Computers in Biology and Medicine 185 (2025) 109492

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Going Smaller: Attention-based models for automated melanoma diagnosis

Sana Nazari *, Rafael Garcia

Computer Vision and Robotics Group, University of Girona, Plaça de Sant Domènec, 3, Girona, 17004, Spain

ARTICLE INFO

Keywords:
Melanoma detection
Automated melanoma diagnosis
Skin cancer
Attention-based models
Compact models

ABSTRACT

Computational approaches offer a valuable tool to aid with the early diagnosis of melanoma by increasing both the speed and accuracy of doctors' decisions. The latest and best-performing approaches often rely on large ensemble models, with the number of trained parameters exceeding 600 million. However, this large parameter count presents considerable challenges in terms of computational demands and practical application. Addressing this gap, our work introduces a suite of attention-based convolutional neural network (CNN) architectures tailored to the nuanced classification of melanoma. These innovative models, founded on the EfficientNet-B3 backbone, are characterized by their significantly reduced size. This study highlights the feasibility of deploying powerful, yet compact, diagnostic models in practical settings, such as smartphone-based dermoscopy, and in doing so revolutionizing point-of-care diagnostics and extending the reach of advanced medical technologies to remote and under-resourced areas. It presents a comparative analysis of these novel models with the top three prize winners of the International Skin Imaging Collaboration (ISIC) 2020 challenge using two independent test sets. The results for our architectures outperformed the second and third-placed winners and achieved comparable results to the first-placed winner. These models demonstrated a delicate balance between efficiency and accuracy, holding their ground against larger models in performance metrics while operating on up to 98% less number of parameters and showcasing their potential for real-time application in resource-limited environments.

1. Introduction

Melanoma, the most dangerous type of skin cancer, arises from the uncontrolled growth of melanocytes. Early detection and precise classification are crucial for improving patient outcomes and reducing mortality rates [1]. In recent years, computer vision and deep learning techniques have dramatically improved the localization and classification of skin lesions using different image modalities [2].

However, the field of automated skin cancer detection faces several limitations. One of the most significant of these is that the more effective models are often ensembles of large and complex architectures that include transformers with hundreds of millions of parameters. What is more, there has been a consistent trend for these models to become increasingly larger in size over time. While such ensemble models demonstrate high performance, their practical application in the real world becomes challenging due to the significant computational resources required to deploy them [2,3]. Since resource and time efficiency considerations are crucial in a scenario where dermatologists attach a dermoscope to their smartphone, there is a clear need to balance model complexity and equipment requirements to ensure practicality in clinical or real-world applications of melanoma detection systems.

A further challenge lies in developing robust algorithms that can distinguish melanoma from other lesions. Some benign ordinary moles such as dysplastic nevus and melanoma share significant similarities in appearance, making it difficult to distinguish between them, even for trained dermatologists. Melanoma is also relatively rare compared to benign moles. Consequently, datasets containing melanoma instances are limited, resulting in an imbalance between nevus and melanoma cases within these datasets.

This lack of sufficient data is not exclusive to melanoma, however; many publicly available datasets exhibit a scarcity of images for other skin cancers, including basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). Additionally, the number of samples representing pre-cancerous lesions like actinic keratosis (AK) is also limited in these datasets [4]. Such an insufficiency of data poses yet another challenge for machine learning models, as they may not be exposed to enough examples of lesions to effectively learn and generalize, potentially impacting their performance and ability to accurately discriminate between benign and malignant samples.

Furthermore, although vision transformers have brought about a revolution in the machine-learning field, their widespread applicability

* Corresponding author.
E-mail address: sana.nazari@udg.edu (S. Nazari).

https://doi.org/10.1016/j.compbiomed.2024.109492
Received 10 May 2024; Received in revised form 8 November 2024; Accepted 26 November 2024
Available online 4 December 2024

to skin lesion datasets in classification tasks is hindered by the fact that these models typically require a large amount of data [5], meaning they suffer from the same scarcity issue already mentioned above. For this reason, there is a need to develop attention-based architectures specifically tailored to the task of skin lesion analysis. This involves designing models that can effectively leverage attention mechanisms while accommodating the aforementioned challenges posed by limited data availability.

In this work, we address the three challenges outlined above and devise a pipeline to overcome these issues. We start by outlining the pre-processing stage and the selection of a loss function capable of mitigating concerns regarding class imbalance. We then present four novel deep CNN-based architectures that incorporate attention mechanisms and are capable of performing multi-class classification of skin lesions. These models have been designed to strike a balance between model complexity and accuracy in performance. The robustness of our model is then demonstrated through the use of diverse evaluation metrics suitable for assessing the performance of models trained on imbalanced datasets with a particular focus on the medical relevance of the problem. Finally, we present a comparative assessment of our models and the top three winning models from the International Skin Imaging Collaboration (ISIC) 2020 challenge [6]. Below is a summary of the main contributions of our work:

- A comparative evaluation of loss functions, specifically designed to address data imbalance.
- The proposal of four novel light-weight attention-based models for detecting melanoma.
- A demonstration of real-world performance assessment of the designed models through testing on two independent test sets.
- Detailed comparison of our novel models with existing top-performing models.

The remaining sections of the article are organized as follows. Section 2 provides an overview of the related literature. The Methodology Section 3 starts with a comprehensive explanation of the pre-processing methods implemented in this study, followed by details of the integrated attention modules. Section 3.3 describes the architecture of our proposed models, and Section 3.4 provides details of the selected loss functions. In the Experimental results Section 4, we first provide details of our datasets 4.1, training strategy and hyper-parameter tuning 4.2, and then the results of our experiments 4.3. This is followed by a comparative discussion of our findings in Section 5. Finally, Section 6 briefly reviews our methods and results, offering additional insights and discussing future work.

2. Related work

An extensive search was performed of the Web of Science to identify studies that trained classifiers for melanoma diagnosis, using datasets that incorporated both ISIC 2019 and ISIC 2020. Many of the reviewed studies focused on training binary classifiers. In most cases, the evaluation of these works was done using a test set derived from the original combined dataset [7–12]. Some studies also used smaller, more balanced subsets extracted from the ISIC 2019 and 2020 datasets [7,12]. A summary of the reviewed studies is presented in Table 1.

Some articles used well-established deep CNN architectures, such as EfficientNet-B6, MobileNet or InceptionV3 [7,8,11,12], with pre-trained weights sourced from ImageNet. Others introduced innovative methodologies and pipelines for melanoma diagnosis. Saeed et al. [10] used a VGG16 to extract features from images and then a Support Vector Machine (SVM) to classify the extracted features, while Dong et al. [9] proposed a new classifier based on an EfficientNet-B5 backbone. Additionally, the winning model of the ISIC 2020 challenge was improved in a later study and tested on an independent test set provided by the same authors [13]. This model comprised an ensemble of 18 pre-trained deep CNN models trained on different image sizes and incorporating patient metadata.

We also searched for works specifically exploring the application of attention mechanisms to CNN models in automated melanoma diagnosis. To improve the ability of CNNs to learn features for image classification tasks, several attention-based techniques have been developed. Wang et al. [14] introduced a Residual Attention Network (RAN), where attention weights were learned by trainable convolutional layers. Another influential work in image classification is the Squeeze-and-Excitation Network (SENet) proposed by Hu et al. [15], which uses channel-wise multiplication between the attention weights. Woo et al. [16] presented the Convolutional Block Attention Module (CBAM) to use channel-wise and spatial-wise attention to boost the representation learned by the network.

To avoid overfitting on small training datasets to classify skin lesions, Zhang et al. [17] proposed an attention-based method called ARL-CNN, which uses attention weights during network training. He et al. [18] designed a mixed attention mechanism, DeMAL-CNN, which considers both spatial and channel-wise attention information to classify skin lesions. Additionally, Zenghui et al. [19] introduced an auxiliary learning approach that incorporates a dual attention mechanism, auxiliary learning with loss functions at different stages of the model and a data oversampling approach. In recent studies, class-wise attention was introduced by Naveed et al. [20], who integrated it into a DenseNet-121 model, achieving an AUC of 0.99 on the HAM10000 dataset [21]. This performance surpassed that of existing related works. Similarly, Tan et al. [22] incorporated a global–local attention module into a ResNet-50 model to address class imbalance, resulting in an F1-score of 0.874 on the ISIC 2018 dataset. Their module was compared to other attention modules, such as SENet and CBAM, and demonstrated superior effectiveness. Lastly, in their study Omeroglu et al. [23] applied soft attention to the Xception architecture.

A significant proportion of the models examined in the reviewed articles were either quite large in terms of number of parameters or failed to assess the real-world performance of their models on an independent set of data. The size and complexity of a model can have critical implications for their practical deployment. Large models with an excessive number of parameters may require substantial computational resources, making them less feasible for deployment in resource-constrained environments. Moreover, the absence of independent testing poses a potential risk to generalization of the models, while models trained and evaluated solely on the same dataset may unintentionally capture dataset-specific patterns rather than learning underlying features.

3. Methodology

In this section, we provide a detailed explanation of our methodology, starting with details of our pre-processing approaches. We then explore the attention blocks and attention mechanisms within CNNs. The aim of this exploration is to clarify the role of attention in enhancing the discriminative capabilities of our models. Next, we analyze several loss functions to understand the differences that influence their performance on datasets with unbalanced class distributions. And finally, we discuss our proposed classifier models and architectures in detail, explaining the design complexities and the rationale behind the choices made in shaping these models.

3.1. Pre-processing

Pre-processing is a crucial step in developing a deep model since it has a significant impact on the information that the model learns during training. This process is especially important in the context of skin lesion diagnosis because most classes have very little data available. Additionally, the presence of artifacts in some of the images further highlights the importance of careful pre-processing. The strategies used

Table 1
Details of similar classification studies combining ISIC 2019 and ISIC 2020 datasets. # Samples: number of samples; Test Split: the percentage of data split for testing; #Parameters: number of parameters; Ind. Test: whether an independent set of images was used to evaluate the models; ROC-AUC: area under the ROC curve for the test set.

