Interpretability of a Deep Learning Based Approach for the Classification of Skin Lesions into Main Anatomic Body Sites

Joanna Jaworek-Korjakowska [1,*] , **Andrzej Brodzicki** [1,*] , **Bill Cassidy** [2] , **Connah Kendrick** [2] and **Moi Hoon Yap** [2]

1 Department of Automatic Control and Robotics, AGH University of Science and Technology, 30-059 Kraków, Poland

2 Department of Computing and Mathematics, Manchester Metropolitan University, John Dalton Building, Chester Street, Manchester M1 5GD, UK; B.Cassidy@mmu.ac.uk (B.C.); Connah.Kendrick@mmu.ac.uk (C.K.); M.Yap@mmu.ac.uk (M.H.Y.)

* Correspondence: jaworek@agh.edu.pl (J.J.-K.); brodzicki@agh.edu.pl (A.B.); Tel.: +48-12-617-50-65 (J.J.-K.); +48-12-617-44-66 (A.B.)

Simple Summary: The detection of skin moles driven by current deep learning based approaches yields impressive results in the classification of malignant melanoma. It has been observed that the specific criteria for in situ and early invasive melanoma highly depend on the anatomic site of the body. To address this problem, we propose a deep learning architecture based framework to classify skin lesions into the three most important anatomic sites, including the face, trunk and extremities, and acral lesions. In this study, we take advantage of pretrained networks, we perform in depth analysis on database, architecture, and result regarding the effectiveness of the proposed framework. Experiments confirm the ability of the developed algorithms to classify skin lesions into the most important anatomical sites with 91.45% overall accuracy for the EfficientNetB0 architecture, which is a state-of-the-art result in this domain.

Abstract: Over the past few decades, different clinical diagnostic algorithms have been proposed to diagnose malignant melanoma in its early stages. Furthermore, the detection of skin moles driven by current deep learning based approaches yields impressive results in the classification of malignant melanoma. However, in all these approaches, the researchers do not take into account the origin of the skin lesion. It has been observed that the specific criteria for in situ and early invasive melanoma highly depend on the anatomic site of the body. To address this problem, we propose a deep learning architecture based framework to classify skin lesions into the three most important anatomic sites, including the face, trunk and extremities, and acral lesions. In this study, we take advantage of pretrained networks, including VGG19, ResNet50, Xception, DenseNet121, and EfficientNetB0, to calculate the features with an adjusted and densely connected classifier. Furthermore, we perform in depth analysis on database, architecture, and result regarding the effectiveness of the proposed framework. Experiments confirm the ability of the developed algorithms to classify skin lesions into the most important anatomical sites with 91.45% overall accuracy for the EfficientNetB0 architecture, which is a state-of-the-art result in this domain.

1. Introduction

During the last few years it has been widely observed that malignant melanoma, the deadliest form of skin cancer, is becoming increasingly aggressive due to a combination of environment, genetics, and lifestyle. Most skin cancer cases are related to ultraviolet (UV) light damaging the DNA in skin cells. The statistics released by the American Cancer Society are alarming. It is projected that the number of new melanoma cases will increase

by 5.8% in 2021 [1]. Furthermore, it is estimated that 207,390 cases of melanoma will be diagnosed in the U.S. in 2021, including 106,110 cases in situ (noninvasive) and 101,280 invasive cases, penetrating the epidermis into the skin's second layer. The staggering rates show that global action including redefining of medical diagnostic algorithms and early diagnosis and novel treatment methods are needed in order to achieve control of melanoma mortality rate reduction and prevention of severe cases.

The most widely used medical diagnostic algorithms include pattern analysis, the ABCD rule of dermoscopy, and the so-called seven-point checklist, which are based on a critical, simultaneous assessment of so-called dermoscopic criteria. Argenziano et al. confirmed that diagnostic algorithms improved the rate of diagnosing pigmented skin lesions by 10–30% [2]. However, due to the lack of access to large datasets, the algorithms have not been adapted and adjusted for skin changes depending on the place of origin. It has been observed that the criteria for melanoma in situ and early invasive melanoma is highly dependant on the anatomic site of the lesion origin for the three main anatomic sites including (1) trunk with extremities, (2) face, and (3) palms and soles (acral lesions) [2].

The currently proposed computer-aided methods have been designed to extract and calculate significant features based on the entire dermoscopic dataset and distinguish between benign and malignant skin lesions. However, when dealing with melanoma originating in different parts of the body, no detailed research studies have been published so far.

This study aims to perform an experimental study in order to determine the ability of algorithms to recognize the anatomical site based only on dermoscopic images. We propose a novel framework for distinguishing between pigmented skin lesions based on site-specific dermoscopic characteristics of skin lesions originating in different anatomic sites of the body. We achieve this goal with the application of pretrained convolutional neural networks (CNN), their interpretability, and connection to the domain knowledge.

The information about the body location of the analysed skin lesion can be exploited as an additional channel in the CNN based architecture or as a parameter determining the selection of the next step of the classification system in the two-stage decision making process. Furthermore, it can be very beneficial to add such an algorithm to prove whether the assigned location seems to be correct or not. During a body examination, several lesions are analyzed for one patient (sometimes even more than 20). There are systems that require marking the place of origin right after taking the medical image and those that require adding anatomical site annotation at a later stage, after registering all skin moles. It seems that the automatic checking of the origin of the skin mole can be valuable and result in more accurate detection of malignant melanoma. Moreover, automatic information about the place of origin of a section for the histopathological examination may also be helpful in assessing the lesion if it has not been provided at an earlier stage.

The novelty of this work can be summarized as follows:

- We present a new approach based on the adjusted pretrained EfficientNetB0 network architecture for the classification of skin moles into anatomic sites of the body, which confirms that melanoma-specific criteria occurring in particular sites enable differentiation between them.
- We compare the outcomes of state-of-the-art pretrained models including VGG19, ResNet50, Xception, DenseNet121, and EfficientNetB0. We visualize the feature distribution extracted by each architecture.
- We propose a new approach for model interpretability based on comparing Grad-CAM heatmaps with the segmentation ground-truth for assessing the skin lesion classification process.
- We compare and estimate the correlation between feature importance and domain knowledge.

