Document downloaded from:

http://hdl.handle.net/10251/179752

This paper must be cited as:

Del Amor, R.; Launet, L.; Colomer, A.; Moscardó, A.; Mosquera-Zamudio, A.; Monteagudo, C.; Naranjo Ornedo, V. (2021). An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. Artificial Intelligence in Medicine. 121:1-12. https://doi.org/10.1016/j.artmed.2021.102197

The final publication is available at

https://doi.org/10.1016/j.artmed.2021.102197

Copyright Elsevier

Additional Information

An Attention-based Weakly Supervised framework for Spitzoid Melanocytic Lesion Diagnosis in Whole Slide Images

Rocío del Amor[1], Laëtitia Launet[1], Adrián Colomer[1], Anaïs Moscardó[2], Andrés Mosquera-Zamudio[2], Carlos Monteagudo[2] and Valery Naranjo[1]

[1]Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, 46022, Valencia, Spain

[2]Pathology Department. Hospital Clínico Universitario de Valencia, Universidad de Valencia, Valencia, Spain

Abstract

Melanoma is an aggressive neoplasm responsible for the majority of deaths from skin cancer. Specifically, spitzoid melanocytic tumors are one of the most challenging melanocytic lesions due to their ambiguous morphological features. The gold standard for its diagnosis and prognosis is the analysis of skin biopsies. In this process, dermatopathologists visualize skin histology slides under a microscope, in a highly time-consuming and subjective task. In the last years, computer-aided diagnosis (CAD) systems have emerged as a promising tool that could support pathologists in daily clinical practice. Nevertheless, no automatic CAD systems have yet been proposed for the analysis of spitzoid lesions. Regarding common melanoma, no system allows both the selection of the tumor region and the prediction of the benign or malignant form in the diagnosis. Motivated by this, we propose a novel end-to-end weakly supervised deep learning model, based on inductive transfer learning with an improved convolutional neural network (CNN) to refine the embedding features of the latent space. The framework is composed of a source model in charge of finding the tumor patch-level patterns, and a target model focuses on the specific diagnosis of a biopsy. The latter retrains the backbone of the source model through a multiple instance learning workflow to obtain the biopsy-level scoring. To evaluate the performance of the proposed methods, we performed extensive experiments on a private skin database with spitzoid lesions. Test results achieved an accuracy of 0.9231 and 0.80 for the source and the target models, respectively. In addition, the heat map findings are directly in line with the clinicians' medical decision and even highlight, in some cases, patterns of interest that were overlooked by the pathologist.

Keywords: Spitzoid lesions, Attention convolutional neural network, Inductive transfer learning, Multiple instance learning, Histopathological whole-slide images

1. Introduction

According to the World Health Organization, nearly one in three diagnosed cancers is a skin cancer [1]. The most dangerous skin cancer is melanoma which is responsible for 80 percent of skin cancer-related deaths [2]. Melanoma is an aggressive melanocytic neoplasm with numerous resistance mechanisms against therapeutic agents. In most melanocytic tumors, a precise pathological distinction between benign (nevus) and malignant (melanoma) is possible. However, there are still uncommon melanocytic lesions that represent a diagnostic challenge for pathologists. Among these, one of the most challenging lesions to diagnose is the so-called 'spitzoid melanocytic tumors' (SMTs), composed of spindled and/or epithelioid melanocytes with a large nucleus [3].

The final diagnosis of SMTs is confirmed by skin biopsies. The skin tumor is excised, laminated, stained with Hematoxylin and Eosin (H&E) and finally stored in crystal slides. Then, dermatopathologists analyze the sample under the microscope [3]. During the analysis of spitzoid lesions, different histopathological characteristics can be observed depending on the malignancy degree, see Figure 1. The regions with benign spitzoid lesions generally have a confluence of melanocytes in well-defined and organized nests. Figure 1 (a)-(b) shows sub-regions of a benign spitzoid melanocytic lesion. These regions show cellular and architectural maturation (both melanocytes and nests decrease in size to-

Email address: `madeam2@upvnet.upv.es` ()

wards the base of the lesion) throughout the dermis. In this case, this type of benign lesion is known as compound Spitz nevus. If the lesion only occurs in the epidermis and does not show extension into the dermis it would be called junctional nevus. In the case of spitzoid malignant lesions, cellular disorder is a frequent pattern, the melanocytic nests are ill-defined and are usually devoid of maturation, see Figure 1 (c). Additional features associated with malignancy of spitzoid melanocytic lesions include marked nuclear pleomorphism, pagetoid spread (individual cells or small aggregates of melanocytic cells grow and invade the upper epidermis from below) and a poor circumscription of lesions at their peripheries [4]. Figure 1 (d) shows an example of the pagetoid pattern. In addition to the cellular disorder, there are other local-level features associated with malignancy. Among these patterns, typical (bipolar and symmetrical) and atypical (aberrant mitotic figures, usually asymmetrical and/or multipolar) mitoses stand out. Note that benign melanocytic lesions can also have occasional typical mitoses, particularly in the most superficial areas. Therefore, if we find a typical mitosis in a spitzoid lesion, we should take into account additional factors such as the number of mitoses and their location within the lesion (deep typical mitoses are more suspicious of malignancy than the superficial ones) to determine if the neoplasm is malignant. Typical mitoses are only a sign of cellular proliferation and their mere presence cannot establish that a neoplasm is malignant. However, if numerous typical mitoses ($> 6/mm^2$) are found without evidence of a traumatic event, the probability of malignancy is high. Similarly, the presence of atypical mitoses in a spitzoid tumor favors malignancy. An example of typical and atypical mitoses on a malignant lesion are shown in Figure 1 (e)-(f), respectively. Table 1 summarizes the main features distinguishing normal tissue, tissue with benign and malignant spitzoid lesion. The manual diagnosis process is highly time-consuming and commonly leads to discordance between histopathologists due to the ambiguity of these neoplasms [5]. This is why these lesions represent a formidable diagnostic challenge.

Table 1: Main histological features of normal melanocytes and spitzoid lesions.

Histological features	Normal tissue	Benign spitzoid lesion	Malignant spitzoid lesion
Basal and periodically distributed isolated melanocytes	Yes	No	No
Melanocytic nests	No	Well defined	Ill defined
Pagetoid patterns	No	Rare	Yes/No
Typical mitoses	No	No/Few	Common (usually numerous)
Atypical mitoses	No	No	Yes/No
Necrosis	No	No	Yes/No
Ulceration	No	Very rare	Yes/No
Marked nuclear pleomorphism	No	No	Common

The computer-aided diagnosis systems (CADs) aim to support pathologists in the daily analysis of skin biopsies, reducing both the workload and the inconsistency generated. With the emergence of digital pathology, the digitization of histological crystals into whole-slide images (WSIs) has been standardized [6], leading the way to the application of computer vision methods. The development of CADs based on WSI analysis presents important hardware limitations because of their large size. For this reason, the typical approach generally involves extracting small patches from larger WSIs, resulting in thousands of patches per image. The convolutional neural networks (CNN)-based approaches have been extensively tested for the detection of breast cancer [7–9], prostate cancer [9–11] or lung cancer [12, 13]. However, regarding skin cancer diagnosis, specifically for melanoma detection, most research was based on the analysis of dermoscopic images [14–22] and few studies have focused on the analysis of WSIs [23–26]. Hekler et al. [23] used transfer learning on a pre-trained ResNet50 CNN to differentiate between two classes, benign and melanoma tissues. The main limitation of this work is that they are not able to analyze entire WSIs but only a characteristic tumor sub-region. In De Logu et al. [24], a pre-trained Inception-ResNet-v2 network was then used to distinguish cutaneous melanoma areas from healthy tissues. However, this work didn't discriminate melanoma from nevi WSIs. In [25], the authors developed a deep learning system to automatically detect malignant melanoma in the eyelid from histopathological sections. The main limitation of this work is that the input of the algorithm is the tumor region and not the entire WSI image.

To the best of the authors' knowledge, no previous studies have focused on the SMTs distinction based on data-driven approaches. There is only one method based on hand-crafted feature extraction for SMTs identification [26]. In [26], the authors used a machine learning algorithm to assist in the diagnosis of SMT. In this study, a random forest classifier was used on numerical morphological characteristics extracted by the pathologists from histological images [26]. Therefore, the method does not extract features directly from the histological images. As SMTs are uncommon skin lesions, the available data is generally scarce. This is why this study used data from 54 patients.

