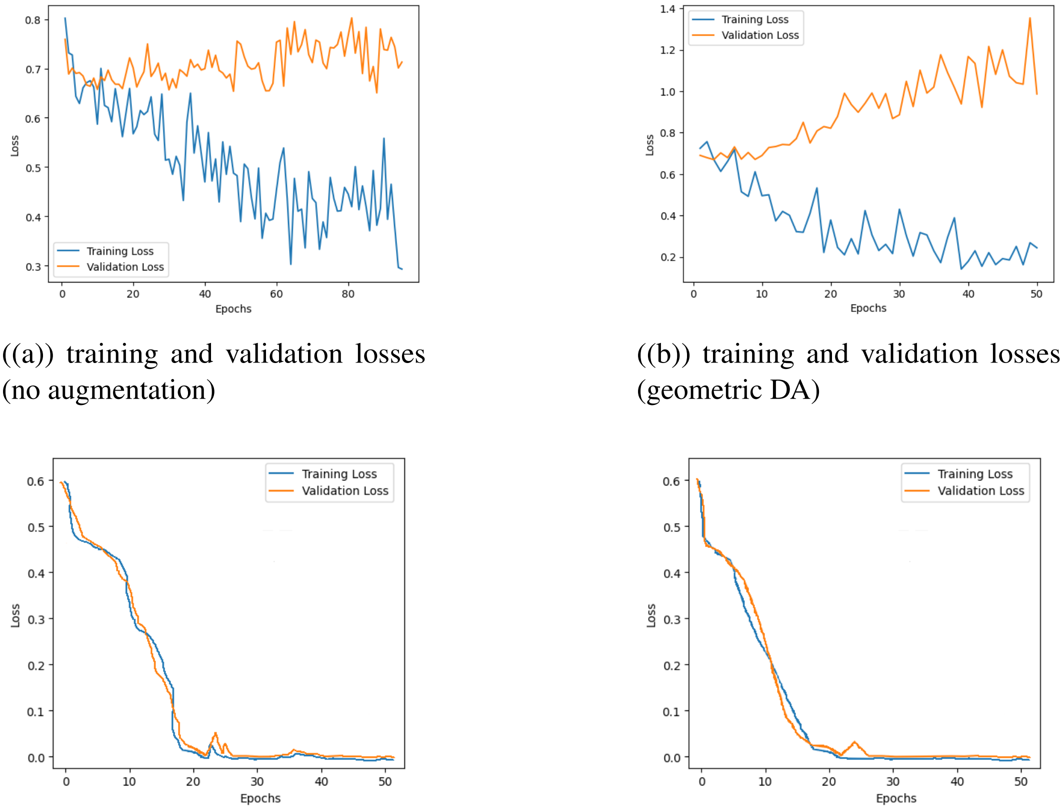
((d)) geometric + diffuser-based DA

Figure 5.2 Test confusion matrices for the different data augmentation techniques

Table 5.2 Classification results on the test set with different data augmentation techniques.

Approach	**Accuracy**	**Sensitivity**	**Specificity**
no augmentation	65.52%	66.18%	64.58%
geometric DA	69.25%	75.98%	59.72%
diffuser-based DA	74.42%	88.19%	64.70%
geometric + diffuser-based DA	**77.01%**	**88.89%**	**68.63%**

much information from one token is used to influence the representation of another. This visualization provides a model interpretation, very useful for physicians and clinical stakeholders. Indeed, the availability of such information - rarely found on deep learning models - is a remarkable point in favor of the proposed approach. This is particularly true due to the intrinsically critical nature of clinical diagnosis



((a)) training and validation losses (no augmentation)

((b)) training and validation losses (geometric DA)

((c)) training and validation losses (diffuser-based DA)

((d)) training and validation losses (geometric + diffuser-based DA)

Figure 5.3 Training and validation losses for the different data augmentation methods

Table 5.3 Accuracy comparison with literature approaches on CBIS-DDSM dataset.

Approach	**Accuracy**
deep multiple instance learning [82]	74.2%
max pooling [83]	67.5%
average pooling [83]	70.3%
region-based group-max pooling [83]	76.2%
global group-max pooling [83]	76.7%
proposed	**77.01%**

and decision making, for which the proposed model offers a valid and explainable support.

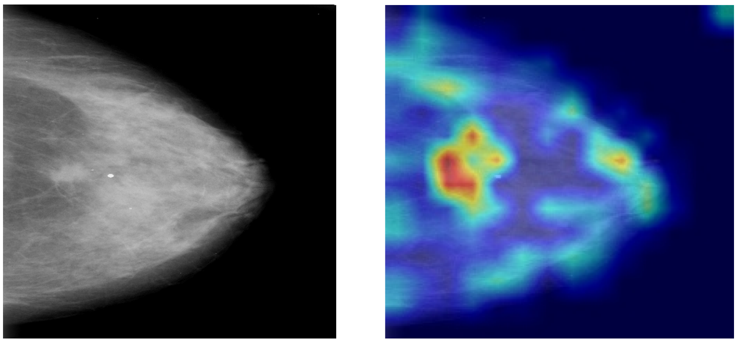


Figure 5.4 Example of attention map on a malignant lesion image

5.6 Discussion

This application example presents an effective ViT-based method to recognize the presence of a tumour in mammographic images. The dataset used is CBIS-DDMS, known in the literature. However, this dataset has limited images, some of which are noisy. To overcome this difficulty, this work implemented a model that combines data augmentation with the interpretation potential provided by ViT to improve classification. Furthermore, the proposed approach can be easily extended to other challenging datasets, especially biomedical data, containing highly heterogeneous information. The proposed augmentation method is based on a geometric data augmentation followed by a DA implemented with a Diffuser architecture. The results show how geometric data augmentation, preserving the original distribution of the data, surpasses the results without augmentation. Furthermore, the results are further improved using a Diffuser trained on the geometric augmented dataset.

Chapter 6

Attention Interpretation and Explainability

The crucial role of attention in Transformer-based architectures is undoubtedly the key factor behind the success of such models across different domains. In this section, the most relevant methods developed to probe and interpret attention in Vision Transformers will be discussed, with the purpose of investigating the reasons behind the capabilities of the architecture that make it such an effective and versatile tool.

In particular, since ViT's main strength is the ability to gather knowledge about the global picture context, the work will try and explore the way the model learns locality - which is equally as important in any computer vision application. Furthermore, similarites and differences between Vision Transformers and CNNs will be analyzed, in order to gain a better understanding of the models' internal representations of images.