1.1. Motivation and Clinical Definition

The main motivation to undertake this research is the difficulty observed in correct visual assessment of dermoscopic images by inexperienced dermatologists who typically achieve sensitivity and specificity at around 62–63% [2]. Furthermore, the varied appearance and relevance of melanoma-specific criteria present in skin lesions originating in different anatomic sites can cause serious problems during visual assessment. In recent years, the diagnostic criteria have been proposed and tested by several authors [3–5].

In Table 1, we present the most important melanoma-specific criteria for melanoma in situ and early invasive melanoma, which contribute to the diagnosis where the frequency of the criteria is >70% [2]. For thick and advanced melanomas, the preformed anatomic structures responsible for the site-specific dermoscopic appearance are already destroyed and are independent of the various sites. Table 1 shows the dermoscopic criteria that are commonly observed in skin lesions heavily dependent on the anatomic site of the body. For trunk and extremities, the more common melanoma-specific criteria include multi-component pattern and atypical pigment networks, in contrast to the face where reticular patterns and atypical pigment pseudonetworks are always present. For skin moles located on palms and soles, the presence of parallel-ridge patterns is considered highly important (Figure 1).

Table 1. Common melanoma-specific criteria for melanoma in situ and invasive melanoma detection according to the anatomic site of the body based on [2].

Anatomic Site	Criterion	Description	Frequency
Trunk, extremities	Multicomponent pattern	Combination of few dermoscopic structures	Very common
	Atypical pigment network	Irregular brown to black network	Very common
	Irregular dots and globules	Black or brown oval structures	Common
	Irregular streaks	Irregular linear structures	Common
	Irregular pigmentation	Pigmented areas with irregular size and distribution	Common
Face	Reticular pattern	Diffuse pigmentation of the erpidermis or papillary dermis	Always present
	Atypical pigment pseudonetwork	Advanced morphological structures by melanoma progression	Always present
Palms and soles	Parrallel-ridge pattern	Pigmentation along the cristae superficiales	Very common
	Irregular dots/globules	Black or brown oval structures	Common
	Irregular pigmentation	Pigmented areas with irregular size and distribution	Common
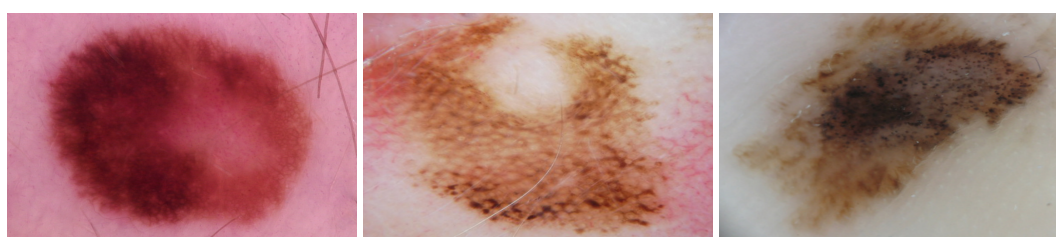
Figure 1. Dermoscopic images of in situ or early invasive melanomas presenting different dermoscopic features according to the anatomic site: (**left**) melanoma on the leg characterized by an atypical pigment network and irregular streaks, (**middle**) melanoma on the face characterized by reticular pattern, and (**right**) acral melanoma characterized by parallel-ridge pattern and irregular pigmentation.

1.2. Related Studies

In recent years, numerous clinical decision-support systems and computer-aided diagnostic systems have emerged for the automatic diagnosis of melanocytic lesions. These systems implement deep neural networks capable of classification of malignant and benign lesions. To the best of our knowledge, this study represents the first attempt to classify skin lesions into three main anatomic sites and proposes a new benchmark for the classification of skin lesions dedicated separately for each subtype. These subtypes include trunk with extremities, face, and palms and soles (acral lesions). However, we present the most recent studies concerning the classification of skin lesions from the respective anatomical regions.

Yu et al. [6] created a VGG-16 network trained on dermoscopic images of hands and feet consisting of acral melanoma and benign nevi confirmed by histopathological examination. This binary classification network demonstrated true positive, true negative, and area-under-the-curve measures similar to expert dermatologists and was able to outperform junior physicians. However, the dataset used was comparatively small—a total of 724 dermoscopic images consisting of 350 images of acral melanoma and 374 images of benign nevi.

Le et al. [7] devised a ResNet50 ensemble network for the classification of seven skin lesion types, including melanoma. This network used class weighing with a focal loss function to address the class imbalance of the HAM10000 dataset used for training their network. They achieved top-1, top-2, and top-3 accuracy, 93%, 97%, and 99%, respectively. This work observed that the gradual removal of the surrounding skin using U-Net segmentation resulted in increasingly reduced network performance. This suggests that the skin textures surrounding lesions are an important contributing factor to network accuracy and may be a vital pointer to any future networks trained to identify lesions by anatomical site.

Winkler et al. [8] investigated the diagnostic performance of FotoFinder Moleanalyzer Pro [9]—a commercially available CNN. Their experiment involved a binary classification (malignant/benign) for different melanoma localizations and subtypes using six dermoscopic datasets, which included melanomas of acral skin. This study noted that for acral melanomas, the system showed reduced sensitivity at high specificity.

Han et al. [10] created a localization network comprising a blob detector, a fine-image selector, and disease classifier. Their heterogeneous dataset comprised unprocessed photographs of malignant and benign lesions, which included lesions located on the head and neck. This study noted the limitations of using only dermoscopic images to train deep learning models that would be used in real-world settings due to the large number of complex shapes present on the human body, including acne and acne scars.