Inspired by the main limitations of the studies focused on melanoma detection and more specifically on SMTs diagnosis, in this work, we put forward a novel semi-supervised inductive transfer learning strategy to

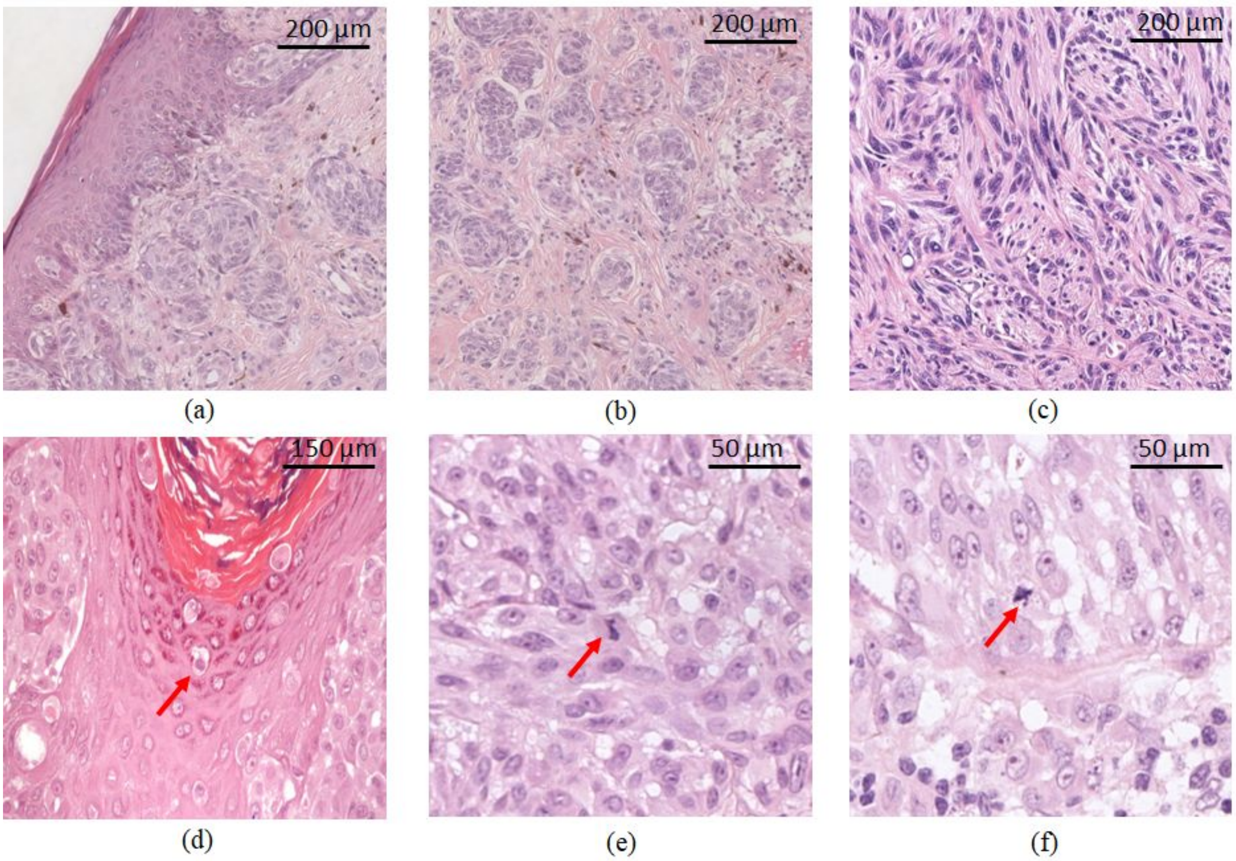


Figure 1: Representative patches extracted from WSIs presenting different spitzoid melanocytic lesions; (a)-(b): Benign spitzoid nevus containing well-defined melanocytic nests in an organized fashion; (c): Malignant lesion representative of the cellular disorder with ill-defined large tumor nest; (d): Malignant lesion with pagetoid spread, very common in this type of lesions; (e): Typical mitosis; (f): Atypical mitosis.

conduct both the local automatic detection of tumor regions and the global prediction of an entire biopsy. In summary, the main contributions of this work are:

- Spitzoid histological images are used for the first time to develop an automatic feature extractor.
- A new attention-based backbone is proposed to extract more accurate features.
- A novel framework based on inductive transfer learning to solve at the same time ROI selection and malignancy detection is developed.
- Multiple instance learning-based solutions are formulated in a novel framework for spitzoid lesion detection using biopsy-level labels.
- A wide clinical interpretability of the results achieved with the proposed methods is provided.

The rest of the paper is structured as follows. Section 2 details the related work regarding inductive transfer learning and multiple instance learning strategies, then the underlying methodologies of the present work, finally highlighting the improvement introduced in medical research. In Section 3, we present the data used in this work, CLARIFYv1, a private database comprised of skin WSIs from patients with spitzoid tumors. In Section 4, we describe the proposed methodology, mainly composed of two stages: i) development of a source model in charge of performing a patch-level classification to select tumor regions and ii) a target model based on a multiple instance learning approach to predict the malignancy degree at the biopsy level. Sections 5, 6 and 7 provide information on the performance outcomes related to the different classification tasks. Finally, in Section 8 we present our conclusions along with the future work.

2. Related work

A. *Inductive transfer learning*

Given a source domain D_S with a corresponding source task T_S, and a target domain D_T with a corresponding task T_T, transfer learning (TL) is the process of improving the target predictive function $f_T(\cdot)$ by using the related information from D_S and T_S, where $D_S \neq D_T$ or $T_S \neq T_T$ [27]. In the context of this work, we refer to inductive transfer learning (ITL) as the ability of the learning mechanism to enhance the performance on the target task (with a reduced number of labels) after having learned a different but related concept or skill on a previous task in the same domain [28]. The intuition behind this idea is that learning a new task from related tasks should be easier, faster and with better solutions or using less amount of labeled data than learning the target task in isolation. When the source and the target domain labels are available, the inductive transfer learning approach is known as multi-task learning.

Interest in this technique has grown in recent years in applications related to medical issues due to the promising results obtained. In this context, Caruana et al. suggested using multi-task learning in artificial neural networks and proposed an inductive transfer learning approach for pneumonia risk prediction [29]. Silver et al. introduced a task rehearsal method (TRM) as an approach to life-long learning that used the representation of previously learned tasks as a source of inductive bias. This inductive bias enabled TRM to generate more accurate hypotheses for new tasks that have small sets of training examples [30]. Zhang et al. used a technique based on inductive transfer learning to solve two-step classification problems: classification of malignant-nodule and non-nodule, and to classify the Serious-Malignant and the Mild-Malignant in malignant-nodule [31]. Tokuoka et al. provided an inductive transfer learning approach to adopt the annotation label of the source domain datasets to tasks of the target domain using Cycle-GAN based on unsupervised domain adaptation (UDA) [32]. Zhou et al. used an inductive transfer learning method to improve the performance of ocular multi-disease identification. In this case, the source and the target domain data were fundus images, but the source and target domain tasks were diabetic retinopathy lesion segmentation and multi-disease classification, respectively [33]. De Bois et al. used an inductive transfer learning approach to build a better glucose predictive model using a CNN-based architecture. A first model was trained on source patients that may come from different datasets and then, the model was fine-tuned to the target patients. Adding a gradient reversal layer, the patient classifier module made the feature extractor learn a feature representation that was general across the source patients [34].

In that context, we adopt an inductive transfer strategy to accurately classify instances from WSIs. The source model is trained to predict tumor regions by a patch-based CNN using inaccurate annotations with a large number of labels. After that, the backbone of the source model is retrained to classify nevus and malignant biopsies using a target model where the number of labels is reduced as this model is retrained at the biopsy level.

B. *Multiple instance learning*

Multiple instance learning (MIL), a particular form of weakly supervised learning, aims at training a model using a set of weakly labeled data [35]. In MIL tasks, the training dataset is composed of bags, where each bag contains a set of instances. A positive label is assigned to a bag if it contains at least one positive instance. The goal of MIL is to teach a model to predict the bag label. MIL approach has been successfully applied to computational histopathology for tasks such as tumor detection based on WSIs, reducing the time required to perform precise annotations [36–39]. In this vein, [36, 37] assigned the global label (cancerous against non-cancerous) to all patches of a slide. Campanella et al. [36] proposed a MIL-based deep learning system to accomplish the identification of three different cancers: prostate cancer, basal cell carcinoma and breast cancer metastases. In this case, they used an instance-level paradigm obtaining a tile-level feature representation through a CNN. These representations were then used in a recurrent neural network to integrate the information across the whole slide and report the final classification result to obtain a final slide-level diagnosis. Das et al. [37] used an embedded-space paradigm based on multiple instance learning to predict breast cancer. Specifically, they used a deep CNN architecture based on the pre-trained VGG19 network to extract the features of each bag. Then, the bag level representation is achieved by the aggregation of the features through the batch global max pooling (BGMP) layer at the feature embedding dimension. Silva et al. [39] used a novel weakly supervised deep learning model, based on self-learning CNNs, that leveraged only the global Gleason score of gigapixel whole slide images during training to accurately perform both, grading of patch-level patterns and biopsy-level scoring. Other works like [38] treated the tumor areas manually annotated by pathologists as a bag. In this case, the authors proposed a MIL method based on a deep graph convolutional network and feature selection for the prediction of lymph node metastasis using histopathological images of colorectal cancer. To the best of the authors' knowledge, no previous works have taken advantage of the promising MIL-based approaches for the diagnosis of melanocytic tumors yet. Our starting premise is that since there is at least one identifying patch of malignancy in a melanoma lesion, the MIL-based approach could assist in diagnosing a spitzoid lesion based on its whole context lessening the ambiguity between malignant and benign lesions. Additionally, in contrast to the works cited above, as in this study each bag contains the tumor region pseudo-labeled by the source model, the number of noisy labels is reduced, which will facilitate model training-loop since the number of available samples is particularly limited.