6.1 Mean Attention Distance (MAD)

Introduced by Dosovitskiy *et al.* in the seminal Vision Transformer paper, Mean Attention Distance is defined as the geometric distance between two patches scaled by the attention values. This implies that a large MAD value indicates that distant patches get high attention values, whereas a small MAD score denotes the opposite, e.g. high attention values are associated to closely located patches. This quantity was proposed with the purpose of investigating the changes in attention distribution across the image as a function of the Transformer's head and layers.

The steps to compute MAD are the following (described sizes and dimensions are related to a ViT *Base* configuration with $(224 \times 224 \times 3)$ image resolution and (16×16) patches):

- Distance matrix is computed: its size will be (196×196), where $14^2 = 196$ is the total number of (16×16) patches. Each position (i, j) in the matrix is an expression of the distance of patch j with respect to patch i, computed as L2-norm (or Frobenius norm);
- Attention weights from each transformer block are collected (e.g.: for 12 transformer blocks there will be 12 attention weight tensors). Overall size is 12 blocks $\times 196 \times 196 = (12 \times 196 \times 196)$;
- The full attention weight tensor is multiplied by the distance matrix. The final result has a shape of $(12 \times 196 \times 196)$;
- All elements in each row are summed together to get the average distance per patch. Final result has a shape of (12×196);
- All elements in each columns are averaged to get the average attention distance across all 196 patches. Final result has the same size as the number of attention heads (e.g.: 12);
- The procedure is repeated for each head and for all blocks. When the process is complete, each head will have a specific MAD score.

MAD scores for each head will assume values lying in the range $(0, L)$, where L can be evaluated by taking into account the following observations:

- The average distance between two random patches in a square grid with side length K can be expressed as:

$$A = K \cdot \frac{2 + \sqrt{2} + 5\ln(1 + \sqrt{2})}{15} \approx K \cdot 0.521405433 \tag{6.1}$$

- The average pixel distance of a pixel inside a square with side length H from any of the vertices can be expressed as:

$$B = H \cdot \frac{1}{3}[\sqrt{2} + \ln(\sqrt{2}+1)] \approx H \cdot 0.765195716 \tag{6.2}$$

- Quantities A and B are two expected values that can be assumed to represent two statistically independent distance distributions. Thus, superposition can be applied and, for each query pixel selected for MAD evaluation, it can be stated that the maximum value for MAD can be computed as:

$$L = A + B \approx K \cdot 0.512405433 + H \cdot 0.765195716 \tag{6.3}$$

  where K is ViT the image resolution size (e.g.: 224), and H is the patch side length (e.g.: 16).

For instance, when considering the ViT Base architecture, with an image resolution of ($224 \times 224 \times 3$) and a patch size of ($16\times$) the maximum expected value for the Mean Attention Distance will be approximately equal to:

$$\begin{aligned} L &\approx K \cdot 0.512405433 + H \cdot 0.765195716 = \\ &= 224 \cdot 0.512405433 + 16 \cdot 0.765195716 \approx 128 \end{aligned} \tag{6.4}$$

It can be observed that MAD for lower layers can either assume small or large values, depending on the head. This can be interpreted as the model having a wider spatial span of attention, either at a local and global level. As the number of layers increase, MAD ranges become progressively narrower for all heads, indicating that the network pays more attention to the picture in its entirety. Of note, MAD behaviour does not seem to change significantly for the case of hybrid CNN/ViT architectures, in which intermediate features extracted from CNN like ResNet50 where fed as input to the Vision Transformer.

The relationship between Mean Attention Distance, model depth and pretraining data was explored by Raghu et al. [84].

ViT models were either:

- pretrained on the JFT-300M dataset and fine-tuned on ImageNet;

- pretrained and fine-tuned on ImageNet
- pretrained and fine-tuned on JFT-300M [85]

It is possible to notice that MAD scores become lower for the initial heads in the case of a ViT-L/16 pretrained and fine-tuned on ImageNet, indicating that the attention tends to be more widespread from the early stages of the Transformer, as opposed to what previously described. However, this situation no longer occurs for ViT-B/32, which is a smaller and shallower version of the very same architecture. Those observations suggest that when pretraining data is insufficient (as for the case of pretraining and fine-tuning on ImageNet), ViT models are not able to properly learn locality. The deeper the model, the more evident the effect of such event.

Indeed, Mean Attention Distance is displayed in Figure 6.1 for the case of the Vision Transformer model proposed for the ISIC2017 skin lesion classification task, as previously described in Chapter 4. It is possible to interpret the diagram as follows: shallower layers appear to have a narrower attention window, and seem to focus more closely on small subsections and local details of each image. Conversely, deeper layers are able to cast their attention over a wider range, connecting sub-regions in each figure to try and grasp the underlying relationships in order to get the *full picture*. As a last remark, it is possible to observe that MAD scores for all heads lie approximately in the range $(0, L) = (0, 128)$, as previously evaluated in equation 6.4.

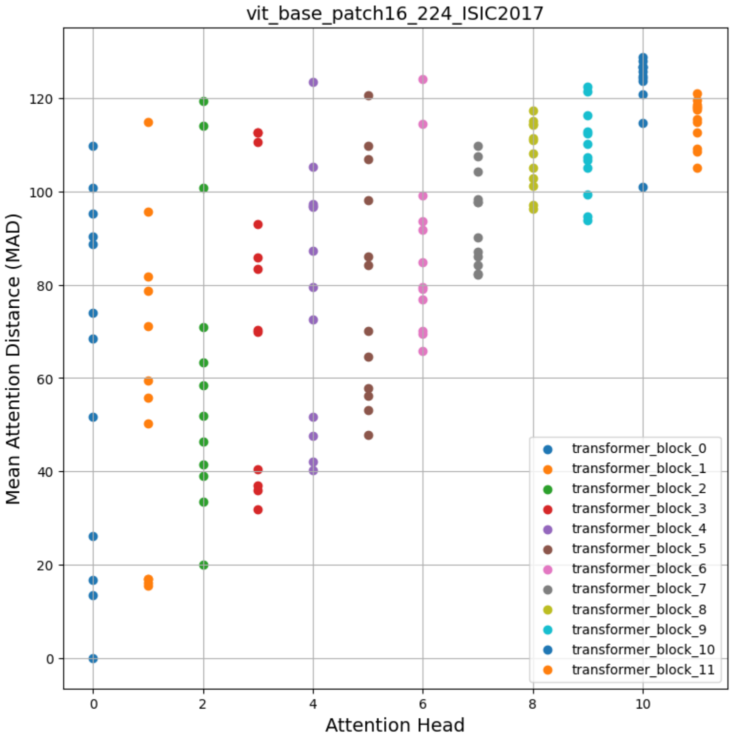


Figure 6.1 Mean Attention Distance (MAD) for the proposed skin lesion classification model