Study	#Samples	Test Split	Model	#Parameters	Ind. Test	ROC-AUC
Balaha and Hassan [7]	11,449	15%	MobileNet	4M	No	0.99
Bandy et al. [8]	58,457	9%	EfficientNet-B6	43M	No	0.99
Dong et al. [9]	58,457	20%	Custom CNN	30M	No	0.97
Saeed et al. [10]	58,457	20%	VGG16-SVM	144M	No	0.92
Jaisakthi et al. [11]	58,457	Not reported	EfficientNet-B6	43M	No	0.96
Mijwil [12]	24,000	20%	Inception-V3	25M	No	0.87
Marchetti et al. [13]	58,457	–	EfficientNetB4-7-SeresNext101-ResNest101	644M	Yes	0.86

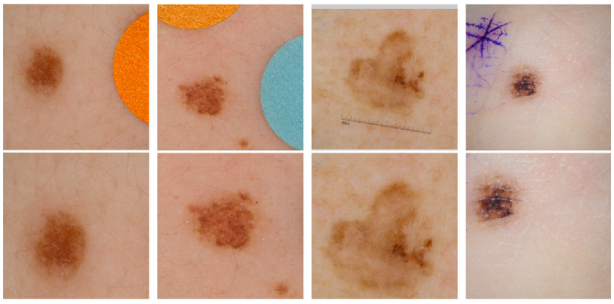

Fig. 1. Sample results of artifact cropping. The top row displays the original images with artifacts present and the bottom row displays the images after cropping.

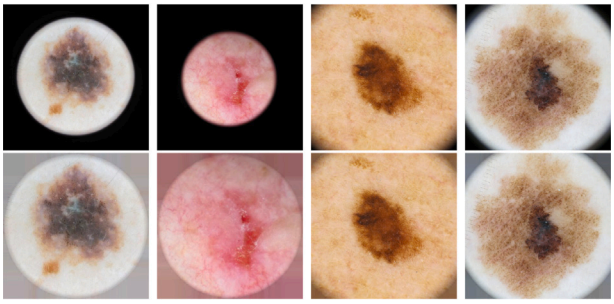

Fig. 2. Sample outcomes of dark corner artifact (DCA) removal. The first row shows the original images, while the second row depicts the corresponding modified samples.

in this phase not only shape the input data but also play a critical role in improving the ability of the model to identify meaningful patterns in limited and potentially noisy datasets.

After examining the training and validation datasets (ISIC 2019 and 2020), we noticed that they contained artifacts that could potentially bias the model, such as band-aids, dermoscopic rulers, dermoscopic dark rings and ink markings (see Fig. 1). Previous studies have shown that these artifacts can cause confusion and distract the model, leading to inaccurate analysis [24]. Moreover, Pewton and Yap [25] investigated the impact of the dark corner artifact (DCA) on the decision-making of convolutional neural networks (CNNs) and developed a dynamic approach to automatically detect and eliminate this artifact from images. This method involves a masking approach to identify the DCA. Following the identification of the DCA, the approach proceeds to either crop or inpaint the affected area, considering both the size and intensity of the artifact. Finally, a GAN-based super-resolution method is employed to augment the quality of the modified image.

In our work, the ruler, band-aid, and marker artifacts were addressed by cropping the affected samples, which at times proved challenging due to their proximity to the lesions. The goal was to minimize the impact of these artifacts on the model's perception and focus on the essential features of the skin lesions. Fig. 1 provides visual examples of the outcomes of this targeted cropping approach. In addition, the DCA removal method was implemented on our dataset to address the dermoscopy dark ring artifacts, resulting in the modification of a total of 10,513 samples. Fig. 2 depicts examples of the final outcomes achieved through the DCA removal method.

Following this, we applied a diverse set of data augmentation methods to enhance the training and validation dataset for melanoma classification. These techniques included random transformations such as transposition, vertical and horizontal flips, and adjustments to brightness and contrast. Additionally, our augmentation process incorporated various types of blurring (directional motion blur, median blur and Gaussian blur), random noise application (Gaussian noise) and distortion methods (optical distortion, grid distortion and elastic transformation). Furthermore, contrast-limited adaptive histogram equalization (CLAHE), hue, saturation and value adjustments, shift-scale rotation and cutout operations were also implemented. Finally, all images were resized to 384 × 384 pixels [26,27].

In the final phase, we tested all models with and without test-time augmentation (TTA). TTA involves applying 20 distinct transformations to each image in the test set and then averaging the predictions from all 20 transformed instances to obtain the final prediction for each image. To ensure consistency across our analysis, when applied, the same TTA methodology was used in all models, including experiments conducted with the winning ISIC models.

3.2. Attention modules

The role of attention in human perception is well-established, a key feature of it being the human visual system's tendency to avoid processing an entire scene simultaneously. Instead, humans employ a sequential approach, capturing partial glimpses and selectively focusing on salient parts to enhance their understanding of visual structures [28]. Vision transformers have become a powerful tool in computer vision, using attention mechanisms to selectively focus on important parts of an image. This ability to capture long-range relationships and global contexts makes transformers particularly well-suited for tasks such as image classification [29].

Although transformers are powerful tools, they can be problematic when used on imbalanced datasets with few data due to their self-attention mechanism-based architecture, which requires a large amount of diverse data for them to learn and generalize patterns. When faced with imbalanced datasets and small samples for some classes, transformers may have trouble identifying meaningful representations, which can result in biased attention and possibly overfitting to the majority class. CNNs, on the other hand, are exceptionally efficient at learning spatial relationships and hierarchical features within images, which makes them more appropriate for tasks where there is a limited amount of data and a class imbalance [30,31]. Moreover, their built-in inductive biases help CNNs learn with fewer data.

In CNN, the neurons of the first layer are dedicated to capturing features within a defined area of the image, known as the receptive field of each neuron. If the filters employed have a size of 3 × 3, this area coincides with the filter dimensions. If in subsequent CNN layers each neuron undergoes convolution with the same 3 × 3 area from the preceding layer, then they contribute to an enlarged receptive

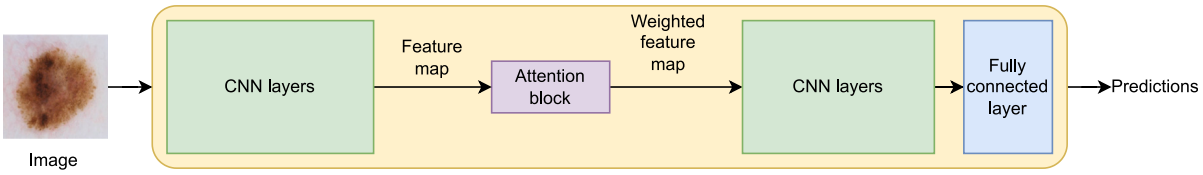

Fig. 3. The architecture of an attention-based CNN.

field in the input image. As the network deepens, individual neurons correspond to progressively larger receptive fields. The array of neurons along the third dimension forms a learned feature vector representing the receptive field within the image. Varied receptive fields represent distinct regions in the image, enabling the application of attention at this level through a weighted combination of the feature vectors. By training the model to learn the relative importance of these corresponding feature vectors, the network learns to pay attention to more important regions within the image [32].

In our proposed architectures, we incorporate attention blocks to enhance feature extraction in convolutional neural networks. The process involves passing the features, initially extracted by the CNN, through an attention mechanism. Within this attention block, a weight vector is computed based on the inherent characteristics of the features. Subsequently, these weights are applied to the original feature vector, creating a weighted feature representation. The resulting weighted feature vector serves as the output of the attention block. The overall pipeline of the proposed attention-based architecture is depicted in Fig. 3. More details of our proposed architectures are provided in Section 3.3. All the attention modules we implemented are lightweight and have very few extra parameters and calculations. A detailed explanation of each attention block we employed in our models is provided below:

3.2.1. Spatial attention

This block generates a spatial attention map, leveraging inter-spatial relationships within features. To compute spatial attention, average-pooling and max-pooling operations are applied along the channel axis, and their results are concatenated to form an efficient feature descriptor. The concatenated descriptor undergoes a convolution operation, yielding a spatial attention map denoted as $M_s(F) \in \mathbb{R}^{H\times W}$, indicating where to emphasize or suppress attention [16,33]. The spatial attention is computed as shown in Eq. (1):

$$M_s(F) = \sigma\left(f_{k\times k}\left(\left[\text{AvgPool}(F); \text{MaxPool}(F)\right]\right)\right) \tag{1}$$

where σ is the sigmoid function and $f_{k\times k}$ represents a convolution operation with a filter size of $k \times k$ (we chose k=7).

3.2.2. Convolutional block attention

The channel attention mechanism applies max-pooling and average-pooling across the spatial scope of the feature map. Then it passes the features through a multi-layer perceptron (MLP) to produce a corresponding channel attention map [16,32]. Finally, the derived features are combined, enabling the calibration of feature sensitivity in the channel dimension. The channel-wise attention mechanism is calculated as presented in Eq. (2), where σ is the sigmoid function:

$$M_c(F) = \sigma\left(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))\right) \tag{2}$$

The Convolutional Block Attention Module (CBAM) integrates spatial attention with channel attention [16], facilitating the model's enhanced focus on informative features. This attention mechanism incorporates both spatial and channel attention in a sequential manner, first employing channel attention and then spatial attention. By leveraging this dual-attention strategy, CBAM enables the model to selectively emphasize relevant channels and spatial regions. CBAM involves the sequential computation of a 1D channel attention map and a 2D spatial attention map, denoted as $M_c \in \mathbb{R}^{C\times 1\times 1}$ and $M_s \in \mathbb{R}^{1\times H\times W}$ respectively, based on an intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$. The attention process of CBAM is shown in Eq. (3), where $\odot$ represents element-wise multiplication:

$$\begin{aligned} F' &= M_c(F) \odot F, \\ F'' &= M_s(F') \odot F' \end{aligned} \tag{3}$$

In this procedure, the initial feature map F undergoes sequential processing. First, it is fed into the channel attention module M_c, and the resulting output is element-wise multiplied with the original F, generating F'. Subsequently, F' is input to the spatial attention module, Ms. The output of this step is then multiplied element-wise by F'. This outcome denoted as F'', represents the output of the entire attention block.