3. Materials

To evaluate the proposed learning methodology, we resort to a private database, CLARIFYv1, with histopathological skin images from different body areas that contain spitzoid melanocytic lesions. The database is composed of 53 biopsies from 51 different patients who signed the pertinent informed consent. The number of patients used in this study is relatively limited because these lesions are uncommon among the population. The tissue samples were sliced, stained and digitized using the Ventana iScan Coreo scanner at 40x magnification obtaining WSIs. The slides were analyzed by an expert dermatopathologist at the University Clinic Hospital of Valencia (CM). Specifically, 21 of the 51 patients under study were diagnosed as malignant melanocytic lesions (melanoma) and the rest as benign melanocytic lesions (nevus).

The global tumor regions, areas with spitzoid lesions, were annotated by the pathologists (AM, AM-Z and CM) using an in-house software based on the OpenSeadragon libraries [40]. With these annotations, WSIs were divided into regions of interest or ROI (tumor region) and non-interest regions (the rest of the WSI). Note that the tumor region denotes the part of the biopsy where the spitzoid lesion is found. After defining the tumor regions, the pathologist classified them as benign or malignant. Figure 2 shows the annotation of benign and malignant regions. To streamline the annotation task, these annotations were performed in a coarse way, so in some sub-regions there are tumor discontinuities not considered. This fact is shown in Figure 2 (b) and (d), where these patches have not patterned related to the tumor lesion.

In order to process the large WSIs, these were down-sampled to $10x$ resolution, divided into patches of size 512x512x3 with a 50% overlap among them. Aiming at pre-processing the biopsies and reduce the noisy patches, a mask indicating the presence of tissue in the patches was obtained by applying the Otsu threshold method over the magenta channel. Subsequently, the patches with less than 20% of tissue were excluded from the database. A summary of the database description is presented in Table 2. Note that, due to the irregular morphology of these lesions, the tumor shape is very different among patients, with the number of patches per patient varying considerably.

Table 2: CLARIFYv1 database description. Amount of whole slide images with their respective biopsy label (first row), number of patches of each tumor region (second row) and number of non-interest region (third row).

	Benign	Malignant
# WSI	30	21
# Tumor patches	3652	4726
# Non tumor patches	5842	8139

4. Methodology

The methodological core of the proposed approach is a semi-supervised CNN classifier able to detect the tumor region in a WSI and classify it into either benign or malignant spitzoid lesions. The proposed workflow is composed of a source and a target model, (θ^s) and (θ^t) respectively. The first model (θ^s) allows to automatically obtain the patches with significant features of spitzoid neoplasms, Figure 3. Tumor patches selected by the first model are then transferred to a second model (θ^t), Figure 4. This second model discerns malignant and benign biopsies using a MIL paradigm.

4.1. Source model: ROI selection

The objective of this stage is to build a 2D-CNN architecture able to extract discriminatory features from WSI patches to distinguish tumor regions.

A. *Backbone*

(1) Feature extractor. The patch-level feature extractor $G_f : x \rightarrow F$ is a CNN which maps an image x into an F feature volume. Since the deep learning models trained from scratch report worse performance in comparison to fine-tuned models when the

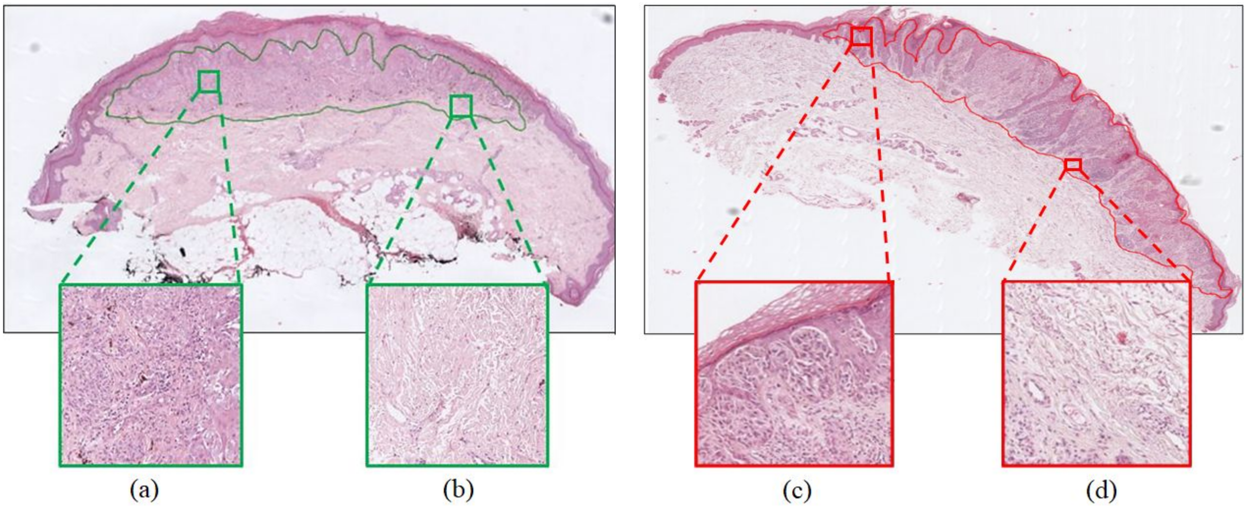


Figure 2: Annotation of a benign and a malignant spitzoid lesion. Patches (a) and (c) show characteristic patterns of the tumor region, benign and malignant respectively. Although patches (b) and (d) are inside the interest region annotated by the pathologists, these patches correspond to reactive stroma and do not contain tumor cells.

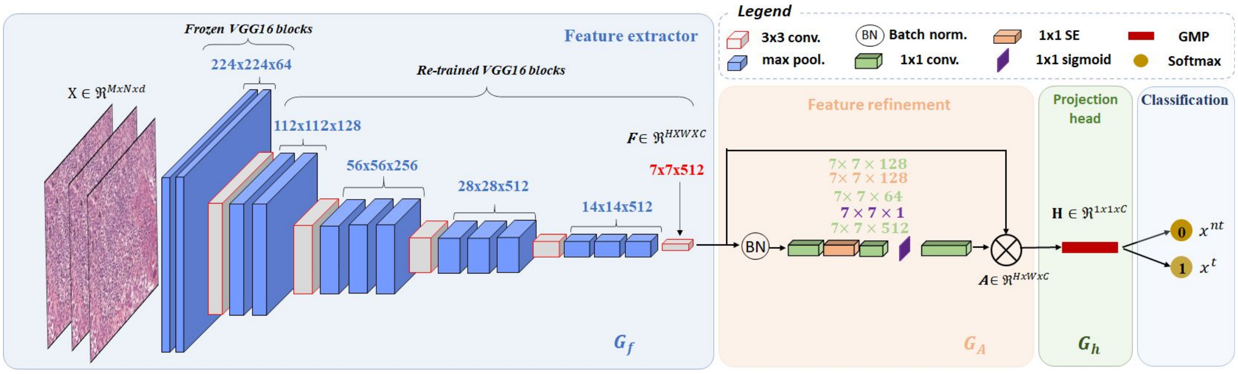


Figure 3: Overview of the proposed source model to conduct the tumor region detection. Blue and orange frames correspond to the base encoder network consisting of feature extraction and refinement. Note that VGG16 has been used as the feature extractor. After that, a projection head (green frame) maps the embedded representations in a lower-dimensional space to maximize the agreement in the classification stage (cyan frame).

amount of available data is limited, we fine-tuned several well-known architectures: VGG16 [41], ResNet50 [42], InceptionV3 [43] and MobileNetV2 [44]. All architectures were pre-trained with around 14 million natural images corresponding to the ImageNet dataset. For the feature extraction stage, the base model is extracted from those pre-trained models and partially re-trained. Since the patterns of the ImageNet dataset are very different from the histological ones (the value of the Frechet Inception Distance metric is around 68), it is optimal to keep the low-level features only (contours, combination of basic colors, general shapes, etc.). To this end, the weights of the first convolutional blocks from the pre-trained model are frozen, while the rest are re-trained to adapt the model to the specific application. The layer from which the freezing strategy is applied is empirically optimized for each architecture and it is specified in the experimental part of the paper, Section 5. Therefore, given a histological image $x \in \mathbb{R}^{M\times N\times d}$, where $M \times N \times d = 224 \times 224 \times 3$, a feature-embedded map $F \in \mathbb{R}^{H\times W\times C}$ is provided by the feature extractor. It is denoted as $F = G_f(x; \sigma^s)$ where σ^s is the set of trainable parameters of this source model.