6.2 Image Representations: Differences And Similarities Between ViTs and CNNs

In their paper, Kornblith et al. [86] introduced Centered Kernel Alignment (CKA) as a quantifiable way to compare neural network image representations. CKA is defined as:

$$CKA(\mathbf{K},\mathbf{L}) = \frac{HSIC(\mathbf{K},\mathbf{L})}{\sqrt{HSIC(\mathbf{K},\mathbf{K})HSIC(\mathbf{L},\mathbf{L})}} \tag{6.5}$$

where **K** and **L** are Gram matrices, calculated as:

$$\mathbf{K} = XX^T \tag{6.6}$$

$$\mathbf{L} = YY^T \tag{6.7}$$

X and Y are the image representations from different neural models. HSIC indicates the Hilbert-Schmidt independence criterion, according to which X and Y are independent if, and only if, the Hilbert-Schmidt norm, defined as:

$$||C_{xy}||_{\mathscr{F}}{}^2 = \sum_i \lambda_i^2 = tr[C_{xy}^T C_{xy}] \tag{6.8}$$

is equal to zero (the term C_{xy} is the covariance matrix).

HSIC is defined as:

$$HSIC(\mathbf{K}, \mathbf{L}) = vec(\mathbf{K'}) \cdot vec(\mathbf{L'})/(m-1)^2 \tag{6.9}$$

where **K'=HKH** and **L'=HLH** are the centered Gram matrices, and $\mathbf{H} = \mathbf{I}_n - 1/n\mathbf{1}\mathbf{1}^T$ is the centering matrix.

CKA exhibits two interesting properties:

- invariance to orthogonal representations or transformations, meaning that any permutation of nodes within the network will not affect CKA value;
- invariance to isotropic scaling, which means that it does not change when image dimensions are scaled.

CKA was evaluated comparing either ViT-L/16 to ViT-H/14 as well as ResNet50 to ResNet152. Vision Transformer layers show more uniform similarities among each other. On the other hand, ResNet layers similarities appear to behave differently according to the network depth. In other words, uniform similarities present among lower layers appear to be different from uniform similarities appearing among the upper layers.

Indeed, a comparison between ViT-L/16 and ResNet50, and ViT-H/14 and ResNet50, shows that ViTs compute similar features as ResNets, but with a smaller set of layers. Of note, Vision Transformer features seem to propagate in a more faithful manner with respect to ResNet.

When computed for the model proposed in Chapter 3, the plot for Centered Kernel Alignment appears as shown in Figure 6.2

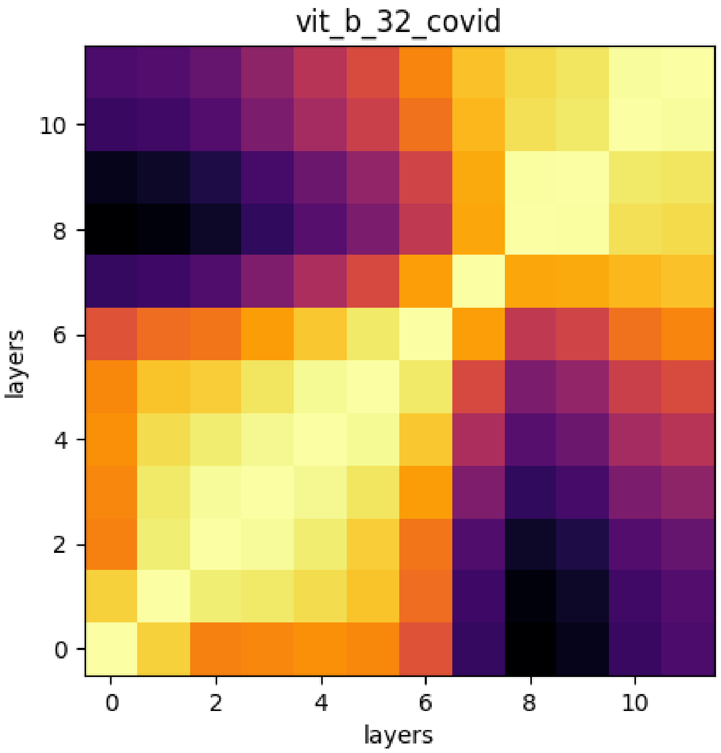


Figure 6.2 Centered Kernel Alignment (CKA) for the proposed chest X-ray classification model

It can be noted that image representations across network layers exhibit a tight similarity throughout the entire architecture, highlighting the capability to propagate features in a uniform manner. In general, this effect is even more evident on deeper versions of the Vision Transformer.

6.3 Skip Connections

The influence of skip connections in Vision Transformer architectures has an impact on the uniformity of representations discussed in the previous section. This was shown again by Raghu et al. by defining the impact of skip connections as:

$$\|z_i\|/\|f(z_i)\| \tag{6.10}$$

where z_i is the vector of hidden representations for the ith layer (e.g., the skip connection itself), and $f(z_i)$ is the transformation of z_i when it goes through a ViT block - which will be either a self-attention block or a MultiLayer Perceptron.

It was observed that there's a pretty clear transition between spatial tokens and CLS token around the sixth block. Skip connections seem to propagate the CLS token at first, carrying spatial tokens from the sixth block onwards.

When it comes to ResNets, norm ratios are generally low. On the other hand, they show significantly high values for the CLS token in the lower layers, and for self-attention in the higher layers. In addition, it was shown that removing skip connections has an impact on the uniformity of representations.

6.4 Explainability

Model explainibilty was approached before the development of Transformers, with the purpose of interpreting and understanding model predictions. Reaching such goal is usually increasingly complex as models grow deeper, since they tend to contain more and more parameters.