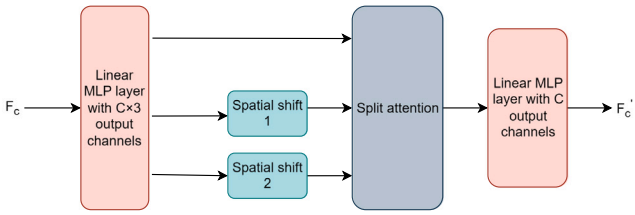

Fig. 4. Architecture of the Spatial-Shift MLPv2 attention block. F_c denotes the feature map with C channels and F_c' denotes the modified feature map with the same channels.

3.2.3. S2-MLPv2 attention

The spatial-shift (S2) MLP backbone consists of a patch-wise fully connected layer, spatial-shift MLP blocks and a fully connected layer [34]. S2-MLP can be enhanced by expanding the feature map using an MLP layer and then dividing the expanded feature map into three splits. Each split is then independently shifted, and the split feature maps are finally merged through split attention. This attention strategy is known as the spatial-shift MLPv2 (S2-MLPv2) [35], and the overall structure of its attention block is depicted in Fig. 4.

3.2.4. Spatial group-wise enhance attention

The Spatial Group-wise Enhance (SGE) module is designed to highlight multiple active areas with different high-order semantics. SGE can adjust the importance of each sub-feature by generating an attention factor for each spatial location in each semantic group. This enables each individual group to autonomously enhance its learned expression and suppress possible noise [36].

3.3. Proposed attention-based models

Our experimental framework is based on the EfficientNet architecture, which has demonstrated efficacy in the field of skin cancer diagnosis [3]. We selected the EfficientNet-B3 [37] model due to its optimal balance between model complexity and parameter efficiency, which enables it to capture and represent the significant features of our data. This is a crucial aspect, because large models require substantial computational resources, while very small models may struggle to capture essential features of dermoscopic images.

The fundamental component of EfficientNet models is the inverted residual block with squeeze-and-excitation (MBConv), which is composed of a depth-wise convolution, an additional 1 × 1 convolution and a squeeze-and-excitation module that uses global average pooling (GA) and two fully-connected layers to adaptively re-calibrate the channel-wise feature responses. Skip connections and the Swish (SiLU) activation function are also used by the inverted residual block to improve network performance [37]. As Fig. 5 illustrates, EfficientNet-B3 is composed of seven major blocks, each containing multiple MBConv blocks. In total, EfficientNet-B3 contains 26 MBConv blocks. These blocks are distinguished by the filter size used in the convolutional layers within the block, which can range from 1 × 1, 3 × 3 to 5 × 5, and whether the block includes an inverted residual connection. The network has approximately 12 million parameters and 2 billion floating point operations (FLOPs).

Throughout our experiments, we explored the integration of attention mechanisms within the architecture of EfficientNet- B3. Applying attention mechanisms after each major block, we experimented with both single and multiple applications of attention. Furthermore, we conducted tests by combining various attention blocks at different levels within the model's architecture. This comprehensive exploration allowed us to assess the impact of attention mechanisms on feature extraction and representation, providing insights into their effectiveness at different stages of the EfficientNet-B3 model.

Our study revealed that in most cases, the model's performance significantly degraded when attention was applied after the initial block (block 0). This decline in efficacy could be attributed to the model's early developmental stages at this juncture. Applying attention during this early phase appeared to lead the model astray, to discard crucial information and misdirect its focus towards less pertinent aspects of the image. This finding highlights the sensitivity of attention mechanisms to the model's progression and underscores the need for judicious placement to avoid detrimental effects on performance. This observation is similar to the results presented in the work of Alhichri et al. [32], which was conducted on EfficientNet-B3 using different attention mechanisms and image modality. Ultimately, we opted for four attention-based models that exhibited superior performance in our evaluations.

Our initial novel attention-based model (Spatial-B3) was constructed by integrating spatial attention blocks into the EfficientNet-B3 base architecture. These attention blocks were strategically introduced after each major block, except for block 0. This design choice was made to leverage the potential of spatial attention in highlighting important regions within the feature maps generated by the major blocks. The architecture of this model is illustrated in Fig. 6. Similarly, the second model (SGE-B3) proposed in our study entails the incorporation of SGE attention blocks placed after the execution of blocks 1 through 6, as visually depicted in Fig. 7.

To formulate the third model (S2-CBAM-B3), we implemented CBAM blocks after each major block, including block 0. Additionally, following the execution of block 5, just before introducing a CBAM block, we incorporated an S2-MLPv2 attention block (see Fig. 8). This specific configuration seeks to take advantage of the well-positioned CBAM and S2-MLP attention mechanisms to enhance the model's ability to capture spatial and channel features and optimize its attention focus across different processing stages.

In our last model (S2-MLP-B3) iteration, we opted for a simplified approach by integrating a singular block of S2-MLPv2 attention immediately after block 5. The architecture for this model is represented in Fig. 9.

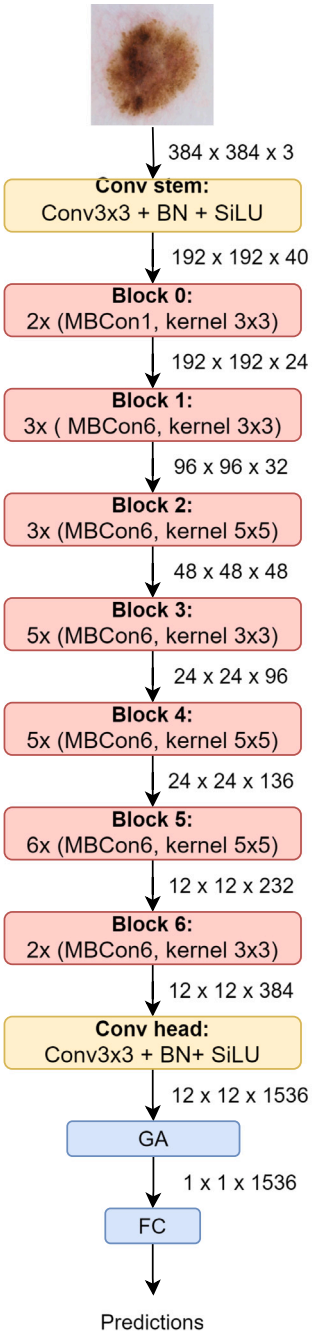

Fig. 5. The architecture of the EfficientNet-B3 model. BN: batch normalization layer; SiLU: swish activation function; GA: global average pooling layer; FC: fully-connected layer.

3.4. Loss function

One of the greatest obstacles in the field of medical imaging, particularly in the context of diagnosing skin cancer, is data scarcity. This is particularly problematic for rare skin lesions, for which there are very few samples available. The problem with data scarcity for these rare skin conditions is that conventional models tend to overfit the majority class. In our case, the dataset is also highly imbalanced (see Section 4.1 for the details). Using a suitable loss function is crucial to overcome this challenge and keep the model from excessively favoring the majority class.

A popular classification loss function, the Categorical Cross-Entropy (CE) loss, suffers from a certain limitation when it comes to imbalanced classification: it prioritizes the majority class during training. To overcome this problem, other loss functions, like weighted CE, Focal loss [38] or Label Distribution Aware Marginal loss (LDAM) [39] are preferable. These variations assign different weights to classes based

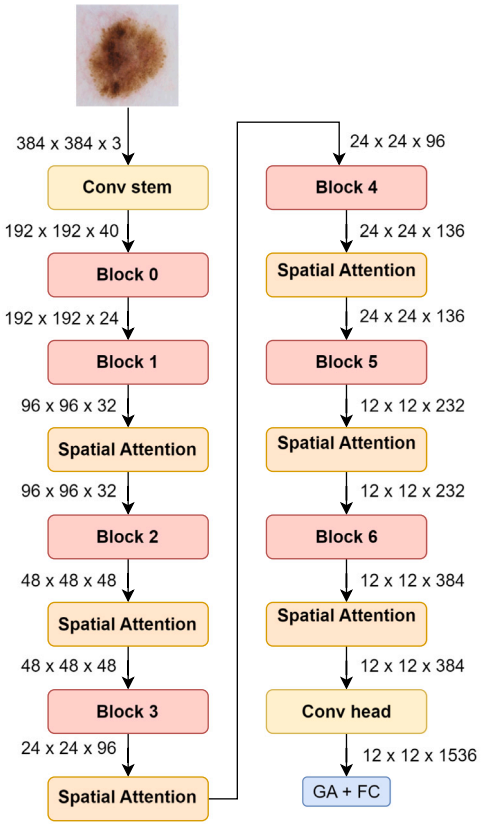

Fig. 6. Spatial-B3: Attention-based model using spatial attention blocks.

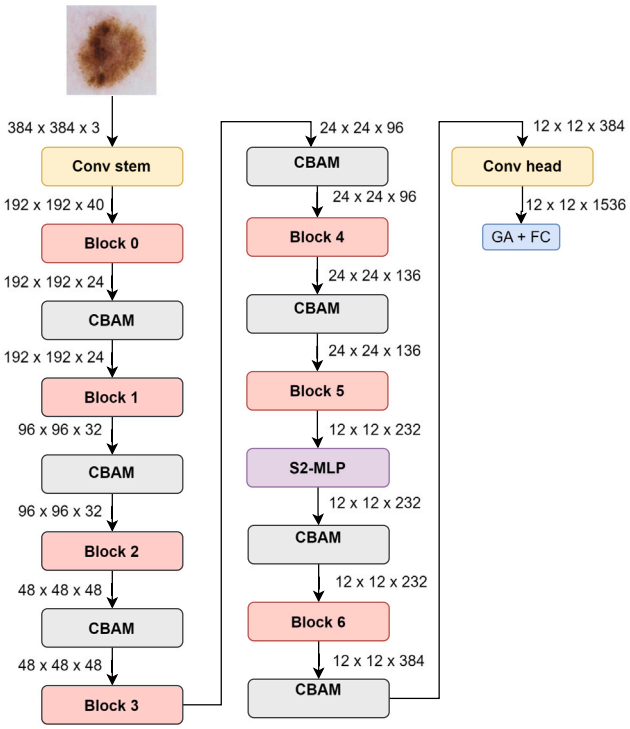

Fig. 8. S2-CBAM-B3: Attention-based model combining S2-MLP with Convolutional Block Attention Module (CBAM) blocks.