(2) Feature refinement (SeaNet). Medical images always contain some irrelevant information that can disrupt the decision-making. For this reason, to solve ambiguous classification problems, it is essential to refine the features extracted by the CNN model. To this end,

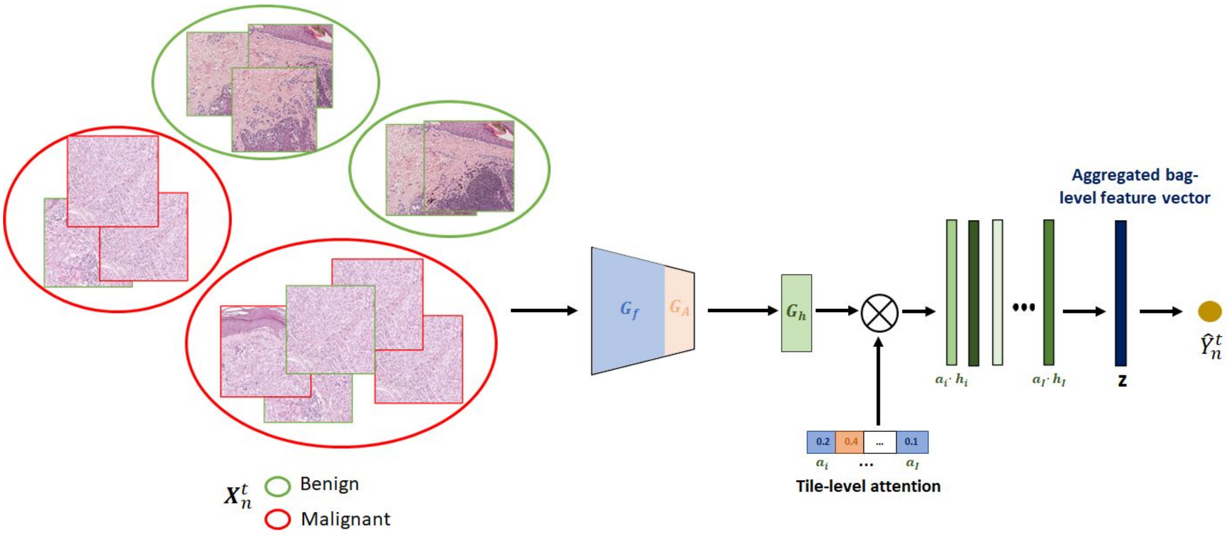


Figure 4: Pipeline showing the embedded-level approach for spitzoid melanocytic lesion classification. The weights of the pre-trained feature extractor and feature refinement of the source model (σ^s and δ^s) are used to initialize this approach. After that, we use the output of the projection head and tile-level attention to weight the patches in the prediction of a whole biopsy. Using an aggregated bag-level feature vector we classify the entire biopsy.

an attention module $G_A(F;\delta^s)$ was proposed to mimic the clinical behavior by focusing on the key features for the prediction, $G_A : F \rightarrow A$. In this case, the input of the attention module corresponds to the output feature map generated by the feature extractor, $F \in \mathbb{R}^{H\times W\times C}$. The proposed attention module works as a kind of autoencoder composed of 1×1 convolutions in which the filters are decreased and increased, respectively. Therefore, the feature maps obtained at the output of each of these convolution layers will have the same spatial dimension as the previous feature map, with the difference that the number of channels will have been changed to accomplish a combination of the features. In order to explore the dependencies existing among the different feature channels as well as the contextual information, the blocks called 'Squeeze-and-Excitation' (SE) [45] were implemented between the different convolutional reduction layers of the attention module, see Figure 5.

The input to the SE block, $G \in \mathbb{R}^{H\times W\times R}$, is embedded into a $s \in \mathbb{R}^{1\times 1\times R}$ vector by a global average pooling (GAP) layer, which provides a global distribution of responses by channels. Note that the number of filters R, corresponds to the number of channels at the output of the convolutional layers of the attention module. In the following step, s is transformed into $\hat{s} = \phi(W_2(\partial(W_1 s)))$ where ϕ is the sigmoid activation function, $W_1 \in \mathbb{R}^{\frac{R}{r}\times R}$ and $W_2 \in \mathbb{R}^{R\times \frac{R}{r}}$ are the weights of two completely fully-connected layers (FC) and ∂ is the Relu activation function. The parameter r is the reduction ratio for dimensionality reduction, in this case $r = 4$, indicating the bottleneck. After the sigmoid activation, the activations of $\hat{s}$ are ranged to [0,1] and it is used to recalibrate the input $\boldsymbol{G} = [g_1, g_2, ..., g_c]$ where $g_i \in \mathbb{R}^{HxW}$. The output feature map of this block is $\boldsymbol{G}_{se} = [\hat{s_1}g_1, \hat{s_2}g_2, ..., \hat{s_c}g_c]$.

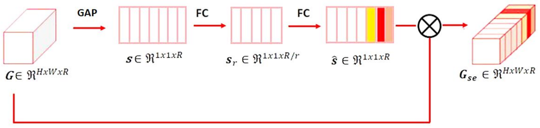


Figure 5: Architecture of the Squeeze-and-Excitation blocks used to exploit the dependencies between feature channels.

The last reduction layer of the attention module has the sigmoid as activation function to recalibrate the inputs and force the network to learn useful properties from the input representations. After increasing the number of filters to the same number as the input layer to this module, the output of the attention module is pondered with the output of the feature extractor obtaining a refined feature map $A \in \mathbb{R}^{H\times W\times C}$.

B. *Projection head module*

In this paper, we instantiate a projection head network, $G_h : A \rightarrow Z$, that maps the representations A to an embedding vector Z where the classification stage is addressed in a lower-dimensional space. In this case, different configurations already applied in the literature were tested in Section 5. In contrast to other widely used approaches such as the flattening of the activation volume resulting from the final convolutional block and the class prediction through consecutive fully-connected layers, the global max pooling (GMP) and the global average pooling (GAP) layers reduce the number of parameters decreasing the complexity of the model. At the end of the convolutional network, a softmax-activated dense layer is applied to address the tumor region identification.

4.2. Target model: WSI prediction

The target model aims to classify spitzoid lesions under an embedded-space paradigm using the biopsy-level labels for learning. To that end, our main goal is to find a compact embedding for the instances of a bag/WSI and combine these instance embeddings to a single embedding that represents the entire bag, see Figure 4.

Specifically, we denote each individual bag as $X^t_n = \left\{x^t_{n,1}, ..., x^t_{n,i}, x^t_{n,I_n}\right\}$, where $x^t_{n,i}$ is the i-th predicted tumor instance by the source model and I_n denotes the total number of predicted tumor region patches in a slide. Note that I_n can vary across bags. Hence, the objective of the target model becomes to obtain the label of a slide ($\hat{Y}^t_n$) from the tumor instances predicted by the source model ($x^t_{n,i}$), which can be defined as follows:

$$\hat{Y}^t_n = f(\left\{x^t_{n,1}, ..., x^t_{n,i}, ..., x^t_{n,I_n}\right\}, \omega^t) \quad (1)$$

where ω^t denotes the target model weights.

In order to find an embedding representation of each bag, we use the pre-trained backbone and the projection head module of the source model. In this manner, following an inductive learning strategy, the backbone already has prior knowledge concerning basic features of the histological database. After embedding each bag, $\mathbf{h}_n = G_h(G_A(G_f(X^t_n)))$, we obtain a C-dimensional feature vector for each instance. The bag label predictor $G_y : \{\mathbf{h}_i\}_{i \in I_n} \rightarrow \hat{Y}^t_n$ aggregates the C-dimensional feature vectors $\{\mathbf{h}_i\}_{i \in I_n}$ into a feature vector $Z_n \in \mathbb{R}^{1 \times C}$ representative of the bag. In the literature, there exist different aggregation functions such as batch global max pooling (BGMP) or batch global average pooling (BGAP). However, such functions are not flexible since they do not have trainable parameters. For this reason, in this work we use a trainable aggregation function [46]. In this case, $G_y(\cdot; \omega^t)$ is characterized by a set of trainable parameters $\mathbf{V} \in \mathbb{R}^{L \times C}$ and $\mathbf{w} \in \mathbb{R}^{L \times 1}$. The embedded feature vector per bag is obtained as $Z_n = \sum_{i \in I_n} a_i \cdot \mathbf{h}_i$, where a_i is defined as:

$$a_i = \frac{exp(\mathbf{w}^T tanh(\mathbf{V}\mathbf{h}_i))}{\sum_{j \in I_n} exp(\mathbf{w}^T tanh(\mathbf{V}\mathbf{h}_j))} \quad (2)$$

The attention-based aggregation function is differential and can be trained in a end-to-end manner using gradient descent. Additionally, the attention module not only provides a more flexible way to incorporate information from instances, but also enables us to localize informative tiles. The superiority of this aggregation function for spitzoid prediction will be shown in Section 5. Finally, the Z_n vector attaches to the dense layer with a sigmoid function-activated neuron to obtain the prediction at the biopsy level.