The most common tools developed for explainability can roughly be divided into model specific, like Grad-CAM [87], and model agnostic, such as SHAP [88] or LIME [89]. In the case of Transformers, the first concepts that were exploited by researchers were based on analyzing the attention values that correspond to the CLS token and using such quantities as an explanation for prediction. However, this approach raises two main issues:

- aggregation across heads: given the amount of variability of attention among the Transformer heads, it becomes necessary to define a proper way to average attention across all heads;
- aggregation across layers: attention scores for each patch at higher layers is a mixture of attention from the previous layers, so there's a need to account for this effect.

An approach that attempts to solve the above problems, called *Attention Rollout*, was proposed by Abnar et al. [90]. Such method suggests a simple average of attention for all heads to tackle the first issue, and a matrix multiplication of attention maps to track context (see Figure 6.3).

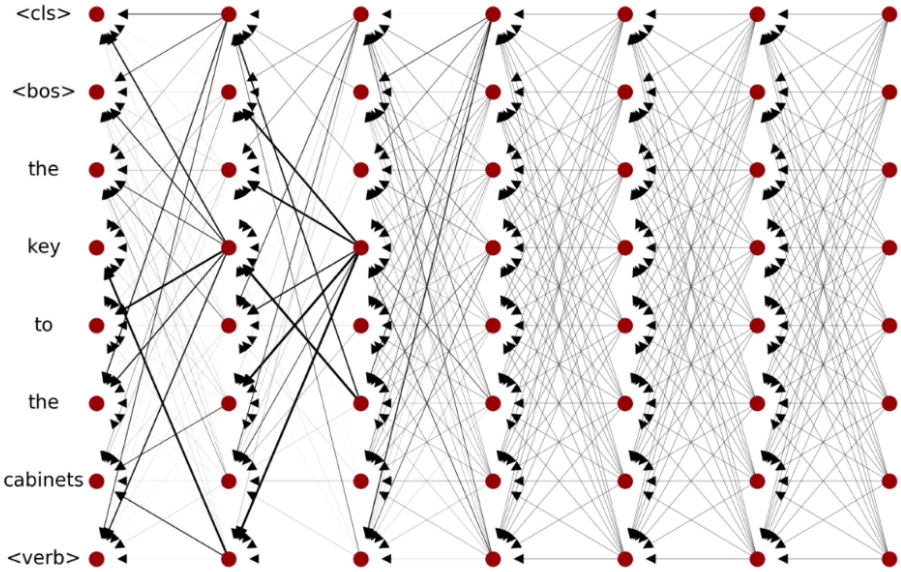


Figure 6.3 Visualization of Attention Rollout on a textual example

This means that attention matrices for all layers are multiplied, adding the identity matrix to each self-attention matrix to account for skip connection, as expressed by the following equation:

$$A = 0.5W_att + 0.5I \tag{6.11}$$

An application of this method on the model proposed in Chapter 5 for mammographic image classification is displayed in Figure 6.4, in which attention is *rolled* from the network output to the input - in this case, a malignant lesion mammogram. This technique allows to effectively represent information propagation through the network layers.

An alternative technique was proposed on the same paper by solving a max-flow problem on the attention graph, the so-called *Attention Flow*. This approach is more computationally expensive: $O(d^2 * n^4)$, where d is the network depth and n is the number of patches in the image. Average attention across heads is still computed as a basic average, just as Attention Rollout. However, this method does not take into account the possible differences in relevance among heads with respect to the

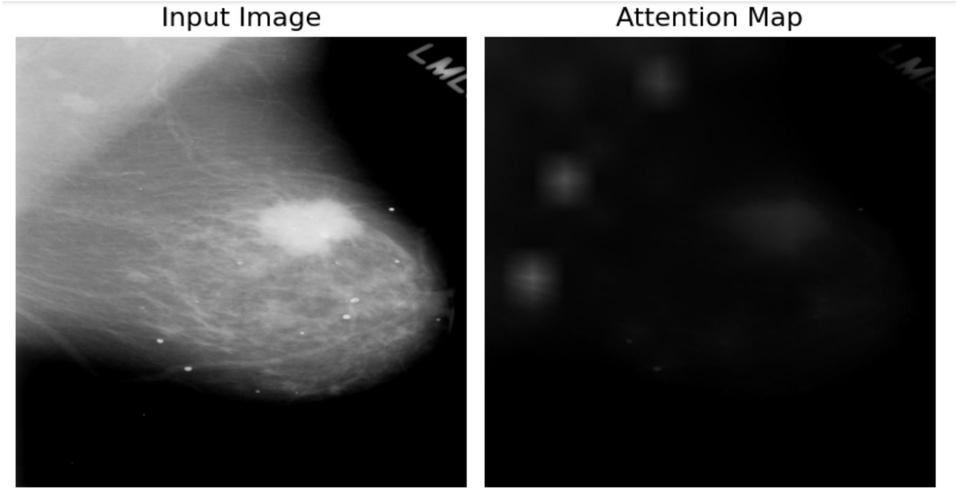


Figure 6.4 Attention Rollout on a malignant lesion mammographic image from the CBIS-DDSM dataset

prediction outcome of the model. Furthermore, the contribution of other network elements other than attention blocks is completely neglected.

To overcome those problems, Chefer et al. proposed TiBA [91], which is based on the concept of averaging attention gradients across heads (see Figure 6.5).

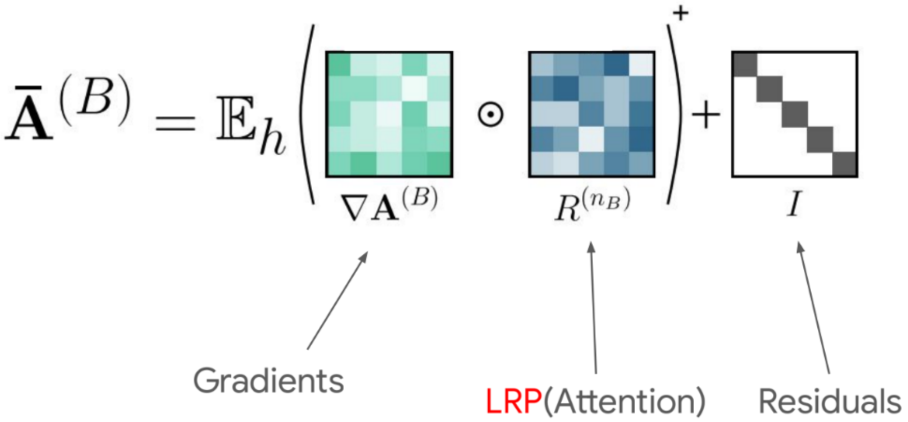

$\bar{\mathbf{A}}^{(B)} = \mathbb{E}_h\left(\nabla\mathbf{A}^{(B)} \odot R^{(n_B)}\right)^+ + I$
Gradients
LRP(Attention)
Residuals

Figure 6.5 Depiction of the TiBA Attention Computation Concept

Specifically, attention gradients are scaled using Layer-Wise Relevance Propagation [92] as a relevance matrix, to account for other layers. An identity matrix is added to include the effect of skip connections. Eventually, layers are aggregated via matrix multiplications, similar to Attention Rollout. Leveraging gradients allows to show class specific attention heatmaps.