González-Cruz et al. [11] also noted limitations of datasets used in deep learning research for melanoma detection. They analyzed a dataset of 2849 high quality dermoscopic images of skin tumours to determine suitability for machine learning analysis. Their findings indicate that a large number of tumours located on the head, neck (76.8%), and trunk (>53.1%) had potential exclusion criteria due to absence of normal surrounding skin and pigmentation.

2. Database Specification

Nowadays, the most widely used dermoscopic skin lesion image database is the fourth ISIC dataset released by [12–14].

The ISIC 2019 dataset contains 33,569 dermoscopic images with patient metadata for the training set, indicating anatomical site of 22,700 lesions from a total of 25,331. Part of the ISIC 2019 dataset comprises the HAM10000 dataset, constituting the majority of dermoscopic images that are associated with the anatomical site. HAM10000 has been released by [12] and contains 11,526 dermoscopic images with metadata indicating anatomical site for 9781 lesions in the training set. The dataset contains 7222 dermoscopic images representing skin lesions originating in three different anatomic sites of the body including 6225 trunk/extremities, 702 face/head, and 295 acral lesions.

Due to highly imbalanced class composition, we augmented acral and face lesions by randomly applying image transformations such as rotation, sheer, and zoom. Each acral image was augmented 21 times, and each face image was augmented nine times, creating 6195 and 6318 artificial images, respectively. Augmentation was completed after we split the data into train, validation, and test subsets to avoid leaking information between subsets.

Data Visualization

In order to understand the distribution of the dataset, we visualize the data distribution of HAM10000 using two-dimensional reduction techniques—Uniform Manifold

Approximation and Projection (UMAP) [15] and the t-distributed Stochastic Neighbor Embedding technique (t-SNE) [16]. UMAP is a manifold learning technique for dimension reduction, and t-SNE is an unsupervised method that maps similarities between high-dimensional data into a probability distribution in such a manner that similar objects have a higher probability, minimizing the Kullback–Leibler divergence between the two distributions [16]. Figure 2 shows the visualisation of dataset distribution using UMAP and t-SNE and the relationship between anatomical sites of the body.
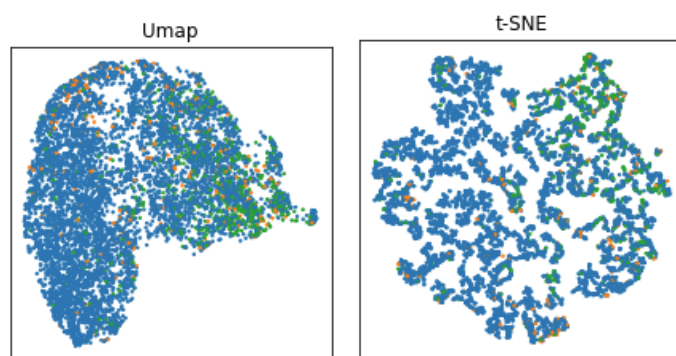
Figure 2. Visualization of HAM10000 dataset distribution on three anatomic sites with UMAP and t-SNE data transformation. The blue dots represent trunk and extremities, orange dots represent acral, and green dots represent face and head skin lesions.

We observe that skin lesions originating on the face form clusters of green dots while acral cases show irregular distribution. In order to analyze the datasets, we have calculated statistical metrics for (IntraC) intra-class and (InterC) inter-class ratio together with the ratio between InterC and IntraC (Ratio), computed using the Euclidean distance. Moreover, we analyzed the Silhouette Coefficient (*Silh.*), which is given by [17] as follows:

$$Silh. = \frac{b - a}{max(a, b)} \tag{1}$$

where *a* is the mean distance between a sample and all other points in the same class, and *b* is the mean distance between a sample and all other points in the next nearest cluster. The best value is 1 and the worst value is −1. Values near zero indicate overlapping clusters. Another relevant metric is the Calinski–Harabasz (CH) index, also known as Variance Ratio Criterion, and it represents the ratio of the sum of between-cluster dispersion and of within-cluster dispersion for all clusters within the dataset. The dispersion is given as the sum of distances squared [18]. Additionally, the Davies–Bouldin index has been calculated, which signifies the average similarity between clusters as a measure that compares the distance between clusters with the size of the clusters themselves and is defined as follows [19]:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{2}$$

where

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{3}$$

and s_i is the average distance between each point of cluster *i* and the centroid of that cluster, d_{ij} is the distance between cluster centroids, and *k* is the number of clusters.

In Table 2, we present the statistical analysis of the HAM10000 dataset in terms of the distribution of lesions regarding the anatomical site of the body. We observe that the complexity of the underlying classification task is very high and that regular machine learning algorithms will not be able to provide sufficient results. A high intra-class distance value indicates that cases are widely distributed in the space and hardly separable. However, as

the inter-class distance is higher, measuring the difference between two classes, it indicates the possibility of separating the data into anatomical sites.

Table 2. Statistical analysis of the dataset including calculations for the entire HAM10000 dataset regarding malignant and benign lesions as well as distribution in the acral and non-acral subsets.

Anatomic Site	# Total Nb.	#Melanoma Cases	Metrics				
			IntraC	InterC	Silh	CH	DB
Acral lesions	295	16	331.30	360.62	0.12	3.87	3.96
Face/head	702	102	317.65	324.655	0.03	6.75	6.84
Trunk/extremities	6225	490	338.26	356.12	0.12	90.27	4.57
HAM10000	7222	608	338.40	353.92	0.10	95.21	4.88

Furthermore, Figure 3 presents the distribution of melanocytic lesions within the disjoint dataset into the anatomic site. We observe that the red dots, representing malignant lesions, form areas and shapes that will be easier to separate than in the entire dataset. This is further confirmed by Table 2, which shows that the Silh. score and DB values indicate a better partition between trunk/extremities and the entire dataset.