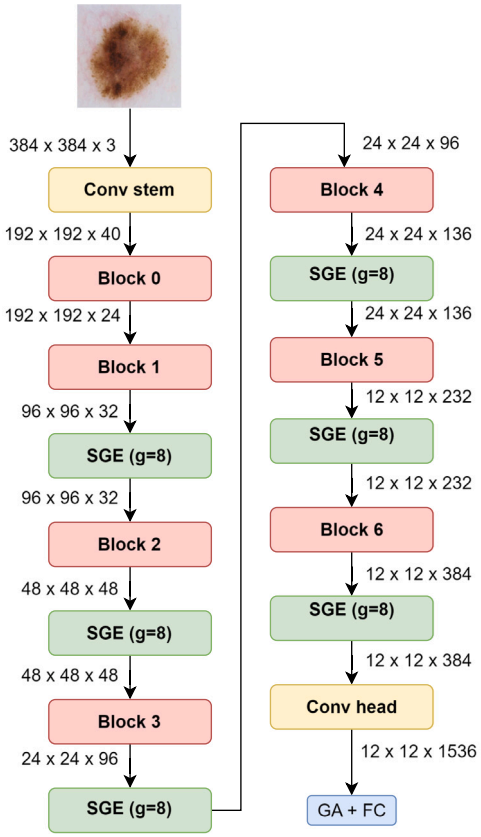

Fig. 7. SGE-B3: Attention-based model using Spatial Group-wise Enhance (SGE) blocks.

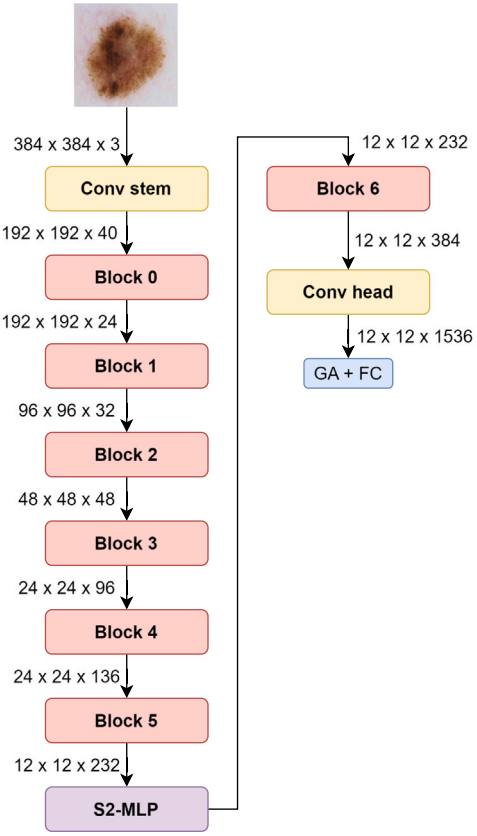

Fig. 9. S2-MLP-B3: Attention-based model using Spatial-shift (S2) MLP blocks.

on their prevalence, which helps lessen the effects of imbalances and promotes more efficient learning across all classes.

Focal loss prioritizes hard (difficult to classify) samples, which are frequently found in minority classes, while LDAM assigns weights based on the class distribution of the dataset. To take advantage of both features, Sadi et al. [40] proposed the Large Margin-aware Focal (LMF) loss. This loss function is a linear combination of Focal loss and LDAM, where the weights are determined by the hyperparameters α and β. Following various experiments, we chose α and β both equal to 0.2. Consequently, the LMF loss function employed in our study is formulated as follows:

$$L_{\mathrm{LMF}} = 0.2 \times (L_{\mathrm{LDAM}} + L_{\mathrm{Focal}}) \tag{4}$$

In this work, we conducted experiments with three different loss functions: weighted cross entropy, Focal loss (gamma = 2), and LMF loss, with the aim of determining the most effective loss function for our dataset. The outcomes, detailed analysis, and findings of these experiments are presented in Section 4.

4. Experimental results

The aim of this section is to provide a thorough explanation of the experiments carried out in our study. First, we detail the specifics of our training and testing datasets, before going on to describe the training settings. Lastly, we present the results of these experiments in detail.

4.1. Datasets

The International Skin Imaging Collaboration (ISIC) was a project that aimed to improve the detection and prevention of melanoma using digital skin imaging [41]. ISIC has organized multiple machine learning challenges to encourage the development of models that could accurately classify melanoma from images.

The ISIC 2019 dataset included 25,331 dermoscopic images of skin lesions and categorized them into the following nine diagnostic classes [42]: melanoma (MEL); melanocytic nevus (NV); basal cell carcinoma (BCC); actinic keratosis (AK); benign keratosis (BKL); dermatofibroma (DF); vascular lesion (VASC); squamous cell carcinoma (SCC); and none of the above. The dataset was used for the ISIC 2019 Challenge, which involved classifying skin lesions with or without metadata. This dataset included all the datasets from the previous ISIC challenges [21,43,44].

The ISIC 2020 dataset was composed of 33,126 images of different resolutions for training and 10,982 for the test set [45]. The training set comprised MEL, BKL, NV and unknown (UNK) benign samples. A total of 2,056 patients were included in this dataset from various locations around the world. In the SIIM-ISIC Melanoma Classification Challenge, competitors were asked to build models to identify melanoma using the dataset and the metadata.

To develop our model, we used the training datasets sourced from both ISIC 2020 and ISIC 2019, combining them to construct our training and validation sets. Our curated dataset ultimately comprised 58,457 images from ISIC 2019 and ISIC 2020, as summarized in Table 2.

From Fig. 10 it is evident that our dataset exhibited a significant imbalance, with a higher percentage of data representing nevus and unknown samples than other classes. To address this imbalance and promote robust evaluation while minimizing bias, we implemented a k-fold cross-validation strategy by dividing the data into five folds, thus ensuring an equal distribution of each class across all folds. During each training, we used one fold as the validation set and the rest as the training set.

To assess the real-world performance of our model, we conducted testing on two distinct datasets. This was done to ensure that the model could effectively handle unique scenarios that it had not encountered in previous learning phases. The first dataset employed for testing was the PROVe-AI set, recently sourced from the ISIC archive [13]. As Table 3 shows, this dataset comprised 603 dermoscopic image samples, with 95 instances identified as melanomas. We compared the labels of the datasets to our predefined classes, categorizing five images with labels that did not match ours as 'other'. The dataset included 603 images assigned across nine classes. The second test set was collected from 623 patients at Hospital Italiano de Buenos Aires (HIBA). As presented in Table 4, this dataset consisted of 1,616 images (1,270 dermoscopy and 346 clinical images) including 253 melanomas [46].

Table 2
Our training and validation dataset details. NV: nevus; MEL: melanoma; BCC: basal cell carcinoma; BKL: benign keratosis; AK: actinic keratosis; SCC: squamous cell carcinoma; DF: dermatofibroma.

Class Label	Samples
UNK	27,125
NV	18,069
MEL	5106
BCC	3323
BKL	2847
AK	867
SCC	628
VASC	253
DF	239
Total	**58457**

Table 3
PROVe-AI dataset details, used as the first test set.

Class Label	Samples
NV	340
BKL	110
MEL	95
AK	19
SCC	13
DF	11
BCC	9
VASC	1
Other	5
Total	**603**

Table 4
HIBA dataset details, used as the second test set.

Class Label	Samples
NV	602
BKL	88
MEL	253
AK	63
SCC	158
DF	61
BCC	340
VASC	51
Total	**1616**

4.2. Experimental setup

During our experiments, we maintained a consistent experimental set-up to ensure reproducibility and comparability across different model evaluations. After careful hyper-parameter tuning, the AdamW optimizer was employed as the optimization algorithm, with an initial learning rate set to 3×10^{-5}. To enhance training dynamics, we used a combination of the Cosine Annealing Warm Restarts and Gradual Warm-up Scheduler techniques.

For image processing, a uniform image size of 384 pixels was adopted across all experiments, including testing. The weight decay parameter was set to 10^{-5} to regulate model training and prevent

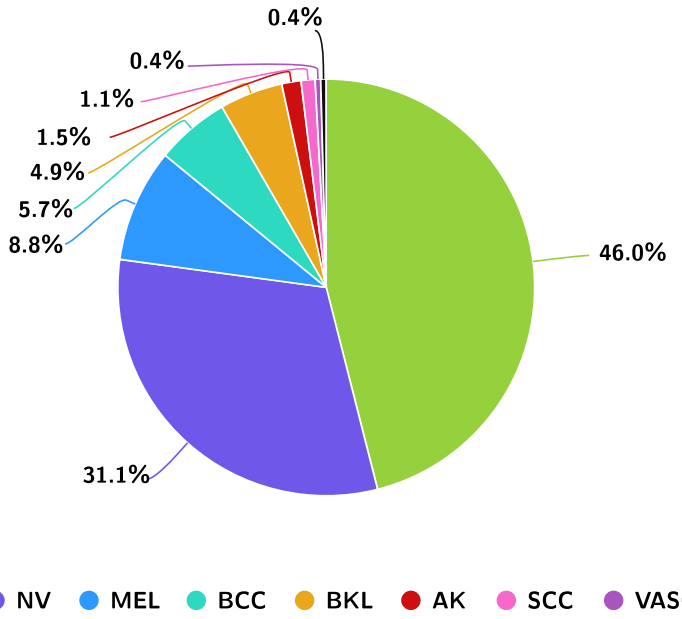

Fig. 10. Pie-chart illustrating train and validation dataset class distribution. From the chart, it is evident that the dataset is highly imbalanced.

overfitting. A batch size of 64 was chosen to balance computational efficiency and model convergence.