The path for class specific attention was also explored by Touvron et al. [93], who showed that the attention layers can be separated into clusters dedicated to image patches only, and to the interaction between patches and the CLS token. This was implemented by introducing the CLS token into the network from a certain layer onwards, and freezing the patch embeddings obtained from the previous layers (Figure 6.6).

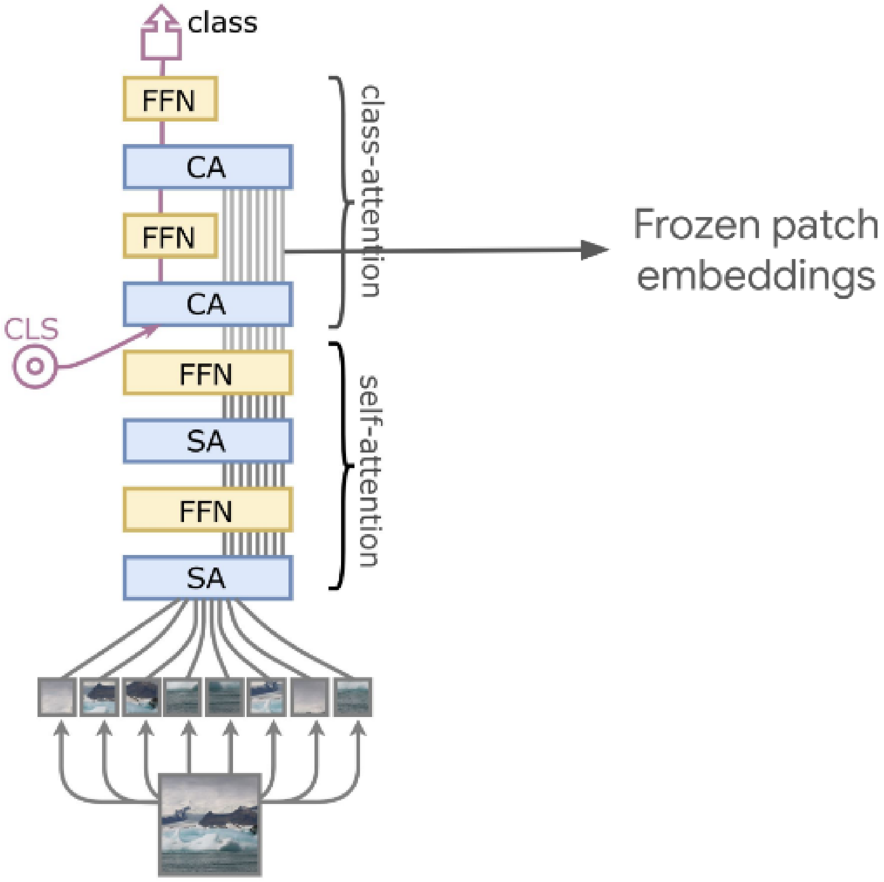


Figure 6.6 Example of modified ViT architecture to obtain class specific attention

The resulting attention maps seem to contain semantic information associated to different elements of the image, as long as the pretraining phase was unsupervised [94]. Conversely, semantic layouts become sparse when pretraining is supervised.

Other relevant aspects that can be inspected through visualization are related to early processing steps at the beginning of the Vision Transformer architecture: patch embeddings and positional encoding. Such operations encrypt image elements into specific sub-elements that retain quantums of information, either related to picture features, rather than their reciprocal spatial relationships. Figure 6.7 shows the first 128 learned projection embeddings for a COVID-19 chest x-ray image input to the model presented in Chapter 3.

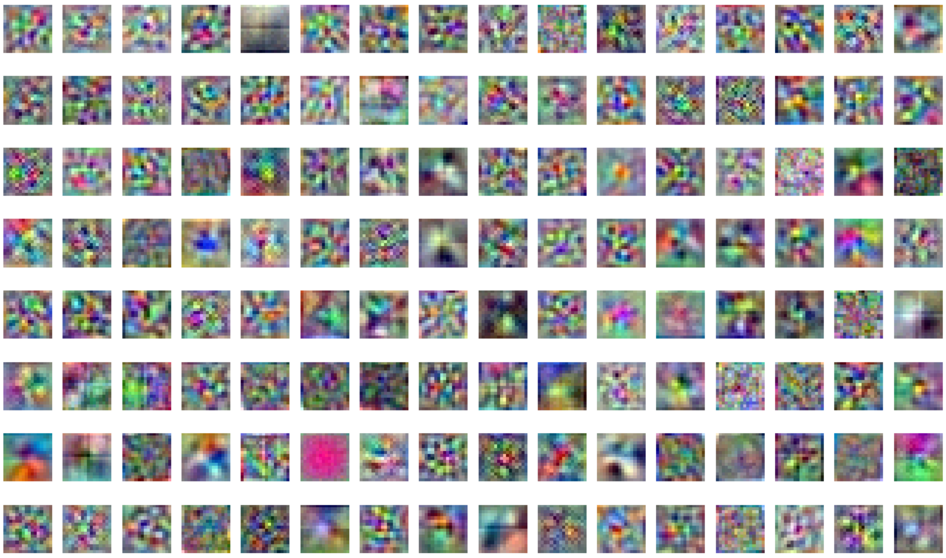


Figure 6.7 Visual representation of the first 128 learned projection embeddings for a COVID-19 chest x-ray image processed by the proposed Vision Transformer model for lung disease classification

It is possible to notice a remarkably striking resemblance to kernels trying to detect peculiar features, such as edges or shapes, in Convolutional Neural Networks. Thus, the early processing stage of Vision Transformer models can be interpreted as a sub-network that tries to capture low-level features, which will be fed to the next layers in order to integrate them into a meaningful context.