5. Ablation Experiments

In this section, we present the results of the different experiments carried out to show the performance of the proposed approach for the different classification tasks: patch-level classification (source model) and WSI prediction (target model). Note that a comparison with the current state-of-the-art methods was not possible as there are no algorithms focused on histological images of spitzoid tumors. Additionally, no public databases of histological images with melanocytic neoplasms have been found to apply our algorithms.

5.1. Database partitioning

Making use of the spitzoid database (CLARIFYv1), we carried out a patient-level data partitioning procedure to separate training and testing sets, aiming at avoiding overestimating the performance of the system and ensuring its ability to generalize. Specifically, 30% of patients were used to test the models, whereas the remainder of the database was employed to train the algorithm. To train the proposed models and optimize the hyperparameters involved in this process, the training set was divided following a 4-fold cross-validation strategy. We used four validation cohorts to optimize both the source and the target models. To encourage the source model to select the most relevant tiles, we used an instance dropout over the non-tumor region, since these represent the majority class. Specifically, instances were randomly dropped during the training, while all instances were used during the model evaluation.

5.2. Source model selection

A. *Backbone optimization*

According to the literature for histopathological image analysis, we compared as feature extractors the well-known ResNet50 and VGG architectures since they have reported the best performance [23, 25]. Additionally, we applied the proposed feature refinement

SeaNet, Squeeze and Excitation Attention Network, on each of these feature extractors in order to evaluate the enhancement introduced. To address an objective comparison of the proposed backbones, we kept the projection head module constant using a GAP layer. In Table 3, we contrast the validation results achieved by the different backbones trained in a binary-class scenario. The comparison was handled by means of different figures of merit, such as sensitivity (SN), specificity (SPC), positive predictive value (PPV), false positive rate (FPR) negative predictive value (NPV), F1-score (F1S), accuracy (ACC) and area under the ROC Curve (AUC). Note that the figures of merit listed above report the results for the average of the validation cohorts in the cross-validation process. Additionally, class activation maps (CAMs) were computed to highlight the regions of interest at patch-level in which the proposed source model paid attention to predict the samples, see Figure 6 and Figure 7. The backbone reporting the best performance during the validation stage was selected as the base encoder network to address the head projection optimization.

Training details. All the contrasting approaches were implemented using Tensorflow 2.3.1 with Python 3.6. Experiments were conducted on the NVIDIA DGX A100 system. NVIDIA DGX A100 is the universal system for all artificial intelligence (AI) workloads, offering unprecedented compute density, performance, and flexibility in a 5 petaFLOPS AI system. After intense experiments, the optimal hyperparameters combination was achieved by training the models for 120 epochs using a learning rate of 0.001 with a batch size of 64. A stochastic gradient descent (SGD) optimizer was applied to minimize the binary cross-entropy (BCE) loss function at each epoch. The base model of the fine-tuned feature extractor was also optimized, selected to freeze the first convolutional block for VGG16 and setting all layers as trainable for ResNet50.

B. *Head projection optimization*

In this section, we report the validation performance using different projection head modules. Specifically, we compare a small multi-layer perceptron (MLP) with one hidden layer of 128 neurons non-linearly activated by the ReLU function, a global max-pooling (GMP) layer and a global average-pooling (GAP) layer, see Table 4. It is important to note that the comparison was conducted using the proposed SeaNet (with VGG16) backbone for all the scenarios.

Training details. The same hardware and software systems as for the backbone section were used to optimize the head projection. Additionally, we use the same learning rate, batch size, loss function and number of epochs as in the previous section. In this case, we only changed the head projection.

5.3. Target model selection

A. *WSI label predictor optimization*

As mentioned throughout the manuscript, the backbone and the projection head module of the target model were optimized during the ROI selection, via the source model. After obtaining an embedded feature vector of each tile in a bag, it is necessary to implement an aggregation function. In this section, we compare the results, when three different aggregation functions were used: batch global max pooling (BGMP), batch global average pooling (BGAP) and batch global attention summary (BGAS), Table 5.

Training details. In order to generate bags and train the algorithms, a maximum of 300 image patches were randomly extracted from the source model prediction. In this case, the optimal results were obtained retraining the whole models during 100 epochs using a learning rate of 0.001 and a batch size of 1, in other words, one slide per batch. To minimize the BCE loss function at every epoch, the SGD optimizer was used.

6. Prediction Results

In this section, we show the quantitative and qualitative results achieved by the proposed strategies during the prediction of the test set. For both methods developed in this work, ROI selection and WSI classification, predictions were performed using the architectures with the best performance during the validation stage.

Quantitative results. Table 6 shows the results reached in the test prediction for the proposed source and target models.

Qualitative results. To qualitatively show the performance of the ROI selection model, we obtained probability heatmaps of representative samples indicating the presence of tumor region in the WSIs, Figure 8.

In the probability maps, for each pixel, the predicted probabilities for the ROI are estimated by bilinearly interpolating the predicted probabilities of the closest patches in terms of euclidean distance to the center of the patches. In addition, using these heatmaps, we visualize the distribution of attention weights, which were calculated for cases correctly classified into benign and malignant neoplasms, see Figure 9.

Table 3: Classification results reached during the validation stage with the proposed fine-tuned architectures. SeaNet: Squeeze-and-Excitation network.

	VGG16	**SeaNet (with VGG16)**	**RESNET50**	**SeaNet (with RESNET50)**
SN	0.8057 ± 0.1247	**0.8310** ± **0.1061**	0.8200 ± 0.1223	0.7494 ± 0.1736
SPC	0.9070 ± 0.0343	**0.9298** ± **0.0185**	0.8850 ± 0.0243	0.9290 ± 0.0422
PPV	0.8448 ± 0.0856	**0.8814** ± **0.0495**	0.8061 ± 0.1005	0.8800 ± 0.0316
FPR	0.0930 ± 0.0343	**0.0702** ± **0.0185**	0.1150 ± 0.0243	0.0828 ± 0.0235
NPV	0.8894 ± 0.0649	**0.9100** ± **0.0232**	0.8830 ± 0.0761	0.8693 ± 0.0516
F1S	0.8183 ± 0.0865	**0.8654** ± **0.0805**	0.8022 ± 0.1126	0.8100 ± 0.0927
ACC	0.8752 ± 0.0357	**0.9031** ± **0.0262**	0.8611 ± 0.0558	0.8770 ± 0.0329
AUC	0.8600 ± 0.0584	**0.8810** ± **0.0566**	0.8400 ± 0.0813	0.8500 ± 0.0737

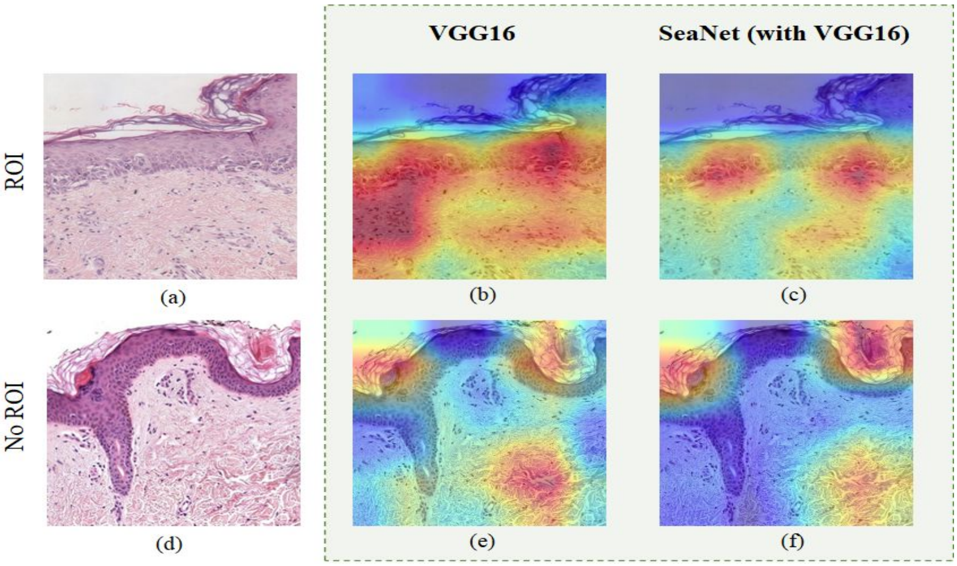


Figure 6: Class activation maps (CAMs) for images correctly classified as tumor region or ROI (first row) and non-tumor regions (second row). First column: original images; Second column: CAMs obtained using the VGG16 model. Third column: CAMs using Squeeze and excitation network (SeaNet) with VGG16 as the backbone. SeaNet model focuses on the most distinctive features and, in this case, pays attention to the pagetoid spread to define a patch as tumorous and to the healthy stromal region for the non-tumoral region.