For each model, we imported EfficientNet-B3 architecture with pre-trained weights trained on the ImageNet database from the Geffnet [47] library. This model possesses a total of 10.7 million parameters. Our final models were multi-class classifiers with nine output channels (predicted labels). However, the multi-class labels were only used to provide label granularity and enhance diagnosis capabilities of the models [48]. For calculation of performance metrics and reporting, we only used the predicted probabilities for the melanoma class. Each experiment was trained for 35 epochs, allowing the models to undergo comprehensive training cycles. During training and validation iterations, the best model based on the evaluation metrics for the validation set was used to calculate the final results. Each prediction in the validation and test phase was performed ten times and the final probability was averaged. This standardized setup ensured a fair and systematic exploration of the chosen models and loss functions. The aim of maintaining a consistent configuration was to facilitate a meaningful comparison of results and draw reliable conclusions from the conducted experiments. The experiments were carried out on an NVIDIA MIG GPU instance with 40 GB of memory capacity and an average runtime of 24 hours for each model.

4.3. Results

In this section, we present a thorough analysis of the results of our experiments. In an ideal scenario, accuracy serves as a reliable metric for assessing model performance, since it directly measures the ratio of correctly predicted instances to the total instances in a dataset. However, when dealing with imbalanced datasets, accuracy provides a misleading view of the model's effectiveness. The F1-score would be a preferred metric to assess model performance when the distribution of classes is highly skewed. However, one drawback of the F1-score is its reliance on the default threshold of 0.5 for precision and recall assessment. In cancer detection tasks such as melanoma detection, the consequences of missing positive cases (false negatives, henceforth FNs) are often more severe than misclassifying negative cases (false positives, henceforth FPs). Therefore, a higher emphasis should be placed on minimizing FNs, even if this leads to an increase in FPs.

The Receiver Operating Characteristic (ROC) curve and associated Area Under the Curve (ROC-AUC) are widely used in melanoma detection and various medical diagnostic applications. The threshold independence of ROC-AUC proves vital in medical scenarios where the optimal decision threshold may vary based on the doctor's desired balance between sensitivity and specificity. On the other hand, when evaluating imbalanced medical datasets, a study conducted by Devries et al. [49] found that ROC-AUC tends to provide overly optimistic results. In contrast, PR-AUC demonstrates resilience to class imbalance and is considered a valuable alternative metric to ROC-AUC in such scenarios [50–52]. We chose to use both metrics in our evaluation to provide an in-depth understanding of the models' performance.

Table 5
Results of EfficientNet-B3 model trained using different loss functions, evaluated on the validation set and averaged for 5-fold in the melanoma classification task. LMF: Large Margin-aware Focal loss; WCE: weighted categorical cross-entropy loss.

Loss Function	Mean ROC-AUC	Mean PR-AUC
LMF	**0.96 ± 0.006**	**0.85 ± 0.009**
WCE	0.95 ± 0.009	0.82 ± 0.011
Focal	0.94 ± 0.008	0.79 ± 0.021

In addition to the above, when developing deep models for cancer detection, it is vital to use assessment metrics that consider the clinical perspective. Physicians often prioritize sensitivity and specificity, as these methods provide a balanced assessment of the model's ability to correctly identify true positives (TP) and exclude true negative (TN) cases, which are crucial in making reliable clinical decisions. In addition, high sensitivity is essential to cancer diagnosis because it ensures accurate patient identification with cancer, which lowers the rate of FNs. For this reason, we also calculated sensitivity and specificity metrics with a fixed threshold of 95% for sensitivity when evaluating the model. In the rest of this section, we examine the results in detail and explain the implications of our research findings.

Firstly, we trained three EfficientNet-B3 multi-class classifiers (starting from pre-trained ImageNet weights) using the three loss functions described in Section 3.4. We then analyzed how these loss functions influenced the performance and outcomes of our experiments during validation. As summarized in Table 5, LMF loss consistently enhanced the ROC-AUC and PR-AUC of the melanoma class during the validation phase. We, therefore, selected this loss function as the default choice.

Subsequently, we proceeded to train and validate four attention-based models, as outlined in Section 3.3, using LMF as our chosen loss function. The trained models were subsequently subjected to testing on

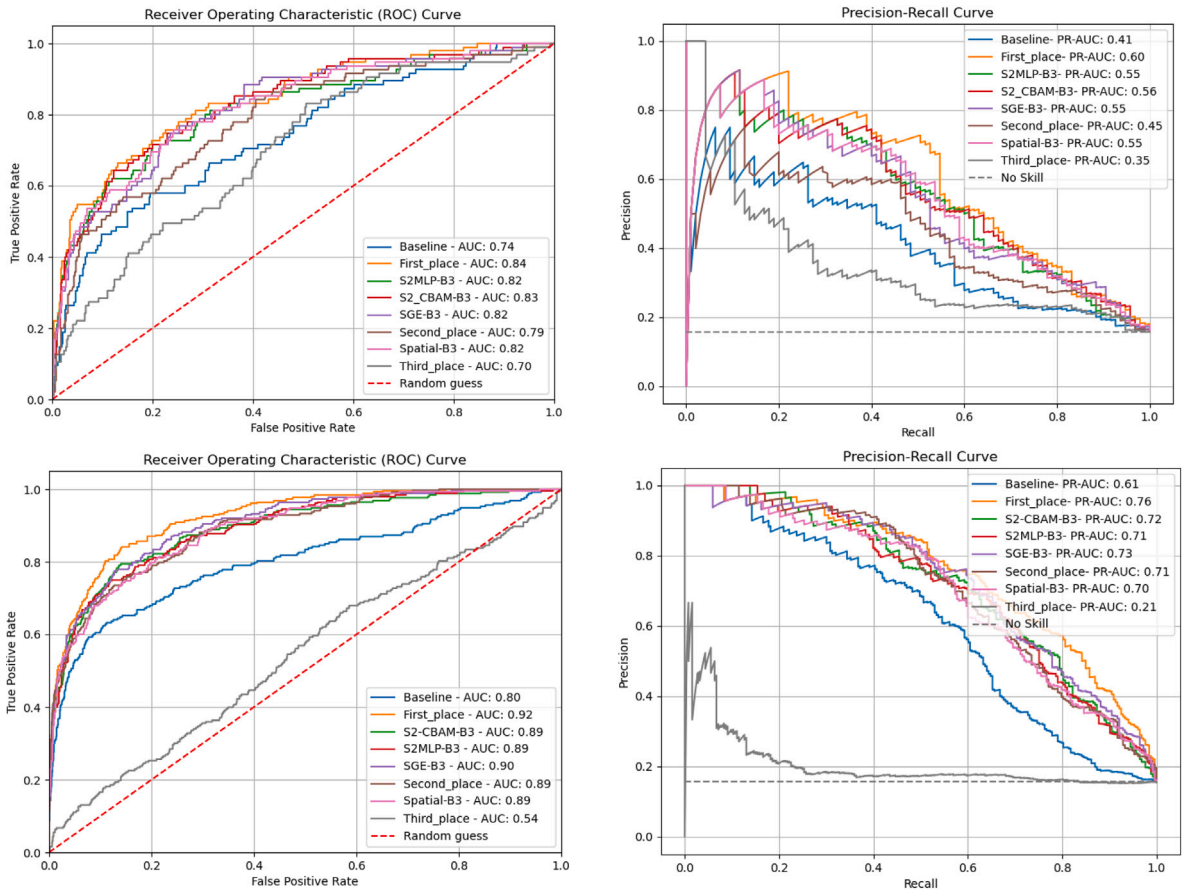

Fig. 11. Comparison of our attention-based models with the top three prize winners of the ISIC 2020 Melanoma Detection Challenge. Top row: results on the PROVe-AI test set. Bottom row: results on the HIBA test set.

a distinct dataset. To demonstrate the capabilities of our models, we conducted a thorough comparison of our results with existing models, benchmarking our work against the top three performers (on the private leaderboard) in the ISIC 2020 melanoma classification challenge. To ensure a fair and precise evaluation, we used the code that had been publicly shared by the respective researchers [13,53,54]. Furthermore, we employed the original model weights they had made available and assessed the performance of their models on our test set using test time augmentation. To better demonstrate the performance of our attention-based models, we also trained a baseline EfficientNet-B3 model using the original hyper-parameters [37] and evaluated its performance. Our analysis included ROC-AUC, PR-AUC, sensitivity, and specificity. The ROC curve and the precision–recall curves of all models are depicted in Fig. 11.

The first-placed prize winner [13] employed an ensemble strategy incorporating 18 models, ranging from EfficientNets-B4 to B7, Squeeze-and-Excitation ResNext101 [15] and ResNeSt101 [55]. The cumulative number of parameters for this ensemble amounted to 644 million. The second-placed prize winner [53] adopted a similar ensemble approach, using three models – an EfficientNet-B7 and two EfficientNet-B6 – with a total of 152 million parameters. Finally, the third-placed winner [54] opted for an ensemble of eight models, all based on EfficientNet-B6, giving rise to 344 million parameters. Notably, each winner trained diverse models with varied image sizes, implementing a 5-fold cross-validation methodology.

To maintain consistency across experiments during the testing phase, we averaged the predicted probabilities from all five folds to generate a single final set of predictions, following the method used by the ISIC winners. The outcomes of this comparative evaluation using the first test set (PROVe-AI) with TTA are outlined in Table 6. Our best-performing model (S2-CBAM-B3) achieved an ROC-AUC of 0.83 and a PR-AUC of 0.56 in detecting melanoma. All of our proposed models achieved a significant improvement in ROC-AUC (from 0.79 to 0.83, all $p < 0.001$) and PR-AUC (from 0.45 to 0.56) compared to the second and third-placed prize winners of the ISIC challenge, while using 92%–96% fewer parameters. Even though our models did not outperform the first-placed winner, they did achieve comparable results, with an average ROC-AUC of 0.82 and PR-AUC of 0.55 using 98% fewer parameters.

In addition, we computed the sensitivity and specificity for all the models using a predetermined threshold of 95% for sensitivity. All of our models attained a sensitivity of 0.96 (95% CI: 0.94–0.97) at the given threshold and outperformed the second and third-placed prize winners with an average specificity of 0.31 (95% CI: 0.10–0.53). Furthermore, our best model (S2-CBAM-B3) achieved a specificity of 0.36 (95% CI: 0.13–0.60). During our experiments, we observed that specificity was considerably low for all models. To better understand the relationship between low specificity and fixed sensitivity thresholds, we tested various thresholds and compared the resulting specificity (results presented in Table 7). This experiment showed how higher sensitivity values lead to lower specificity rates across all models. Additionally, at lower sensitivity thresholds, our models achieved higher specificity compared to the first-place winner.