Indeed, a mere collection of features is seldom sufficient to properly interpret the meaning of a picture. In fact, gathering knowledge on how such features interact with each other can successfully be leveraged to reach a higher level of understanding of visual information. To do so, Vision Transformers apply a learnable positional encoding to each embedded patch, as expressed by the following equations:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{6.12}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{6.13}$$

A visualization of positional encoding for the model proposed in Chapter 5 is shown in Figure 6.8, which depicts similarities among embedded patches once positional encoding has been added.

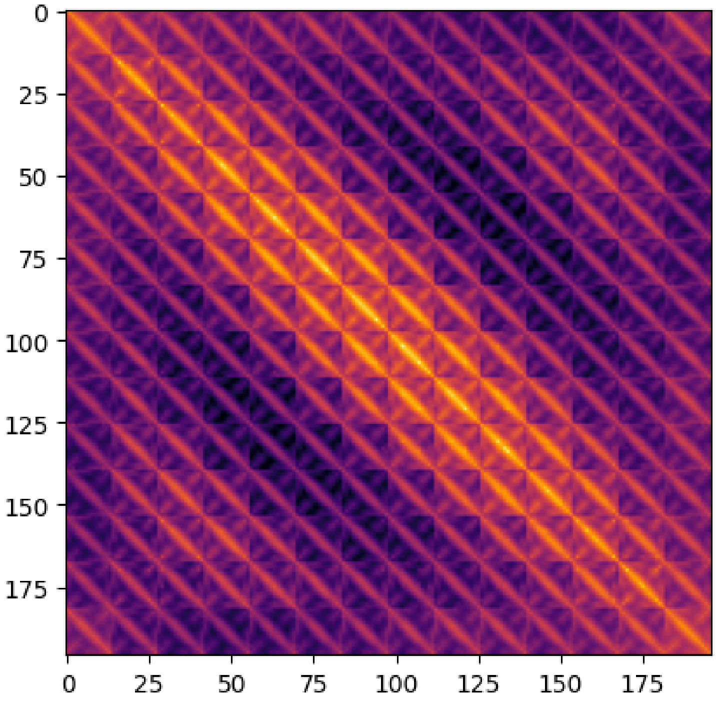


Figure 6.8 Visual representation of positional encoding for the proposed Vision Transformer model for mammographic image classification

The picture shows a bright main diagonal, which trivially indicates that the highest similarity occurs between an embedded patch and itself. However, other dimmer diagonal patterns, parallel to the main one, can be observed, exhibiting an overall pattern that is indeed resembling a sinusoid. All the experiments presented over the case studies and application tasks discussed through the previous chapters undoubtedly prove that the described positional encoding method allows to efficiently capture the spatial relationships among embedded image patches, leading to successful image classification performance over several different domains.

Chapter 7

Conclusion

7.1 Study Limitations

7.1.1 Dataset Reliability

The main limitations of the proposed work are mostly related to the datasets that were used. In fact, it is evident that the effectiveness of the proposed models is only valid with respect to the data samples that were made available to work with. In the end, it all comes down to how representative such samples are of a true distribution in real world cases. Deploying clinical data for public collections and/or challenges, which is supposed to have passed the sieve of expert physicians, should guarantee the validity of the used data across all showed applications. However, stronger data checks and further statistical assessments need to be carried, such as:

1. k-fold cross-validation and Leave-One-Out procedures, in order to exclude any particularly favorable dataset splits;
2. Working together with experts in tight contact and cooperation to further confirm data validity - especially in terms of true distribution representation capabilities;
3. Collecting additional external data, then repeating steps 1 and 2.

7.1.2 Metrics

The choice of performance metrics has an impact on the way results are presented, favouring certain aspects and hindering perspectives on how data can be interpreted.

As an example, a large accuracy score is usually associated to good model performance. However, if a class' cardinality is signficantly higher with respect to the others, the model is likely to identify that category with a higher rate of success. In that case, this will usually occur at the expenses of the model's detection capabilities for the other (underrepresented) classes. Nonetheless, for all the applications presented in this document, a closer look at the confusion matrices and the other metrics of choice (most notably, sensitivity and specificity) clarifies that the proposed models are able to detect all image categories with very large rates of success.

When dealing with clinical image classification applications, the capability of a model to correctly identify all images from the most relevant category (e.g., melanoma) might not be sufficient for the algorithm to be considered efficient and reliable. Indeed, mistakenly classifying a benign lesion as malignant can have equally serious implications as misclassifying a malignant one.

7.2 Current Projects and Future Perspectives

Other relevant projects being currently carried out with similar methodology are:

- The attempt to identify risk factors for rare heart syndromes that can cause sudden cardiac death (SCD) by analyzing scanned ECG charts. Diseseas include channelopathies like Short QT and Brugada syndromes. The scarcity of available images, combined with extremely poor picture quality and a marked heterogeneity, make this problem particularly challenging. In addition, the lack of well defined features associated to SCD, together with the effort to develop an AI-based tool that could unveil hindered details, contribute to make this work remarkably intriguing. [95]
- The identification of plant diseases through leaf image analysis. A success in this kind of application would imply that the proposed methodology can be extended to the analysis of diseases that are no longer limited to human subjects, contributing to safeguard plant species too.

Future perspectives will focus on extensively probing and inspecting attention on the aforementioned projects, and to the above discussed case studies, with the purpose of opening the doors to attention interpretation and model explainability in the clinical and biological images domain.

Bibliography

[1] Lulu Gai, Wei Chen, Rui Gao, Yan-Wei Chen, and Xu Qiao. Using vision transformers in 3-d medical image classifications. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 696–700, 2022.

[2] Faris Almalik, Mohammad Yaqub, and Karthik Nandakumar. Self-ensembling vision transformer (sevit) for robust medical image classification, 2022.

[3] Sangjoon Park, Gwanghyun Kim, Jeongsol Kim, Boah Kim, and Jong Chul Ye. Federated split vision transformer for covid-19 cxr diagnosis using task-agnostic training, 2021.

[4] Sudhakar Tummala, Seifedine Kadry, Syed Ahmad Chan Bukhari, and Hafiz Tayyab Rauf. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, 29(10):7498–7511, October 2022.

[5] Yin Dai, Yifan Gao, and Fayu Liu. TransMed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, July 2021.