7. Discussion

In this section, we make reference to the main contributions detailed throughout the paper and review the results obtained.

In contrast to the state-of-the-art studies for histological images classification, in which the input of the prediction model is the tumor region annotated by the pathologist, in this paper, we propose a framework able to first automatically select neoplastic regions of interest and then predict the malignancy or benignity of spitzoid neoplasms. Note that no previous studies seem to have proposed any automated method for the detection of these challenging neoplasms. Due to the absence of public spitzoid databases, the developed algorithms could not be validated with external databases, which can lead to biased results, according to the database used.

7.1. Source model: ROI selection

A. *About the ablation experiment*

Backbone selection. As a first stage, we carried out an optimization of the feature extractor for the selection of the tumor regions. Considering the limited amount of available samples, we decided to use the fine-tuning technique on the VGG16 and RESNET architectures. Particularly, from Table 3 we can observe that the use of sequential approaches (VGG16) provided slightly better results than architectures with residual blocks

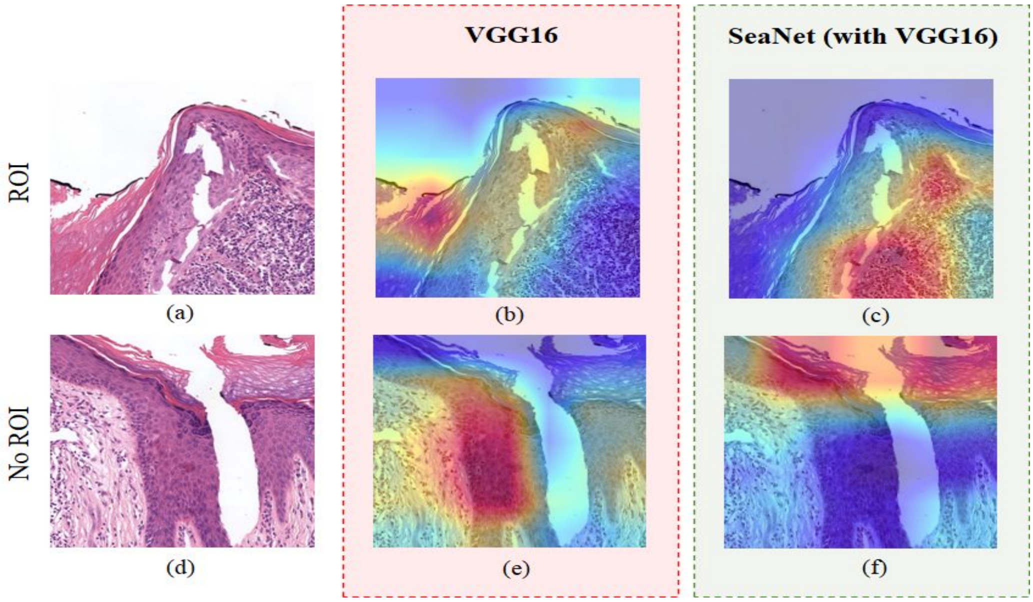


Figure 7: Original images (first column) and Class Activation Maps (CAMs) obtained with the VGG16 model (second column) and the Squeeze and excitation network (SeaNet) with VGG16 (third column). (b) and (e): Patches misclassified by the VGG16 model predicted as no ROI and ROI, respectively; (c) and (f): Patches well classified, ROI and No ROI respectively.

(RESNET). This fact is evidenced in several works in the literature for histopathological analysis where the sequential models used outperform residual ones [10]. Additionally, the proposed SeaNet module, characterized by the refinement of the features via convolutional attention blocks, reported a significant outperforming. Specifically, the SeaNet module via fine-tuning VGG16 architecture achieved the best results. The use of the attention module provides more distinctive feature maps and allows a considerable reduction in the incidence of false positive and false negative samples, leading to improve global metrics. Aiming at qualitatively observing the enhancements introduced by the refinement module, the CAMs of the best models (SeaNet with VGG16 and VGG16 alone) were obtained for correctly classified images (see Figure 6) and for images misclassified by the VGG16 model (see Figure 7). In Figure 6, we can see that both for the prediction of patches belonging to the tumor region (a) and for non-tumor ones (b), the SeaNet activations are focused on smaller regions. For the ROI prediction, the SeaNet (with VGG16) model is mainly focused on the pagetoid pattern present within the epidermis, defining the region as tumor. However, the VGG16 model extends its activations to lymphocytes found within the dermis. In this case, the lymphocytes do not necessarily determine that the region is tumorous, since this small amount of lymphocytes can also be found in healthy regions. Therefore, the VGG16 model without the attention module introduces certain noise in the prediction. Regarding the prediction of non-tumor regions, both models are focused on the epidermis and stromal region of the dermis. Regarding the cases where VGG16 misclassifies tumor regions, Figure 7 (b), the activations are focused on the epidermis region. In this case, the epidermis region has no patterns indicative of a melanocytic lesion, but for a correct classification, the activations would have to be focused on the melanocyte aggregate found in the upper region, as in the case of the SeaNet model, see Figure 7 (c). In this region, we find a large number of melanocytic cells with a high concentration of lymphocytes indicating an inflammatory reaction to a tumor region. For the case of the non-tumor region shown in Figure 7 (d), the VGG16 model erroneously predicts it by focusing on the melanocytic cells found in the epidermis, see Figure 7 (e). Normally, in healthy skin, the dermo-epidermal junction is composed of isolated melanocytic cells with a certain spacing between them. It is representative of a tumor when these cells ascend to the upper layers of the epidermis forming what is known as a pagetoid pattern or infiltrate the dermis forming nests. Furthermore, in this case, the epidermis has no patterns that would be representative of a melanocytic lesion. Unlike the VGG model, the SeaNet (with VGG16) model reports its activations in the epidermal region and based on it establishes the correct prediction, classifying this patch

Table 4: Classification results reached during the validation stage using different projection head modules. SeaNet: Squeeze-and-Excitation network (with VGG16 as backbone), MLP: multi-layer perceptron, GMP: global max-pooling, GAP: global average-pooling.

	SeaNet+MLP	SeaNet+GMP	SeaNet+GAP
SN	0.8716 ± 0.3000	**0.8729 ± 0.0371**	0.8310 ± 0.1061
SPC	0.9076 ± 0.0478	0.9143 ± 0.0131	**0.9298 ± 0.0185**
PPV	0.8460 ± 0.1018	0.8589 ± 0.0710	**0.8814 ± 0.0495**
FPR	0.0927 ± 0.0340	0.0857 ± 0.0131	**0.0702 ± 0.0185**
NPV	0.9100 ± 0.0348	**0.9140 ± 0.0283**	0.9100 ± 0.0232
F1S	0.8606 ± 0.0655	**0.8708 ± 0.0541**	0.8654 ± 0.0805
ACC	0.8940 ± 0.0320	0.9020 ± 0.0164	**0.9031 ± 0.0262**
AUC	0.8800 ± 0.0391	**0.8935 ± 0.2490**	0.8810 ± 0.0566

Table 5: Classification results reached during the validation stage using different aggregation functions. BGMP: batch global max-pooling; BGAP: batch global average-pooling; BGAS: batch global attention summary.

	BGMP	BGAP	BGAS
SN	0.5000 ± 0.3953	0.5833 ± 0.3062	**0.7500 ± 0.2764**
SPC	**0.9000 ± 0.3953**	0.8500 ± 0.1658	0.8500 ± 0.2764
PPV	0.6250 ± 0.4330	0.8375 ± 0.1709	**0.8667 ± 0.1414**
FPR	**0.1000 ± 0.2909**	0.1500 ± 0.3062	0.1500 ± 0.2764
NPV	0.7625 ± 0.1546	0.7848 ± 0.1388	**0.8869 ± 0.1207**
F1S	0.5018 ± 0.3873	0.6000 ± 0.1541	**0.7472 ± 0.1473**
ACC	0.7361 ± 0.0977	0.7361 ± 0.0417	**0.8229 ± 0.0262**
AUC	0.7000 ± 0.1744	0.7167 ± 0.0841	**0.8000 ± 0.0963**

Table 6: Classification results reached during the prediction stage. SM: source model; TM: target model. The proposed source model (SM) was composed of the SeaNet (with VGG16) + global max-pooling (GMP). The proposed target model (TM) used the batch global attention summary (BGAS) layer as an aggregation function.