To further assess the generalizability of our models, we validated their performance on a second dataset without TTA. The diversity of imaging techniques and skin lesions in the HIBA dataset provides an ideal basis for assessing the models across a range of clinical scenarios, thereby strengthening their applicability to real-world settings. As shown in Table 8, our models outperformed the second- and third-place winners on the second test set as well (all $p < 0.001$). Our top-performing model (SGE-B3) achieved a ROC-AUC of 0.91 (95% CI: 0.89–0.93) and a PR-AUC of 0.73 (95% CI: 0.68–0.78), demonstrating

Table 6
Comparing the performance of our proposed models with the top three prize winners of the ISIC 2020 Melanoma Detection Challenge on the PROVe-AI test set. Sensitivity and specificity were calculated with a fixed threshold of 95% for sensitivity. Data in parenthesis correspond to 95% confidence interval.

Model	#Parameters	Image Size	ROC-AUC	PR-AUC	Sensitivity	Specificity
First place [13]	644M	384–896	**0.84 (0.80–0.89)**	**0.60 (0.49–0.71)**	0.96 (0.94–0.98)	**0.37** (0.20–0.56)
S2-CBAM-B3 (ours)	11M	384	0.83 (0.79–0.88)	0.56 (0.46–0.67)	0.96 (0.94–0.98)	0.36 (0.13–0.60)
SGE-B3 (ours)	10.7M	384	0.82 (0.78–0.87)	0.55 (0.45–0.65)	0.96 (0.94–0.98)	0.31 (0.02–0.59)
Spatial-B3 (ours)	11.7M	384	0.82 (0.78–0.87)	0.55 (0.45–0.65)	0.96 (0.94–0.98)	0.31 (0.10–0.55)
S2-MLP-B3 (ours)	11M	384	0.82 (0.77–0.87)	0.55 (0.44–0.66)	0.96 (0.94–0.98)	0.27 (0.12–0.43)
Baseline (EfficientNet-B3)	10.7M	300	0.74 (0.68–0.80)	0.41 (0.32–0.53)	0.96 (0.94–0.98)	0.19 (0.06–0.31)
Second place [53]	152M	512–640	0.79 (0.73–0.84)	0.45 (0.35–0.57)	0.96 (0.94–0.98)	0.24 (0.02–0.46)
Third place [54]	344M	256–768	0.70 (0.64–0.76)	0.35 (0.26–0.45)	0.96 (0.94–0.98)	0.19 (0.01–0.44)

Table 7
Specificity results for different models across various pre-defined sensitivity thresholds, evaluated on the PROVe-AI test set.

Model	95%	94%	93%	92%	91%	90%	89%	88%	87%	86%	85%
First Place [13]	**0.37**	**0.41**	**0.44**	**0.47**	0.49	0.50	0.53	0.55	0.56	0.58	0.60
S2-CBAM-B3 (ours)	0.36	0.40	**0.44**	0.46	0.49	0.51	0.53	0.55	0.58	0.60	0.61
SGE-B3 (ours)	0.31	0.36	0.42	**0.47**	**0.50**	**0.53**	**0.57**	**0.58**	**0.60**	**0.61**	**0.62**
Spatial-B3 (ours)	0.31	0.35	0.39	0.43	0.46	0.48	0.51	0.54	0.56	0.57	0.60
S2-MLP-B3 (ours)	0.27	0.29	0.32	0.35	0.37	0.40	0.44	0.46	0.51	0.53	0.56
Baseline (EfficientNet-B3)	0.19	0.22	0.24	0.27	0.30	0.32	0.35	0.37	0.39	0.41	0.43
Second place [53]	0.24	0.28	0.33	0.37	0.41	0.43	0.47	0.49	0.52	0.54	0.56
Third place [54]	0.19	0.23	0.28	0.31	0.33	0.35	0.38	0.39	0.41	0.43	0.44

Table 8
Comparing the performance of our proposed models with the top three prize winners of the ISIC 2020 Melanoma Detection Challenge on the HIBA test set. Sensitivity and specificity were calculated with a fixed threshold of 95% for sensitivity. Data in parenthesis correspond to 95% confidence interval.

Model	ROC-AUC	PR-AUC	Sensitivity	Specificity
First place [13]	**0.92 (0.91–0.94)**	**0.76 (0.71–0.80)**	**0.95 (0.95–0.96)**	**0.62 (0.53–0.72)**
SGE-B3 (ours)	0.91 (0.89–0.93)	0.73 (0.68–0.78)	0.95 (0.95–0.96)	0.57 (0.48–0.66)
S2-CBAM-B3 (ours)	0.89 (0.87–0.92)	0.72 (0.67–0.77)	0.95 (0.95–0.96)	0.48 (0.34–0.63)
Spatial-B3 (ours)	0.89 (0.87–0.91)	0.70 (0.64–0.75)	0.95 (0.95–0.96)	0.51 (0.41–0.61)
S2-MLP-B3 (ours)	0.89 (0.97–0.92)	0.72 (0.66–0.76)	0.95 (0.95–0.96)	0.52 (0.43–0.61)
Baseline (EfficientNet-B3)	0.54 (0.77–0.84)	0.25 (0.55–0.67)	0.95 (0.95–0.96)	0.17 (0.02–0.26)
Second place [53]	0.89 (0.87–0.91)	0.71 (0.66–0.77)	0.95 (0.95–0.96)	0.46 (0.36–0.56)
Third place [54]	0.54 (0.50–0.58)	0.21 (0.18–0.26)	0.95 (0.95–0.96)	0.02 (0.16–0.42)

Table 9
Comparison of average prediction time for one batch of 64 images and model sizes for our models versus the top three ISIC 2020 prize winners. This result is compiled across 5 folds. s: seconds, GB: gigabytes.

Model	Time (s)	Model Size (GB)
SGE-B3 (ours)	0.33 ± 0.001	0.20
S2-CBAM-B3 (ours)	0.33 ± 0.061	0.21
Spatial-B3 (ours)	0.34 ± 0.015	0.20
S2-MLP-B3 (ours)	0.35 ± 0.035	0.21
Baseline (EfficientNet-B3)	0.33 ± 0.002	0.20
Second place [53]	64.60 ± 1.07	2.75
Third place [54]	161.0 ± 2.10	4.69
First place [13]	183.80 ± 3.21	11.60

strong performance, though still modestly below that of the first-place winner. This consistent trend in results further underscores the robustness of our models.

On the other hand, it should be noted that the small number of parameters enabled our models to achieve significantly faster execution time. As Table 9 illustrates, our models demonstrated a significant efficiency advantage, running 20 times faster than the model that won the ISIC 2020 Challenge and therefore offering a more practical solution for real-time diagnostic application in clinical settings. The enhanced efficiency and compactness of our models also render them ideally suited for integration into smartphone-based dermoscopy systems, bridging the gap between advanced diagnosis and mobile healthcare solutions.

Furthermore, we used GradCAM [56] to provide a closer look at how the models pay attention to specific parts of input images. This facilitated the creation of visualizations highlighting the areas of focus for the model. Through these illustrations, it became evident that our attention-based models exhibit a superior capability in accurately detecting lesion areas with greater confidence when contrasted with the performance of the baseline models (see Fig. 12).

5. Discussion

This study aimed to address three significant challenges in the automated detection of skin lesions: (a) addressing data imbalance; (b) enhancing model robustness; and (c) managing model parameters and computational costs. To overcome data imbalance, we tested three different loss functions designed to deal with this type of dataset, with LMF loss demonstrating superior performance compared to the alternatives. However, our results revealed a persistent bias towards the majority class, despite the improvement in overall model performance attributed to LMF loss. This highlights the ongoing need for additional publicly available data in the field to further enhance model generalization.

To guarantee the robustness of our model, we created a multi-class classifier that can distinguish melanoma from other types of cancer and pre-cancerous lesions. Instead of evaluating our models on a subset of the same dataset, we used two entirely different datasets as our test sets to independently assess their real-world performance. Additionally, we employed relevant metrics to conduct a comprehensive and detailed evaluation of our models, further ensuring a thorough assessment of their performance.

To overcome the third challenge, we introduced four attention-based models by integrating attention blocks at various levels within the EfficientNet-B3 architecture. This innovative approach enabled us to create small yet effective models for practical melanoma diagnosis,