[6] Yingda Xia, Jiawen Yao, Le Lu, Lingyun Huang, Guotong Xie, Jing Xiao, Alan Yuille, Kai Cao, and Ling Zhang. Effective pancreatic cancer screening on non-contrast CT scans via anatomy-aware transformers. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 259–269. Springer International Publishing, 2021.

[7] Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20686–20697, 2022.

[8] Kobiljon Ikromjanov, Subrata Bhattacharjee, Yeong-Byn Hwang, Rashadul Islam Sumon, Hee-Cheol Kim, and Heung-Kook Choi. Whole slide image analysis and detection of prostate cancer using vision transformers. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 399–402, 2022.

[9] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *2021*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10933–10942, 2021.

[10] Haoyuan Chen, Chen Li, Ge Wang, Xiaoyan Li, Md Mamunur Rahaman, Hongzan Sun, Weiming Hu, Yixin Li, Wanli Liu, Changhao Sun, Shiliang Ai, and Marcin Grzegorzek. GasHis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognition*, 130:108827, October 2022.

[11] Yi Zheng, Rushin H. Gindra, Emily J. Green, Eric J. Burks, Margrit Betke, Jennifer E. Beane, and Vijaya B. Kolachalama. A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41(11):3003–3015, 2022.

[12] Sheng Wang, Zixu Zhuang, Kai Xuan, Dahong Qian, Zhong Xue, Jia Xu, Ying Liu, Yiming Chai, Lichi Zhang, Qian Wang, and Dinggang Shen. 3dmet: 3d medical image transformer for knee cartilage defect assessment. In *Machine Learning in Medical Imaging*, pages 347–355. Springer International Publishing, 2021.

[13] Richard J. Chen and Rahul G. Krishnan. Self-supervised vision transformers learn visual concepts in histopathology, 2022.

[14] Jianfeng Zhao, Xiaojiao Xiao, Dengwang Li, Jaron Chong, Zahra Kassam, Bo Chen, and Shuo Li. mfTrans-net: Quantitative measurement of hepatocellular carcinoma via multi-function transformer regression network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 75–84. Springer International Publishing, 2021.

[15] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-ViT. *Multimedia Tools and Applications*, 81(17):24265–24300, March 2022.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[18] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023.

[20] Sergio Cannata, Annunziata Paviglianiti, Eros Pasero, Giansalvo Cirrincione, and Maurizio Cirrincione. Deep learning algorithms for automatic covid-19 detection on chest x-ray images. *IEEE Access*, 10:119905–119913, 2022.

[21] Elliot J Lefkowitz, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, and Donald B Smith. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, 46(D1):D708–D717, 10 2017.

[22] Coronavirus disease (COVID-19). Coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019// (visited on 02 July 2022).

[23] Covid-19 PCR test: how does it work? are there any alternatives? | Auxologico. Covid-19 PCR test: how does it work? are there any alternatives? | Auxologico. https://www.auxologico.com/covid-19-pcr-test-how-does-it-work-are-there-any-alternatives// (visited on 29 October 2021).

[24] Asmaa Abbas, Mohammed M. Abdelsamea, and Mohamed Medhat Gaber. Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network, 2020.

[25] Aditya Borakati, Adrian Perera, James Johnson, and Tara Sood. Diagnostic accuracy of x-ray versus ct in covid-19: a propensity-matched database study. *BMJ Open*, 10(11), 2020.

[26] Cornelia Schaefer-Prokop and Mathias Prokop. Chest radiography in COVID-19: No role in asymptomatic and oligosymptomatic disease. *Radiology*, 298(3):E156–E157, March 2021.

[27] Di Dong, Zhenchao Tang, Shuo Wang, Hui Hui, Lixin Gong, Yao Lu, Zhong Xue, Hongen Liao, Fang Chen, Fan Yang, Ronghua Jin, Kun Wang, Zhenyu Liu, Jingwei Wei, Wei Mu, Hui Zhang, Jingying Jiang, Jie Tian, and Hongjun Li. The role of imaging in the detection and management of COVID-19: A review. *IEEE Reviews in Biomedical Engineering*, 14:16–29, 2021.

[28] Aras M. Ismael and Abdulkadir Şengür. Deep learning approaches for COVID-19 detection based on chest x-ray images. *Expert Systems with Applications*, 164:114054, February 2021.

[29] Iason Katsamenis, Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, and Nikolaos Doulamis. Transfer learning for covid-19 pneumonia detection and classification in chest x-ray images. In *Proceedings of the 24th Pan-Hellenic Conference on Informatics*, PCI '20, page 170–174, New York, NY, USA, 2021. Association for Computing Machinery.

[30] Anis Shazia, Tan Zi Xuan, Joon Huang Chuah, Juliana Usman, Pengjiang Qian, and Khin Wee Lai. A comparative study of multiple neural network for detection of COVID-19 on chest x-ray. *EURASIP Journal on Advances in Signal Processing*, 2021(1), July 2021.

[31] S. Vineth Ligi, Soumya Snigdha Kundu, R. Kumar, R. Narayanamoorthi, Khin Wee Lai, and Samiappan Dhanalakshmi. Radiological analysis of COVID-19 using computational intelligence: A broad gauge study. *Journal of Healthcare Engineering*, 2022:1–25, February 2022.

[32] Woan Ching Serena Low, Joon Huang Chuah, Clarence Augustine T. H. Tee, Shazia Anis, Muhammad Ali Shoaib, Amir Faisal, Azira Khalil, and Khin Wee Lai. An overview of deep learning techniques on chest x-ray and CT scan identification of COVID-19. *Computational and Mathematical Methods in Medicine*, 2021:1–17, June 2021.

[33] Koushik Sivarama Krishnan and Karthik Sivarama Krishnan. Vision transformer based COVID-19 detection using chest x-rays. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, oct 2021.

[34] Debaditya Shome, T. Kar, Sachi Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Saudagar. COVID-transformer: Interpretable COVID-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21):11086, October 2021.

[35] Arnab Kumar Mondal, Arnab Bhattacharjee, Parag Singla, and A. P. Prathosh. xViTCOS: Explainable vision transformer based COVID-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–10, 2022.

[36] Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Multi-task vision transformer using low-level chest x-ray feature corpus for COVID-19 diagnosis and severity quantification. *Medical Image Analysis*, 75:102299, January 2022.