(a)

(b)

(c)

(d)
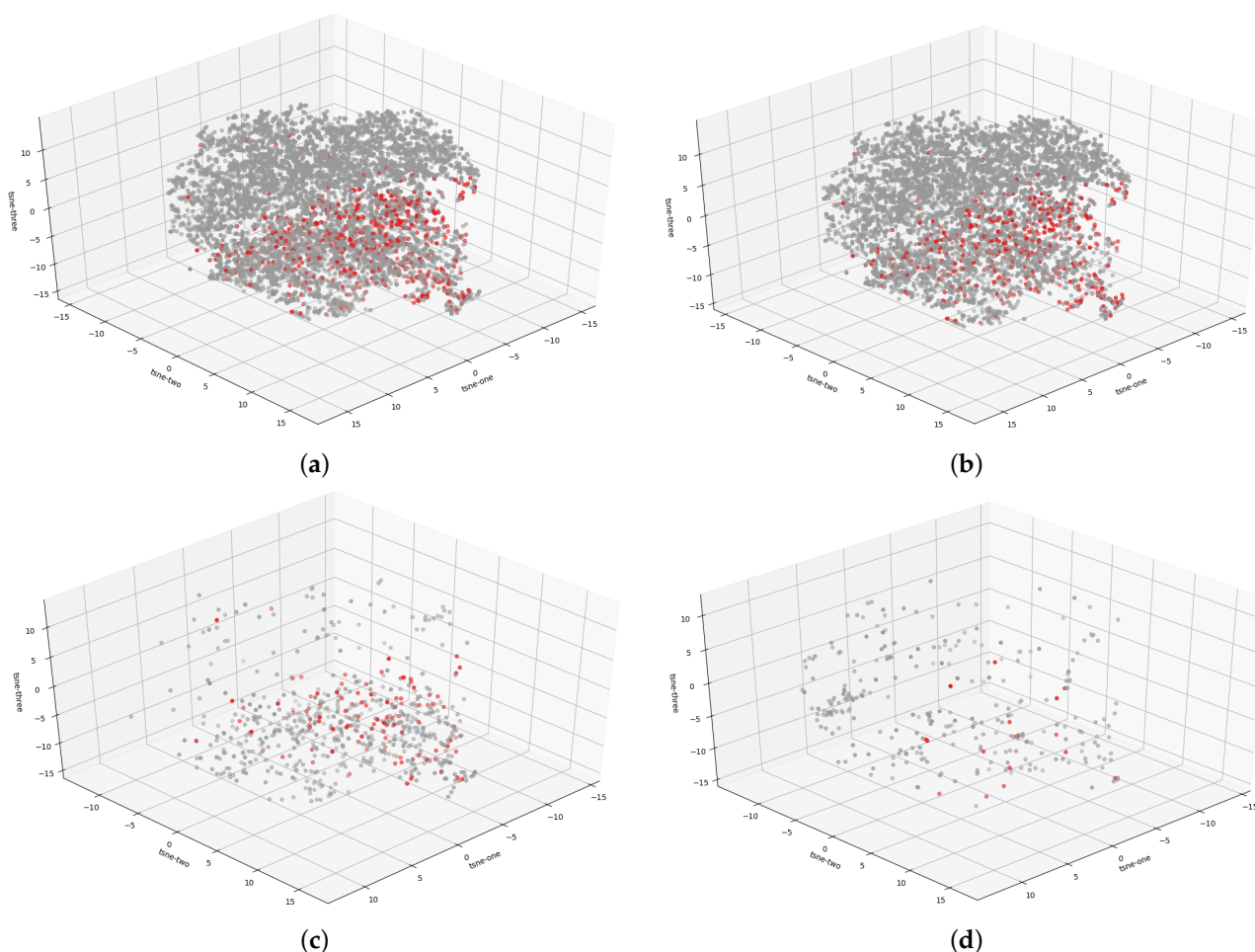
Figure 3. Interpretation of the 3D t-SNE plot visualization of the HAM10000 dataset where the red dots indicate melanoma cases while gray dots represent benign lesions. The following figures present the distribution of malignant and benign cases for the (**a**) entire dataset, (**b**) trunk-extremities dataset, (**c**) face-head dataset, and (**d**) acral dataset.

3. Method

3.1. Determination of the Anatomic Site of the Skin Lesion

An overview of our method is illustrated in Figure 4. We reuse deep CNN models pretrained on the ImageNet dataset for feature extraction using the prepared HAM10000 dataset for the classification of skin lesions into anatomic sites of the body. We adjust the classifier, which has a three layer structure. As a result, we generate classification outcomes for the most widely used pretrained networks and analyze them. We further employ the Grad-CAM algorithm to generate heatmaps in order to conduct multi-task learning model interpretability.

Figure 4. The streamline of our proposed framework. We use pretrained deep learning models including VGG19, ResNet50, DenseNet121, and EfficientNetB0 for feature extraction on the HAM10000 dataset. We employ the extracted features to conduct the multi-class classification task. Finally, we perform model evaluation and interpretation based on the heatmaps.

3.2. Separability Analysis Using Deep Learning

We analyze the capability of the existing deep learning frameworks in discriminating three anatomic sites (trunk and extremities, acral, and face/head). This analysis will inform the design of our proposed method. For this preliminary analysis, we trained the models for 25 epochs without pretrained models and without data augmentation. Figure 5 presents the visualization of the data distribution by each network. The statistical metrics presented in Table 3 confirm that the three anatomic sites are separable and create clusters, where the intra-class values are lower and inter-class values are much higher. We observed that all of the implemented pretrained networks achieved high values for the CH index, which indicates huge potential in obtaining good results for the classification task. Considering the small size and imbalanced nature of the dataset, we propose several strategies to overcome these challenges in the following section.

Figure 5. Visualization of full dataset feature distributions extracted with VGG19, ResNet50, Xception, DenseNet121, and EfficientNetB0. These graphs visually illustrate the separability of three anatomic sites.

Table 3. Statistical analysis on the separability of three anatomic sites (trunk, acral, and face/head) of the HAM10000 dataset using UMAP visualization on deep learning methods.