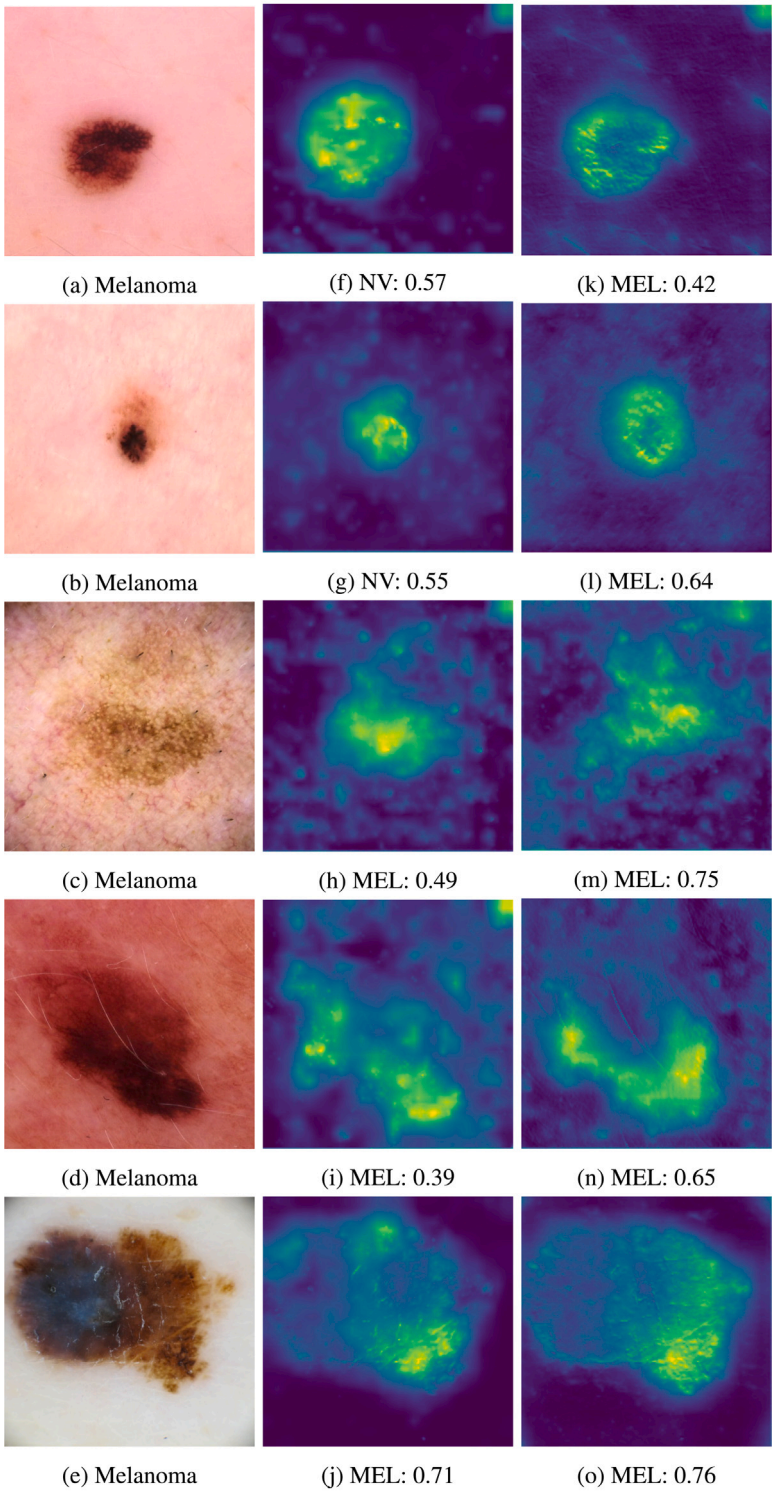

Fig. 12. GradCAM visualizations of our models with predicted labels and prediction confidence. (a, b, c, d, e): melanoma image samples; (f, g, h, i, j): baseline EfficientNet-B3; (k, l, m, n, o): attention-based model. These visualizations revealed that our attention-based models excel over the baseline model in precisely identifying areas of lesions with greater confidence.

outperforming the second and third-placed prize winners of the ISIC Challenge. Furthermore, our top-performing model demonstrated comparable performance in all metrics to the first-placed winner of the ISIC Challenge. Our compact and robust models can provide invaluable assistance to doctors by enabling early and rapid melanoma detection. In addition, their adaptability allows seamless integration with widely available devices, such as smartphones. This not only can help reduce patient waiting time and mortality rates but also improve accuracy and efficiency for healthcare professionals, especially general practitioners.

As discussed earlier in this study, sensitivity is a crucial metric in cancer detection. However, a well-known limitation is that increasing sensitivity leads to a decrease in specificity. In our case, all models (ours and third parties) obtained a very low specificity. This may be due to several factors, including using totally independent test sets,

the diversity of skin lesions and tumors within the unknown class, the limited data available for benign classes (e.g. DF and VASC), and the restrictive sensitivity threshold set at 95%. In many papers in the literature, we observed that results with very high sensitivity and specificity often occur when there is a correlation between the training and test sets. Specifically, when the training and test sets are subsets of the same dataset, the sensitivity and specificity can be unusually high [57]. However, when the test set is completely independent, a trade-off between sensitivity and specificity typically arises and clinicians often prefer higher sensitivity at the expense of lower specificity.

6. Conclusion

Artificial intelligence and deep learning models have proved to be successful in detecting and diagnosing skin lesions, offering valuable assistance to general practitioners and dermatologists by enhancing their diagnostic capabilities. Despite their effectiveness, many top-performing models in the field tend to be large and computationally expensive, posing challenges for practical real-world applications, such as deployment on mobile devices for smartphone-based dermoscopy. Moreover, the robustness of these models is rarely tested on independent datasets, complicating the selection of appropriate models for future use in a real environment.

In this work, we proposed four lightweight attention-based models based on the EfficientNet-B3 backbone, designed for the precise diagnosis and classification of skin lesions. Our models demonstrated notable performance in the melanoma detection task, exhibiting significant improvements in ROC-AUC and PR-AUC compared to the second and third-placed prize winners of the ISIC challenge, while using 92%–96% fewer parameters. Although our models did not improve upon the first-placed prize winner, they did achieve comparable results on both test sets while using 98% fewer parameters. On the PROVe-AI test set, our best model (S2-CBAM-B3) achieved a ROC-AUC of 0.83, a PR-AUC of 0.56, a sensitivity of 0.96 (95% CI: 0.94–0.97) and a specificity of 0.36 (95% CI: 0.13–0.60) at a predetermined threshold of 95% for sensitivity. On the HIBA dataset, our best model (SGE-B3) achieved a ROC-AUC of 0.91, a PR-AUC of 0.73, a sensitivity of 0.95 (95% CI: 0.95–0.96) and a specificity of 0.57 (95% CI: 0.48–0.66), demonstrating its capability to detect true positive cases.

The core objective of our research was to highlight that smaller, well-tuned models with enhanced architecture can achieve competitive results without resorting to large ensembles of complex models. We believe the attention mechanisms explored in this paper have the potential to be applied to more sophisticated models in the future, leading to highly accurate performance and potentially replacing the need for large ensembles altogether.

During our research, we observed a significant data imbalance in publicly available datasets. Despite employing approaches to mitigate this, such as using a specific loss function, our models still tended to overfit to the majority class. Hence, there is a crucial need to acquire more data for the underrepresented classes to facilitate model improvement in the future. One potential approach could be to create synthetic malignant images using Generative Adversarial Networks (GANs). Moreover, certain classes in the ISIC datasets contain a wide variety of lesions, with different shapes and colors. This intra-class diversity can pose challenges and create potential confusion for deep models. It would therefore be interesting and valuable to conduct further investigations that explore skin lesion classification at a sub-class level to address this complexity. Furthermore, even after applying preprocessing techniques, some images still contain artifacts that we could not remove without compromising important features. These artifacts introduce bias into the model, underscoring the need for methods that can effectively address them. Finally, while setting a predefined high sensitivity threshold ensures high accuracy in detecting true positives, real-world applications of these models require greater specificity. This would help reduce false positives and minimize unnecessary medical procedures.

CRediT authorship contribution statement

Sana Nazari: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Rafael Garcia:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially funded through the European Commission's Horizon 2020 program through the iToBoS project (grant number SC1-BHC-06-2020-965221) and the IFUDG 2022 grant. We have published our code in https://github.com/iToBoS/Attention-based-Melanoma-Classifiers.

References

[1] American Cancer Society, About melanoma skin cancer, 2023, https://www.cancer.org/content/dam/CRC/PDF/Public/8823.00.pdf. (Accessed on: 6.5.2024).
[2] M. Naqvi, S.Q. Gilani, T. Syed, O. Marques, H. Kim, Skin cancer detection using deep learning-A review, Diagnostics 13 (11) (2023) http://dx.doi.org/10.3390/diagnostics13111911.
[3] F. Grignaffini, L. Piazzo, M. Troiano, P. Simeoni, F. Mangini, G. Pellacani, C. Cantisani, F. Frezza, Machine learning approaches for skin cancer classification from dermoscopic images: A systematic review, Algorithms 15 (2022) http://dx.doi.org/10.3390/a15110438.
[4] D. Wen, S.M. Khan, A.J. Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C.d.B. Perez, A.K. Denniston, X. Liu, R.N. Matin, Characteristics of publicly available skin cancer image datasets: a systematic review, Lancet Digit. Health 4 (2022) http://dx.doi.org/10.1016/S2589-7500(21)00252-1.
[5] S. Khan, H. Ali, Z. Shah, Identifying the role of vision transformer for skin cancer—A scoping review, Front. Artif. Intell. 6 (2023) http://dx.doi.org/10.3389/frai.2023.1202990.
[6] A. Zawacki, B. Helba, G. Shih, J. Weber, J. Elliott, M. Combalia, N. Kurtansky, N. Codella, P. Culliton, V. Rotemberg, SIIM-isic melanoma classification, 2020, https://kaggle.com/competitions/siim-isic-melanoma-classification. (Accessed on: 6.5.2024).
[7] H. Balaha, A.E.S. Hassan, Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm, Neural Comput. Appl. 35 (2023) http://dx.doi.org/10.1007/s00521-022-07762-9.
[8] A.D. Bandy, Y. Spyridis, B. Villarini, V. Argyriou, Intraclass clustering-based CNN approach for detection of malignant melanoma, Sensors 23 (2023) http://dx.doi.org/10.3390/s23020926.
[9] C. Dong, D. Dai, Y. Zhang, C. Zhang, Z. Li, S. Xu, Learning from dermoscopic images in association with clinical metadata for skin lesion segmentation and classification, Comput. Biol. Med. 152 (2023) http://dx.doi.org/10.1016/j.compbiomed.2022.106321.
[10] M. Saeed, A. Naseer, H. Masood, S.U. Rehman, V. Gruhn, The power of generative AI to augment for enhanced skin cancer classification: A deep learning approach, IEEE Access 11 (2023) 130330–130344, http://dx.doi.org/10.1109/ACCESS.2023.3332628.
[11] S.M. Jaisakthi, P. Mirunalini, C. Aravindan, R. Appavu, Classification of skin cancer from dermoscopic images using deep neural network architectures, Multimedia Tools Appl. 82 (10) (2023) 15763–15778, http://dx.doi.org/10.1007/s11042-022-13847-3.
[12] M.M. Mijwil, Skin cancer disease images classification using deep learning solutions, Multimedia Tools Appl. 80 (2021) http://dx.doi.org/10.1007/s11042-021-10952-7.
[13] M.A. Marchetti, E.A. Cowen, N.R. Kurtansky, J. Weber, M. Dauscher, J. DeFazio, L. Deng, S.W. Dusza, H. Haliasos, A.C. Halpern, S. Hosein, Z.H. Nazir, A.A. Marghoob, E.A. Quigley, T. Salvador, V.M. Rotemberg, Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study), Npj Digit. Med. 6 (1) (2023) 1–11, http://dx.doi.org/10.1038/s41746-023-00872-1.
[14] F. Wang, M. Jiang, C. Qia, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-January, 2017, pp. 6450–6458, http://dx.doi.org/10.1109/CVPR.2017.683.