	SM	TM
SN	0.9285	0.6700
SPC	0.9202	0.8900
PPV	0.8622	0.8000
FPR	0.0798	0.1111
NPV	0.9599	0.8000
F1S	0.8942	0.7300
ACC	0.9231	0.8000
AUC	0.9244	0.7800

as non-characteristic of a spitzoid lesion, see Figure 7 (f).

In any case, the inclusion of the proposed attention module outperforms the popular pre-trained architectures of the state of the art and reduces the number of noisy patches used as input to the target model.

Projection head module selection. After optimizing the backbone, we proceeded to select the projection head module that provided the best results. For this purpose, we tested three projection head modules: multi-layer perceptron (MLP), global average pooling (GAP) and global max pooling (GMP). Table 4 shows that the modules based on GAP and GMP provide very similar and significantly better results than those reported by the MLP. The outperforming of GMP and GAP compared to the fully-connected configuration could be explained by the reduction in the number of weights to be optimized, making the model simpler and more capable of generalizing to new images. Comparing the results provided by GAP and GMP, we can conclude that they are very similar. The main difference between these techniques lies in the method of squeezing the spatial dimension. While GMP considers only the maximum value for the feature map, in the GAP layer the whole spatial region contributes to its output. This explains why the GMP layer enhances SN results and the GAP layer improves SPC results. With the GMP layer, it is more likely to correctly classify a patch belonging to the tumor region, even if it contains a minimal tumor region. However, GAP takes into account the whole context so that regions with small tumor areas are likely to

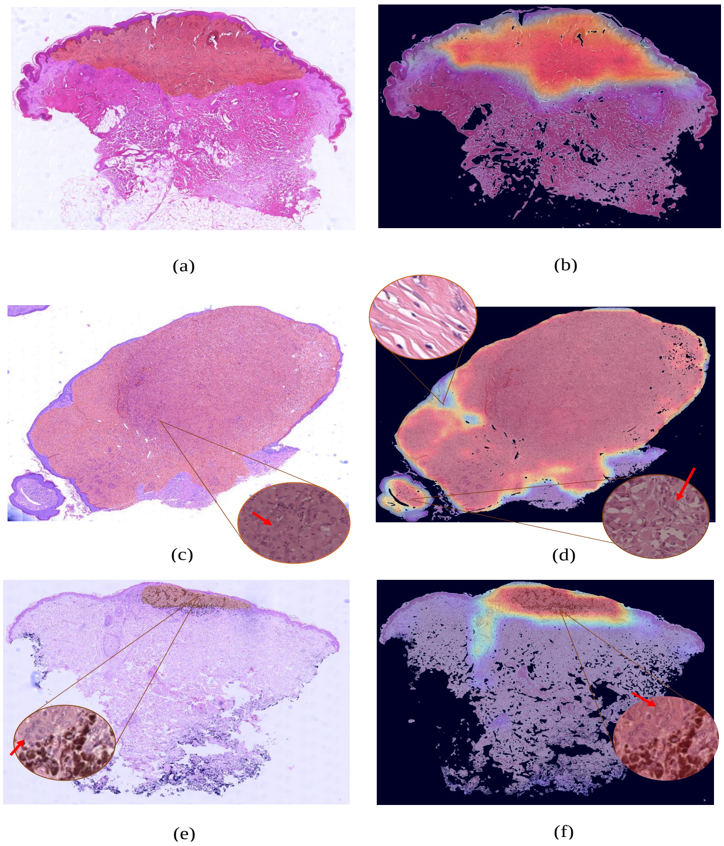
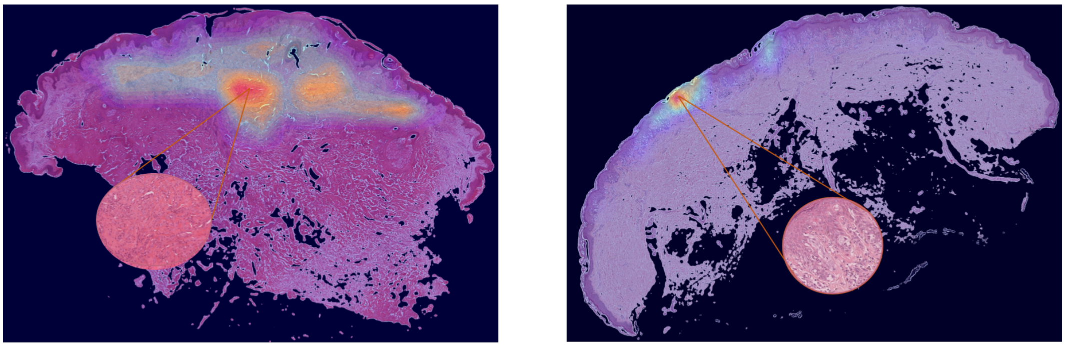


Figure 8: Whole slide image-level prediction for the source model (ROI estimation). (a) Manual annotation by experts; (b) System prediction completely in line with the annotation of (a); (c) Manual annotation by experts with expansion of areas with melanocytic nests characteristic of the lesion; (d) System prediction with certain areas annotated by the pathologists predicted as non-tumor regions. The expansion of the areas where there are no activations demonstrate that there are no melanocytic nests characteristic of the lesion; (e) Manual annotation by experts with expansion of area where melanocytic cells with melanosomes are found; (d) System prediction with expansion in the regions not annotated by the pathologist to demonstrate the presence of tumor cells.

be discarded. Although both show a very similar result, global metrics such as F1S and AUC exhibit a slight improvement with the GMP layer. Therefore, the GMP layer will be preferred as the optimal head projection module.

B. *About the prediction results*

Table 6 shows the results reached by the proposed ROI selection model. All the metrics reported here outperform those obtained in the validation phase. Figure 8 shows the probability maps for the lesion region of three test samples. The majority of the lesion regions predicted by the algorithm are depicted in Figure 8 (b), in which the prediction is completely in line with the annotation performed by the pathologists, Figure 8 (a). Some activation maps, such as those shown in Figure 8 (d), predict certain areas annotated by the pathologists as non-tumor regions. However, if we visualize the expansion of the areas where there are no activations, we can see that there are no melanocytic nests characteristic of the lesion, and therefore, we may be facing a discontinuity of the lesion as explained in Section 3. In contrast, in the lower part of Figure 8 (d), there are activations of tumor regions that have not been annotated by expert pathologists, see Figure 8 (c). However, if these regions are enlarged, it can be concluded that tumor cells are present. At times, due to the large amount of material in a lesion, pathologists can overlook some tumor areas. In the case of Figure 8 (e) and (f), there is also some discrepancy between the annotations performed by the pathologists and the activations predicted by the model. In these figures, we find melanocytic cells with



(a)

(b)

Figure 9: Visualization of the attention weights of the bag aggregation function in heat maps. (a) Benign sample; (b) Malignant sample.

melanosomes that give them their characteristic brown color. It is difficult to differentiate these tumor cells from melanophages (cells with brown staining and all of the same size) that are not tumor cells. In this case, if we zoom the activations of the algorithm (Figure 8 (f)) in those regions not annotated by the pathologist, we can see that there are also tumor cells. Therefore, the developed algorithm could help the decision-making in cases where there is ambiguity for the pathologists. In this context, the developed method enhances the detection of tumor areas.

7.2. *Target model: WSI prediction*

A. *About the ablation experiment*

WSI label predictor optimization. As discussed throughout the document, the backbone used by the target model was optimized during the selection of the source model. Therefore, in this case, it was only necessary to optimize the aggregation function required to perform a prediction using a MIL approach. From Table 5, we can observe that the use of the feature average of all patches containing a bag to obtain the embedded representation provides the best results (BGAP and BGAS aggregation functions). Additionally, the BGAS aggregation function improves the results provided by BGAP thanks to the introduction of optimized attention weights by updating the bag-level predictor weights (ω^t), achieving a validation accuracy of 0.8229. Therefore, we can conclude that the introduction of the attention module allows focusing on more relevant patterns, thus improving the final classification.

B. *About the prediction results*

Table 6 shows the results reached by the proposed target model in the test set. The results are in line with those obtained in the validation phase. Although the results are promising, there are some biopsies that are misclassified by the algorithm. This is because these types of lesions occasionally do not have universally accepted guidelines that can guarantee their specific diagnosis. Figure 9 shows the attention weights of the BGAS aggregation function for benign (Figure 9 (a)) and malignant (Figure 9 (b)) samples. The attention weights were normalized between 0 to 1 in each bag. The red regions in the attention weight maps represent the highest contribution for classification in each bag. Therefore, the bag class label is predicted by only using instances for which the attention values are large. In the case of a benign sample (Figure 9 (a)), the regions contributing to the class establishment are distributed over a wide area of the lesion, these areas being aggregates of melanocytes. However, the large attention weights for a malignant lesion are focused on small region characteristics of malignancy (in this case pagetoid pattern) as shown in Figure 9 (b).