Method	Metrics						
	Intra (Trunk)	Intra (Acral)	Intra (Head)	Inter-Class	Silhoutte	CH	DB
Input	4.2582	4.2815	3.3802	4.5036	−0.0016	251.8074	3.8125
VGG19	7.0763	3.4522	3.7104	7.5375	0.0658	822.4723	1.5865
ResNet50	3.9671	1.4387	1.7728	9.9075	0.4192	3463.3257	0.5819
Xception	3.0442	3.6002	3.0653	17.6211	0.8052	15,137.3670	0.3198
DenseNet121	4.2608	1.9341	1.7729	6.5770	0.3685	1919.5011	0.7422
EffcientNetB0	4.4417	2.2625	2.0617	9.1643	0.5488	3458.3440	0.6102

3.3. Pretrained Model Based Architecture

Due to our limited and imbalanced dataset we take advantage of the transfer learning concept which indicates the effectiveness of reusing pretrained CNN architectures to extract the feature representation. There are several strategies of performing transfer learning including fine-tuning and feature extraction. However, due to our problem specification we propose a CNN based architecture which consists of a pretrained convolutional base and an adjusted classifier. We tested several state-of-the-art architectures including VGG19 [20], ResNet50 [21], DenseNet121 [22] and the latest EfficientNetB0 [23]. EfficientNet models which have been introduced in 2019 by Tan et al. are based on the inverted bottleneck residual blocks of MobileNetV2 and squeeze-and-excitation blocks. They use a compounding scaling method which scales width, depth, and resolution together instead of scaling only one model attribute. The EfficientNetB0 architecture has been proposed by a multi-objective neural architecture search which optimizes both accuracy and floating-point operations. Furthermore, a new activation function, Swish, has been proposed which shows superior performance for deeper networks. Swish is a multiplication of a linear and a sigmoid activation [23]:

$$Swish(x) = x * sigmoid(x) \tag{4}$$

On top of the base, we have adjusted a fully connected classifier that contains the following layers: dense layer with 256 neurons and ReLU activation function, additional dropout layer which randomly sets input units to 0 with frequency of rate 0.7 at each step during training time as a regularization technique for reducing overfitting [24]. The architecture closes with a dense layer with the number of neurons corresponding to the number of classes and Softmax activation function for the predict a multinomial probability distribution.

3.4. Deep Learning Architecture Training

For each of the pretrained architectures including VGG19, ResNet50, Xception, DenseNet121, and EfficientNetB0, we deployed randomized search (RandomizedSearchCV) for optimizing hyperparameters including number of epochs, optimizer, and batch size [25].

The algorithm selected 20 random sets of parameters from an established range, maintaining an equal distance in a search space. We tested batch size and number of epochs from ranges $batch_{size} = 8, 16, \ldots, 512$ and $nb_{epochs} = 5, 10, \ldots, 50$, respectively, and tested several optimizers including RMSprop, SGD, Adadelta, Adam, and Adamax. The learning rate was left at default, as it greatly varies between different optimizers. Hyperparameter optimization was performed using 3-fold cross-validation on a training set. By training our model repeatedly with different parameters from this grid, we were able to select a more narrow area of parameters. Then, we used Grid Search, which performs an exhaustive search on all different hyperparameter combinations, for a much smaller range of parameters. Finally we empirically tuned those numbers further by analysing the model's behaviour on a separate validation set and, for example, stopping the training earlier to avoid overfitting. After deciding the final set of parameters for each network architecture, we trained the models again, five times each, this time also checking the model's performance on a completely separate test set. Achieved results for each training were averaged. Final parameters and results are presented in Table 4.

In Figures 6 and 7, we show the average training and validation accuracy for DenseNet121 and EfficientNetB0 architectures, which are the top two performers, and achieved the highest score in classifying skin lesions into the three main anatomical sites.
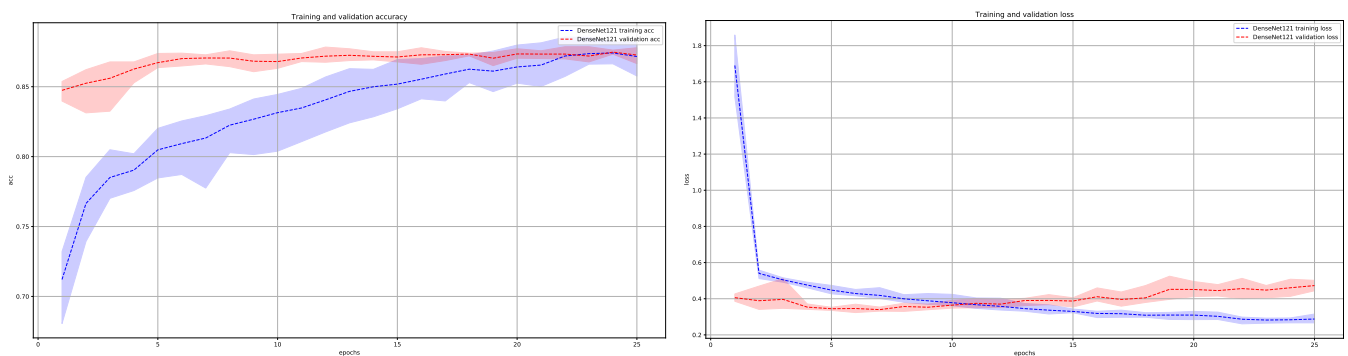
Figure 6. Average training and validation accuracy (**left**) and loss (**right**) during training of DenseNet121 for five times with maximal and minimal deviation areas marked in color (blue for training and red for validation).
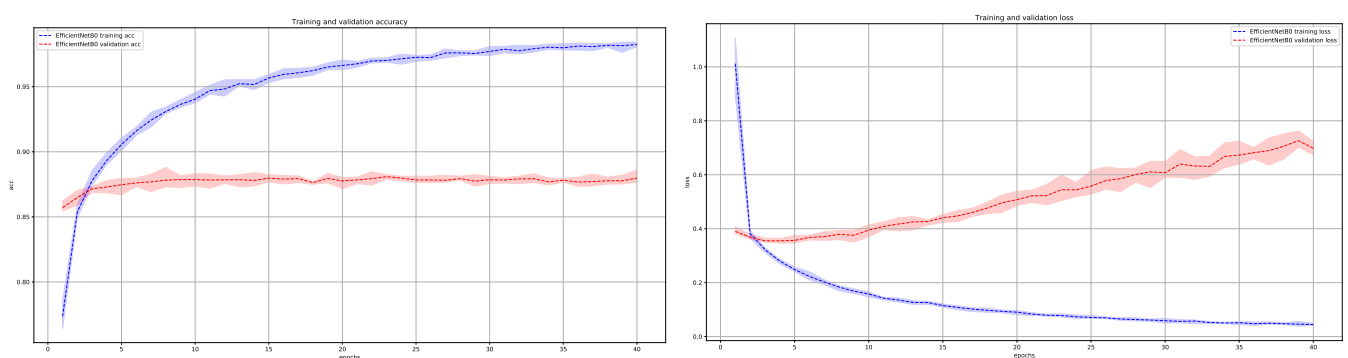
Figure 7. Average training and validation accuracy (**left**) and loss (**right**) during training of EfficientNetB0 for five times with maximal and minimal deviation areas marked in color (blue for training and red for validation).