[15] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/CVPR.2018.00745.

[16] S. Woo, J. Park, J. Lee, I.S. K, CBAM: Convolutional block attention module, in: Computer Vision – ECCV 2018, Springer International Publishing, 2018, pp. 3–19, http://dx.doi.org/10.48550/arXiv.1807.06521.

[17] J. Zhang, Y. Xie, Y. Xia, C. Shen, Attention residual learning for skin lesion classification, IEEE Trans. Med. Imaging 38 (2019) 2095–2103, http://dx.doi.org/10.1109/tmi.2019.2893944.

[18] X. He, Y. Wang, S. Zhao, C. Yao, Deep metric attention learning for skin lesion classification in dermoscopy images, Complex Intell. Syst. 8 (2022) http://dx.doi.org/10.1007/s40747-021-00587-4.

[19] W. Zenghui, L. Qiang, S. Hong, Dual attention based network for skin lesion classification with auxiliary learning, Biomed. Signal Process. Control 74 (2022) http://dx.doi.org/10.1016/j.bspc.2022.103549.

[20] A. Naveed, S. Naqvi, T.M. Khan, I. Razzak, PCA: Progressive class-wise attention for skin lesions diagnosis, Eng. Appl. Artif. Intell. 127 (2024) http://dx.doi.org/10.1016/j.engappai.2023.107417.

[21] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (2018) http://dx.doi.org/10.1038/sdata.2018.161.

[22] L. Tan, H. Wu, J. Xia, Y. Liang, J. Zhu, Skin lesion recognition via global-local attention and dual-branch input network, Eng. Appl. Artif. Intell. 127 (B) (2024) http://dx.doi.org/10.1016/j.engappai.2023.107385.

[23] A.N. Omeroglu, H.M.A. Mohammed, E.A. Oral, S. Aydin, A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification, Eng. Appl. Artif. Intell. 120 (2023) http://dx.doi.org/10.1016/j.engappai.2023.105897.

[24] F. Pahde, M. Dreyer, W. Samek, S. Lapuschkin, Reveal to revise: An explainable AI life cycle for iterative bias correction of deep models, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland, Cham, 2023, pp. 596–606, http://dx.doi.org/10.1007/978-3-031-43895-0_56.

[25] S.W. Pewton, M.H. Yap, Dark corner on skin lesion image dataset: Does it matter? in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2022, pp. 4830–4838, http://dx.doi.org/10.1109/CVPRW56347.2022.00530.

[26] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 60, http://dx.doi.org/10.1186/s40537-019-0197-0.

[27] M. Nanni, S. Brahnam, A. Lumini, Feature transforms for image data augmentation, Neural Comput. Appl. 34 (24) (2022) 22345–22356, http://dx.doi.org/10.1007/s00521-022-07645-z.

[28] H. Larochelle, G.E. Hinton, Learning to combine foveal glimpses with a third-order Boltzmann machine, in: Advances in Neural Information Processing Systems, vol. 23, 2010, URL https://proceedings.neurips.cc/paper_files/paper/2010/file/677e09724f0e2df9b6c000b75b5da10d-Paper.pdf.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, N. Uszkoreit, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR, 2021, http://dx.doi.org/10.48550/arXiv.2010.11929.

[30] J. Devlin, M. Chang, K. Lee, K. 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), 2018, pp. 4171–4186, http://dx.doi.org/10.18653/V1/N19-1423, arXiv preprint arXiv:1810.04805.

[31] A. Gupta, K. Rangarajan, Uncover this tech term: Transformers, Korean J. Radiol. 25 (1) (2024) 113–115, http://dx.doi.org/10.3348/kjr.2023.0948.

[32] H. Alhichri, A.S. Alswayed, Y. Bazi, N. Ammour, N.A. Alajlan, Classification of remote sensing images using EfficientNet-B3 CNN model with attention, IEEE Access 9 (2021) 14078–14094, http://dx.doi.org/10.1109/ACCESS.2021.3051085.

[33] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2016, http://dx.doi.org/10.48550/arXiv.1612.03928, arXiv preprint arXiv:1612.03928.

[34] Y. Tan, L. Xu, C. Yunfeng, S. Mingming, L. Ping, S^2-MLP: Spatial-shift MLP architecture for vision, 2021, https://arxiv.org/abs/2108.01072.

[35] Y. Tan, L. Xu, Y. C., S. Mingming, L. Ping, S^2-MLPv2: Improved spatial-shift MLP architecture for vision, 2021, https://arxiv.org/abs/2108.01072.

[36] L. Yuxuan, L. Xiang, Y. Jian, Spatial group-wise enhance: Enhancing semantic feature learning in CNN, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13845 LNCS, 2023, pp. 316–332, http://dx.doi.org/10.1007/978-3-031-26348-4_19.

[37] M. Tan, Q.V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, 2019, http://dx.doi.org/10.48550/arXiv.1905.11946, ArXiv arXiv:1905.11946.

[38] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2999–3007, URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8417976.

[39] C. Kaidi, W. Colin, G. Adrien, A. Nikos, M. Tengyu, Learning imbalanced datasets with label-distribution-aware margin loss, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, 2019, pp. 1565–1576, http://dx.doi.org/10.48550/arXiv.1906.07413.

[40] A.A. Sadi, L. Chowdhury, N.-W.-B. Jahan, M.N.S. Rafi, R. Chowdhury, F.A. Khan, N. Mohammed, LMFLOSS: A hybrid loss for imbalanced medical image classification, 2022, http://dx.doi.org/10.48550/arXiv.2212.12741, ArXiv arXiv:2212.12741.

[41] T.I.S.I. Collaboration, The international skin imaging collaboration, 2024, https://www.isic-archive.com/. (Accessed on: 6.5.2024).

[42] ISIC 2019 Challenge, ISIC 2019 challenge, 2024, https://challenge.isic-archive.com/landing/2019/. (Accessed on: 6.5.2024).

[43] N.C.F. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. 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), 2018, pp. 168–172, http://dx.doi.org/10.1109/ISBI.2018.8363547.

[44] C. Hernández-Pérez, M. Combalia, S. Podlipnik, N.C.F. Codella, V. Rotemberg, A.C. Halpern, O. Reiter, C. Carrera, B. Helba, S. Puig, V. Vilaplana, J. Malvehy, BCN20000: Dermoscopic lesions in the wild, Sci. Data 11 (641) (2024) http://dx.doi.org/10.1038/s41597-024-03387-w.

[45] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, P. Soyer, A patient-centric dataset of images and metadata for identifying melanomas using clinical context, Sci. Data 8 (2021) 34, http://dx.doi.org/10.1038/s41597-021-00815-z.

[46] M.A.R. Lara, M.V.R. Kowalczuk, M.L. Eliceche, M.G. Ferraresso, D.R. Luna, S.E. Benitez, L.D. Mazzuoccolo, A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population, Sci. Data 10 (1) (2023) http://dx.doi.org/10.1038/s41597-023-02630-0.

[47] R. Wightman, Gen-efficientnet-pytorch: PyTorch implementation of GenEfficientNet, 2019, https://github.com/rwightman/gen-efficientnet-pytorch. (Accessed on: 6.5.2024).

[48] Z. Chen, R. Ding, T. Chin, D. Marculescu, Understanding the impact of label granularity on CNN-based image classification, in: 2018 IEEE International Conference on Data Mining Workshops, ICDMW, 2018, pp. 895–904, http://dx.doi.org/10.1109/ICDMW.2018.00131.

[49] Z. Devries, E. Locke, M. Hoda, D. Moravek, K. Phan, A. Stratton, S. Kingwell, E.K. Wai, P. Phan, Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability, SPINE J. 21 (7) (2021) 1135–1142, http://dx.doi.org/10.1016/j.spinee.2021.02.007.

[50] H.R. Sofaer, J.A. Hoeting, C.S. Jarnevich, The area under the precision-recall curve as a performance metric for rare binary events, Methods in Ecol. Evol. 10 (4) (2019) 565–577, http://dx.doi.org/10.1111/2041-210X.13140.

[51] G. Fu, L. Yi, J. Pan, Tuning model parameters in class-imbalanced learning with precision-recall curve, Biometrical J. 61 (3, SI) (2019) 652–664, http://dx.doi.org/10.1002/bimj.201800148.

[52] B. Ozenne, F. Subtil, D. Maucort-Boulch, The precision recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases, J. Clin. Epidemiol. 68 (8) (2015) 855–859, http://dx.doi.org/10.1016/j.jclinepi.2015.02.010.

[53] I. Pan, SIIM-ISIC melanoma classification: 2nd place, 2020, https://github.com/i-pan/kaggle-melanoma?tab=readme-ov-file. (Accessed on: 6.5.2024),

[54] C. Rota, Kaggle SIIM-isic melanoma classification: 3rd place solution overview, 2020, https://github.com/Masdevallia/3rd-place-kaggle-siim-isic-melanoma-classification. (Accessed on: 6.5.2024),

[55] Z. Hang, W. Chongruo, Z. Zhongyue, Z. Yi, Z. Haibin, S. Yue, H. Tong, M. Jonas, R. Manmatha, L. Mu, S. Alexander, ResNeSt: Split-attention networks, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2022-June, 2022, pp. 2735–2745, http://dx.doi.org/10.1109/CVPRW56347.2022.00309.

[56] R.S. Ramprasaath, C. Michael, D. Abhishek, V. Ramakrishna, P. Devi, B. Dhruv, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2020) http://dx.doi.org/10.1007/s11263-019-01228-7.

[57] R.H. Patel, E.A. Foltz, A. Witkowski, J. Ludzik, Analysis of artificial intelligence-based approaches applied to non-invasive imaging for early detection of melanoma: A systematic review, CANCERS 15 (19) (2023) http://dx.doi.org/10.3390/cancers15194694.