8. Conclusion

In this work, we propose an inductive transfer learning framework able to perform both ROI selection and malignant prediction in spitzoid melanocytic lesions using WSIs. Our proposed framework is composed of a source model in charge of selecting the patches with characteristic lesion patterns. The source model introduces an attention module able to refine the features of the latent space to maximize the classification agreement. Using the backbone of the source model as a patch-level feature extractor and under a multiple instance learning approach, the target model predicts the malignancy degree by taking as input the tumor patches predicted by the first model. This innovative approach carried out in an end-to-end manner reported promising results for both ROI selection and WSI classification, achieving a testing accuracy of 0.9231 and 0.8000 for the source and the target models, despite the limited number of samples. Thus, our framework bridges the gap with respect to the development of automatic diagnostic systems for spitzoid melanocytic lesions. In future research lines, efforts should focus on improving the discrimination of malignancy and benignity with the acquisition of new samples and enhancements to the implemented attention module in the multiple instance learning approach.

Acknowledgements

We gratefully acknowledge the support from the Generalitat Valenciana (GVA) with the donation of the DGX A100 used for this work, action co-financed by the European Union through the Operational Program of the European Regional Development Fund of the Comunitat Valenciana 2014-2020 (IDIFEDER/2020/030).

Funding

This work has received funding from Horizon 2020, the European Union's Framework Programme for Research and Innovation, under grant agreement No. 860627 (CLARIFY), the Spanish Ministry of Economy and Competitiveness through projects PID2019-105142RB-C21 (AI4SKIN) and SICAP (DPI2016-77869-C2-1-R), Instituto de Salud Carlos III by the project PI20/00094 and GVA through project PROMETEO/2019/109. The work of Rocío del Amor has been supported by the Polytechnic university of Valencia (PAID-01-20). Laëtitia Launet and Andrés Mosquera-Zamudio have been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 860627.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Z. Apalla, A. Lallas, E. Sotiriou, E. Lazaridou, D. Ioannides, Epidemiological trends in skin cancer, Dermatology practical & conceptual 7 (2) (2017) 1–6.

[2] R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2017, CA: A Cancer Journal for Clinicians 67 (1) (2017) 7–30.

[3] T. Wiesner, H. Kutzner, L. Cerroni, M. C. Mihm Jr, K. J. Busam, R. Murali, Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy, Pathology 48 (2) (2016) 113–131.

[4] R. L. Barnhill, The spitzoid lesion: rethinking spitz tumors, atypical variants,'spitzoid melanoma'and risk assessment, Modern pathology 19 (2) (2006) S21–S33.

[5] S. Lodha, S. Saggar, J. T. Celebi, D. N. Silvers, Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting, Journal of cutaneous pathology 35 (4) (2008) 349–352.

[6] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, B. Yener, Histopathological image analysis: A review, IEEE reviews in biomedical engineering 2 (2009) 147–171.

[7] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, Jama 318 (22) (2017) 2199–2210.

[8] A. Rakhlin, A. Shvets, V. Iglovikov, A. A. Kalinin, Deep convolutional neural networks for breast cancer histology image analysis, international conference image analysis and recognition 10882 (2018) 737–744.

[9] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, J. Van Der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Scientific reports 6 (1) (2016) 1–11.

[10] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, V. Naranjo, Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection, Computer Methods and Programs in Biomedicine 195 (2020) 105637.

[11] O. J. del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, H. Müller, Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score, Medical Imaging 2017: Digital Pathology 10140 (2017) 101400O.

[12] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, M. Snyder, Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features, Nature communications 7 (1) (2016) 1–10.

[13] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nature medicine 24 (10) (2018) 1559–1567.

[14] N. C. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, J. R. Smith, Deep learning ensembles for melanoma recognition in dermoscopy images, IBM Journal of Research and Development 61 (4/5) (2017) 5–1.

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

[16] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, et al., Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, Annals of Oncology 29 (8) (2018) 1836–1842.

[17] R. C. Maron, M. Weichenthal, J. S. Utikal, A. Hekler, C. Berking, A. Hauschild, A. H. Enk, S. Haferkamp, J. Klode, D. Schadendorf, et al., Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks, European Journal of Cancer 119

(2019) 57–65.

[18] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, T. Holland-Letz, et al., Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task, European Journal of Cancer 113 (2019) 47–54.

[19] S. H. Kassani, P. H. Kassani, A comparative study of deep learning architectures on melanoma detection, Tissue and Cell 58 (2019) 76–83.

[20] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, et al., A deep learning system for differential diagnosis of skin diseases, Nature Medicine 26 (6) (2020) 900–908.

[21] A. Astorino, A. Fuduli, P. Veltri, E. Vocaturo, Melanoma detection by means of multiple instance learning, Interdisciplinary Sciences: Computational Life Sciences 12 (1) (2020) 24–31.

[22] C. Yu, S. Yang, W. Kim, J. Jung, K.-Y. Chung, S. W. Lee, B. Oh, Acral melanoma detection using a convolutional neural network for dermoscopy images, PloS one 13 (3) (2018) 1–14.

[23] A. Hekler, J. S. Utikal, A. H. Enk, C. Berking, J. Klode, D. Schadendorf, P. Jansen, C. Franklin, T. Holland-Letz, D. Krahl, et al., Pathologist-level classification of histopathological melanoma images with deep neural networks, European Journal of Cancer 115 (2019) 79–83.

[24] F. De Logu, F. Ugolini, V. Maio, S. Simi, A. Cossu, D. Massi, et al., Recognition of cutaneous melanoma on digitized histopathological slides via artificial intelligence algorithm, Frontiers in oncology 10 (2020) 1559.

[25] L. Wang, L. Ding, Z. Liu, L. Sun, L. Chen, R. Jia, X. Dai, J. Cao, J. Ye, Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning, British Journal of Ophthalmology 104 (3) (2020) 318–323.

[26] C. Devalland, Spitzoid lesions diagnosis based on smote-ga and stacking methods, Advanced Intelligent Systems for Sustainable Development (AI2SD'2019): Volume 2-Advanced Intelligent Systems for Sustainable Development Applied to Agriculture and Health 1103 (2020) 348.

[27] K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big data 3 (1) (2016) 1–40.

[28] R. Vilalta, C. Giraud-Carrier, P. Brazdil, C. Soares, Inductive transfer, Springer US (2010) 634–683.

[29] R. Caruana, Multitask learning, Machine learning 28 (1) (1997) 41–75.

[30] D. L. Silver, R. E. Mercer, The task rehearsal method of life-long learning: Overcoming impoverished data, Conference of the Canadian Society for Computational Studies of Intelligence (2002) 90–101.

[31] S. Zhang, F. Sun, N. Wang, C. Zhang, Q. Yu, M. Zhang, P. Babyn, H. Zhong, Computer-aided diagnosis (cad) of pulmonary nodule of thoracic ct image using transfer learning, Journal of digital imaging 32 (6) (2019) 995–1007.

[32] Y. Tokuoka, S. Suzuki, Y. Sugawara, An inductive transfer learning approach using cycle-consistent adversarial domain adaptation with application to brain tumor segmentation, Proceedings of the 2019 6th International Conference on Biomedical and Bioinformatics Engineering (2019) 44–48.

[33] Y. Zhou, B. Wang, L. Huang, S. Cui, L. Shao, A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability, IEEE Transactions on Medical Imaging (2020) 818–828.

[34] M. De Bois, M. A. El Yacoubi, M. Ammi, Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people, Computer Methods and Programs in Biomedicine 199 (2021) 105874.

[35] C. L. Srinidhi, O. Ciga, A. L. Martel, Deep neural network models for computational histopathology: A survey, Medical Image Analysis (2020) 101813.

[36] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature medicine 25 (8) (2019) 1301–1309.

[37] K. Das, S. Conjeti, J. Chatterjee, D. Sheet, Detection of breast cancer from whole slide histopathological images using deep multiple instance cnn, IEEE Access (2020) 213502–213511.

[38] Y. Zhao, F. Yang, Y. Fang, H. Liu, N. Zhou, J. Zhang, J. Sun, S. Yang, B. Menze, X. Fan, et al., Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 4837–4846.

[39] J. Silva-Rodriguez, A. Colomer, J. Dolz, V. Naranjo, Self-learning for weakly supervised gleason grading of local patterns, IEEE journal of biomedical and health informatics (2021) 3094–3104.

[40] Openseadragon, Archivo situacionista hispano, url http://openseadragon.github.io/ (1999).

[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014) 213502–213511.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 770–778.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 2818–2826.

[44] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, Proceedings of the IEEE conference on computer vision and pattern recognition (2017) 4700–4708.

[45] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition (2018) 7132–7141.

[46] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, International conference on machine learning (2018) 2127–2136.