4. Experimental Results

4.1. Statistical Metrics

We compare the ability of state-of-the-art algorithms in classifying dermoscopic images of skin moles into three main anatomic sites of the body, including trunk/extremities, face/head, and acral lesions on five state-of-the-art deep learning networks, i.e., VGG19, ResNet50, Xception, DenseNet121, and the latest EffcientNetB0. The evaluation of the

implemented and optimized architectures has been performed by using 20% of the dataset. The test results have been calculated five times and averaged.

The following performance metrics have been calculated based on the confusion matrix: accuracy (*ACC*), precision (*PPV*, positive predicted value), recall (*SE*, Sensitivity), and F1-score, where we specify the following: *TP* (true positive), *FN* (false negative), *TN* (true negative), and *FP* (false positive) values.

Accuracy, which measures statistical bias and systematic error, refers to the closeness of the measurements to a specific value and can be expressed as follows.

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

Precision refers to random errors, and it is a measure of statistical variability, which describes the closeness of the measurements to each other and can be written as follows.

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

Recall measures the proportion of actual positives that are correctly identified as such and is defined as follows.

$$SE = \frac{TP}{TP + FN} \tag{7}$$

F1-score (also F-score) considers both the precision and the recall of the test to compute the final score and is a measure of the test's accuracy. The F-score can be expressed as follows.

$$F_1 = \frac{2 \cdot PPV \cdot SE}{PPV + SE} \tag{8}$$

4.2. Effectiveness of the Proposed Framework

From the results presented in Table 4, we can conclude that all models were able to correctly recognize anatomic sites with high accuracy. Table 4 presents the evaluation metrics for each network architecture for the best set of training hyperparameters (optimised using grid search method described in Section 3.4). EfficientNetB0 achieved 91.45% accuracy and 91.5% F1-score, precision, and recall, which were the best results when trained with 45 epochs, batch size of 128, and the Adamax optimizer [26]. High precision and recall indicate the overall good performance of the model, with no visible biases. From the group of other architectures, only DenseNet121 managed to overcome the barrier of 90%, with others being slightly worse.

Table 4. Anatomic body site classification results for different neural network architectures with optimal set of parameters for each network and input images resized to 224 × 224.

| Architecture	Optimal Training Hyperparameters			Metrics			
	Optimizer	Batch Size	Epochs	Accuracy	Precision	Recall	F1
VGG19	SGD	64	25	0.883	0.89	0.89	0.89
ResNet50	SGD	32	50	0.898	0.90	0.90	0.90
Xception	Adadelta	128	35	0.867	0.87	0.87	0.86
DenseNet121	Adam	64	25	0.902	0.90	0.90	0.90
EfficientNetB0	**Adamax**	**128**	**45**	**0.9145**	**0.915**	**0.915**	**0.915**

In addition to mentioned statistical metrics, we also assessed the effectiveness of the proposed framework using various visualisation and interpretability techniques, including our own metric, which we further describe in the next section.

4.3. Model Interpretability Based on Heatmaps Analysis

In order to improve model explainability, we used the Grad-CAM visualization algorithm [27], which creates a heatmap that shows which parts of the input image contributed

the most to the classification. Furthermore, we performed an overlapping of the heatmaps with the segmentation ground-truth provided by Tschandl et al. [28].