[37] Yassir Edrees Almalki, Abdul Qayyum, Muhammad Irfan, Noman Haider, Adam Glowacz, Fahad Mohammed Alshehri, Sharifa K. Alduraibi, Khalaf Alshamrani, Mohammad Abd Alkhalik Basha, Alaa Alduraibi, M. K. Saeed, and Saifur Rahman. A novel method for COVID-19 diagnosis using artificial intelligence in chest x-ray images. *Healthcare*, 9(5):522, April 2021.

[38] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.

[39] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M. Zughaier, Muhammad Salman Khan, and Muhammad E.H.

Chowdhury. Exploring the effect of image enhancement techniques on COVID-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132:104319, May 2021.

[40] A. m. freeman and j. townes r. leigh, “viral pneumonia”, encycl. respir. med. four-volume set, pp. 456–466, jul. 2021.

[41] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[43] François Chollet. Xception: Deep learning with depthwise separable convolutions, 2017.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[45] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[47] Don P. Mitchell and Arun N. Netravali. Reconstruction filters in computer-graphics. *SIGGRAPH Comput. Graph.*, 22(4):221–228, jun 1988.

[48] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers, 2022.

[49] Giansalvo Cirrincione, Sergio Cannata, Giovanni Cicceri, Francesco Prinzi, Tiziana Currieri, Marta Lovino, Carmelo Militello, Eros Pasero, and Salvatore Vitabile. Transformer-based approach to melanoma detection. *Sensors*, 23(12), 2023.

[50] Wan Hu, Lanlan Fang, Ruyu Ni, Hengchuan Zhang, and Guixia Pan. Changing trends in the disease burden of non-melanoma skin cancer globally from 1990 to 2019 and its predicted level in 25 years. *BMC cancer*, 22(1):836, 2022.

[51] Piyu Parth Naik. Cutaneous malignant melanoma: A review of early diagnosis and management. *World Journal of Oncology*, 12(1):7, 2021.

[52] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine*, 215:106620, 2022.

[53] Tanja B Jutzi, Eva I Krieghoff-Henning, Tim Holland-Letz, Jochen Sven Utikal, Axel Hauschild, Dirk Schadendorf, Wiebke Sondermann, Stefan Fröhling, Achim Hekler, Max Schmitt, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Frontiers in medicine*, 7:233, 2020.

[54] N Vikranth Kumar, P Vijeeth Kumar, K Pramodh, and Yepuganti Karuna. Classification of skin diseases using image processing and svm. In *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, pages 1–5. IEEE, 2019.

[55] Anurag Kumar Verma, Saurabh Pal, and Surjeet Kumar. Classification of skin disease using ensemble data mining techniques. *Asian Pacific journal of cancer prevention: APJCP*, 20(6):1887, 2019.

[56] Junsong Xie, Zezhi Wu, Renju Zhu, and Hong Zhu. Melanoma detection based on swin transformer and simam. In *2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 5, pages 1517–1521. IEEE, 2021.

[57] Dipu Chandra Malo, Md Mustafizur Rahman, Jahin Mahbub, and Mohammad Monirujjaman Khan. Skin cancer detection using convolutional neural network. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0169–0176. IEEE, 2022.

[58] Ghanta Sai Krishna, Kundrapu Supriya, Meetiksha Sorgile, et al. Lesionaid: Vision transformers-based skin lesion generation and classification. *arXiv preprint arXiv:2302.01104*, 2023.

[59] The International Skin Imaging Collaboration. The International Skin Imaging Collaboration. https://www.isic-archive.com// (visited on 11 May 2023).

[60] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, 2018.

[61] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[62] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[63] Yonis Gulzar and Sumeer Ahmad Khan. Skin lesion segmentation based on vision transformers and convolutional neural networks—a comparative study. *Applied Sciences*, 12(12):5990, 2022.

[64] Xinzi He, Ee-Leng Tan, Hanwen Bi, Xuzhe Zhang, Shijie Zhao, and Baiying Lei. Fully transformer network for skin lesion analysis. *Medical Image Analysis*, 77:102357, 2022.

[65] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 251–266. Springer, 2020.

[66] Wenjun Wu, Sachin Mehta, Shima Nofallah, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9:163526–163541, 2021.

[67] Suliman Aladhadh, Majed Alsanea, Mohammed Aloraini, Taimoor Khan, Shabana Habib, and Muhammad Islam. An effective skin cancer classification mechanism via medical vision transformer. *Sensors*, 22(11):4008, 2022.

[68] Soumyya Kanti Datta, Mohammad Abuzar Shaikh, Sargur N Srihari, and Mingchen Gao. Soft-attention improves skin cancer classification performance. *medRxiv*, 2021.

[69] Bin Zhang, Shenyao Jin, Yili Xia, Yongming Huang, and Zixiang Xiong. Attention mechanism enhanced kernel prediction networks for denoising of burst images. *arXiv preprint arXiv:1910.08313*, 2019.

[70] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[71] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification, 2017.

[72] Francesco Attena, Lucia Abagnale, and Angela Avitabile. Online information about mammography screening in italy from 2014 to 2021. *BMC Women's Health*, 22(1):1–6, 2022.

[73] Ernest Usang Ekpo, Maram Alakhras, and Patrick Brennan. Errors in mammography cannot be solved through technology alone. *Asian Pacific journal of cancer prevention: APJCP*, 19(2):291, 2018.

[74] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.

[75] Arshi Parvaiz, Muhammad Anwaar Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali, and Muhammad Moazam Fraz. Vision transformers in medical computer vision—a contemplative retrospection. *Engineering Applications of Artificial Intelligence*, 122:106126, 2023.

[76] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4:170177, December 2017.

[77] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[78] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[79] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[80] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[81] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[82] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International conference on medical image computing and computer-assisted intervention*, pages 603–611. Springer, 2017.

[83] Xin Shu, Lei Zhang, Zizhou Wang, Qing Lv, and Zhang Yi. Deep neural networks with region-based pooling structures for mammographic image classification. *IEEE transactions on medical imaging*, 39(6):2246–2255, 2020.

[84] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022.

[85] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017.

[86] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019.

[87] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019.

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

[89] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[90] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

[91] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2021.

[92] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016.

[93] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers, 2021.

[94] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[95] Eros Pasero, Fiorenzo Gaita, Vincenzo Randazzo, Pierre Meynet, Sergio Cannata, Philippe Maury, and Carla Giustetto. Artificial intelligence ecg analysis in patients with short qt syndrome to predict life-threatening arrhythmic events. *Sensors*, 23(21), 2023.