In Figure 8, we present two examples for each anatomic site with their corresponding heatmaps for pretrained architectures. Regions on which the network focuses are marked in bright colors superimposed on the input image. From these images, we can draw several conclusions. Firstly, we observe that the proposed architectures do not always concentrate on the region of interest. For VGG19 and ResNet50, the classification is mostly based on the surrounding area resulting in a low Softmax score (p value) within the range of 0.4–0.7, while DenseNet121 and EfficientNetB0 calculated the final score based on the skin lesion area and achieved the highest p value close to one. Furthermore, EfficientNetB0, which achieves the best results, tends to take very large areas into consideration instead of focusing on a single area.
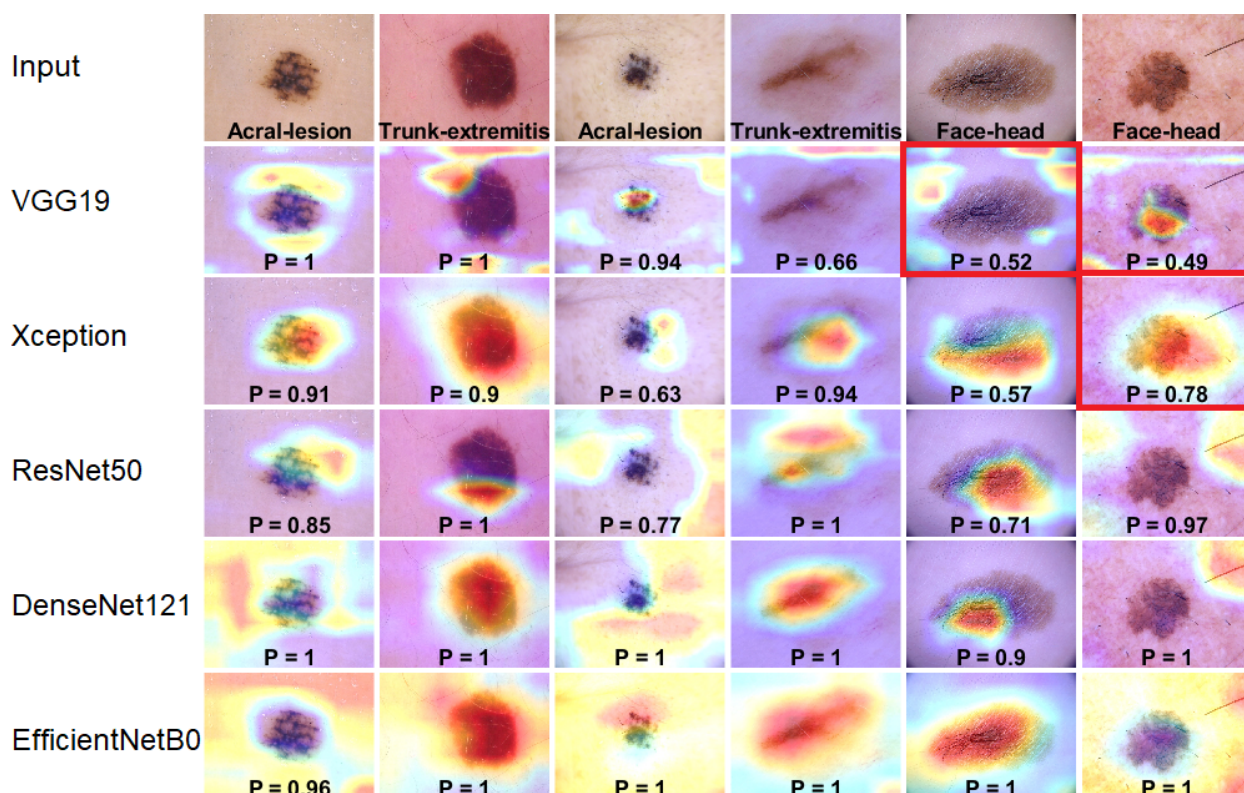

Figure 8. Grad-CAM visualization results. We compare the visualization results for each integrated pre-trained network based on the classification of skin lesion into the anatomic site of the body. The input image is shown on the top, and P denotes the Softmax score.

Acral cases were found to be mostly classified based on the background of the skin, which is connected to the papillary pattern occurring in palms and soles. Trunk and face skin lesion images are classified based on the area of the lesion. These results provide strong evidence of the importance of differentiating between skin lesions originating in different parts of the body.

Moreover, we have proposed and calculated an overlapping index that compares the areas between heatmaps and segmentation ground-truth images. It confirms to what extent the classification is based on the area of the skin lesion. The $Heatmap_{index}$ is defined as the sum of intensity pixels in the heatmap within the segmentation area divided by the sum of all pixels in the heatmap. The formula is given by the following:

$$Heatmap_{index} = \frac{\sum_{(x,y)\epsilon|H\cap S|} H(x,y)}{\sum_{(x,y)\epsilon H} H(x,y)} \cdot 100\% \tag{9}$$

where H is the heatmap image, and S is the binary segmentation mask. Based on the proposed overlap coefficient, we can observe (see Figure 9) and confirm that the classification has been mostly performed based on the skin lesion area for skin lesions originating in trunk/extremities and face, while the acral lesions have been classified based on the surroundings.
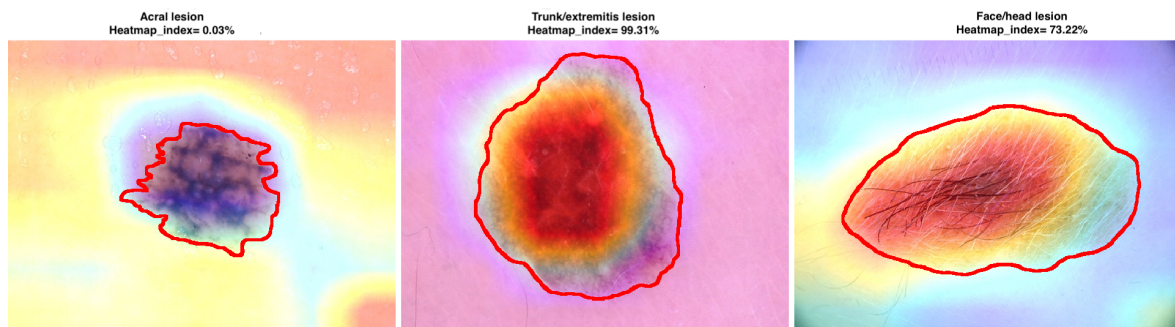
Figure 9. Examples of skin lesions originating in different anatomic sites with the corresponding heatmap created by the EfficientNetB0 model and segmentation ground-truth marked with red color. $Heatmap_{index}$ indicates to what extent the classification is based on the area of the skin lesion.

4.4. Software and Hardware

This research study was conducted using Python 3.7 programming language with Keras 2.3 [29] and scikit-learn [30] libraries. The models were trained on a NVIDIA RTX 2070 Super GPU (8 GB) with 48 GB RAM and Intel i7 Processor.

5. Conclusions

In this study, we developed a deep learning architecture based framework capable of skin lesion classification of the three main anatomical sites trained on the HAM10000 dataset. The network was shown to have high accuracy (>91%) in the classification of face, trunk and extremities, and acral anatomical regions. Furthermore, a heatmap analysis was used to determine locations on dermoscopic images in which the network based its classification on. The resulting architecture shows that features within dermoscopic images can be used to determine anatomical locations of skin lesions.

Author Contributions: Conceptualization, J.J.-K. and M.H.Y.; methodology, J.J.-K.; software, A.B.; validation, A.B., J.J.-K. and M.HY.; formal analysis, J.J.-K.; investigation, J.J.-K. and M.HY; resources, A.B., B.C., C.K., J.J.-K. and M.H.Y.; data curation, A.B., B.C., C.K., J.J.-K. and M.H.Y.; writing—original draft preparation, A.B., B.C., C.K., J.J.-K. and M.H.Y.; writing—review and editing, A.B., B.C., C.K., J.J.-K. and M.HY.; visualization, A.B., B.C., C.K., J.J.-K. and M.H.Y.; supervision, J.J.-K. and M.H.Y; project administration, J.J.-K. and M.H.Y; funding acquisition, J.J.-K. and M.H.Y. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in this article.

Conflicts of Interest: The authors declare no conflict of interest..

References

1. Society, A.C. American Cancer Society. Cancer Statistics Center. 2021. Available online: https://cancerstatisticscenter.cancer.org/#!/ (accessed on 14 October 2021).
2. Argenziano, G.; Soyer, P.; De Giorgi, V.; Piccolo, D.; Carli, P.; Delfino, M.; Ferrari, A.; Hofmann-Wellenhof, R.; Massi, D.; Mazzocchetti, G.; et al. *Interactive Atlas of Dermoscopy*; Edra Medical Publishing & New Media: Milan, Italy, 2000.
3. Argenziano, G.; Fabbrocini, G.; Carli, P.; de Giorgi, V.; Sammarco, E.; Delfino, M. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Arch. Dermatol.* **1998**, *134*, 1563–1570. [CrossRef]
4. Argenziano, G.; Soyer, H.; Chimenti, S.; Talamini, R.; Corona, R.; Sera, F.; Binder, M.; Cerroni, L.; de Rosa, G.; Ferrara, G.; et al. Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. *J. Am. Acad. Dermatol.* **2003**, *48*, 679–693. [CrossRef]
5. Nachbar, F.; Stolz, W.; Merkle, T.; Cognetta, A.; Vogt, T.; Landthaler, M.; Bílek, P.; Braun-falco, O.; Plewig, G. The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions. *J. Am. Acad. Dermatol.* **1994**, *30*, 551–559. [CrossRef]
6. Yu, C.; Yang, S.; Kim, W.; Jung, J.; Chung, K.Y.; Lee, S.W.; Oh, B. Acral melanoma detection using a convolutional neural network for dermoscopy images. *PLoS ONE* **2018**, *13*, e0193321.
7. Le, D.N.; Le, H.X.; Ngo, L.; Ngo, H.T. Transfer learning with class-weighted and focal loss function for automatic skin cancer classification. *arXiv* **2020**, arXiv:2009.05977.
8. Winkler, J.K.; Sies, K.; Fink, C.; Toberer, F.; Enk, A.; Deinlein, T.; Hofmann-Wellenhof, R.; Thomas, L.; Lallas, A.; Blum, A.; et al. Melanoma recognition by a deep learning convolutional neural network—Performance in different melanoma subtypes and localisations. *Eur. J. Cancer* **2020**, *127*, 21–29. [CrossRef]
9. GmbH, F.S. FotoFinder Moleanalyzer Pro & AI; 2013 FotoFinder Moleanalyzer Pro. Available online: https://www.fotofinder.de/en/technology/artificial-intelligence (accessed on 14 October 2021).
10. Han, S.S.; Moon, I.J.; Lim, W.; Suh, I.S.; Lee, S.Y.; Na, J.I.; Kim, S.H.; Chang, S.E. Keratinocytic Skin Cancer Detection on the Face Using Region-Based Convolutional Neural Network. *JAMA Dermatol.* **2020**, *15*, 29–37. [CrossRef] [PubMed]
11. González-Cruz, C.; Jofre, M.; Podlipnik, S.; Combalia, M.; Gareau, D.; Gamboa, M.; Vallone, M.; Faride Barragán-Estudillo, Z.; Tamez-Peña, A.; Montoya, J.; et al. Machine Learning in Melanoma Diagnosis. Limitations about to be Overcome. *Actas Dermo-Sifiliográficas (Engl. Ed.)* **2020**, *111*, 313–316. [CrossRef]
12. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [CrossRef]
13. Gutman, D.A.; Codella, N.C.F.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Mishra, N.K.; Halpern, A.C. 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 Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 168–172.
14. Combalia, M.; Codella, N.C.F.; Rotemberg, V.; Helba, B.; Vilaplana, V.; Reiter, O.; Halpern, A.C.; Puig, S.; Malvehy, J. BCN20000: Dermoscopic Lesions in the Wild. *arXiv* **2019**, arXiv:1908.02288.
15. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861. [CrossRef]
16. Maaten, L.V.D.; Hinton, G.E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
17. Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
18. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.—Theory Methods* **1974**, *3*, 1–27. [CrossRef]
19. Davies, D.L.; Bouldin, D. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
20. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
23. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
24. Srivastava, N.; Hinton, G.E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
25. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
26. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2015**, arXiv:1412.6980.
27. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
28. Tschandl, P.; Rinner, C.; Apalla, Z.; Argenziano, G.; Codella, N.C.F.; Halpern, A.; Janda, M.; Lallas, A.; Longo, C.; Malvehy, J.; et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **2020**, *26*, 1229–1234. [CrossRef] [PubMed]

29. Chollet, F. Keras. 2015. Available online: https://keras.io/ (accessed on 14 October 2021).
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.