IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 12 July 2024, accepted 30 July 2024, date of publication 6 August 2024, date of current version 16 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3439365

RESEARCH ARTICLE

A Large Dataset to Enhance Skin Cancer Classification With Transformer-Based Deep Neural Networks

MIRCO GALLAZZI, SARA BIAVASCHI, ALESSANDRO BULGHERONI, TOMMASO M. GATTI, SILVIA CORCHS, AND IGNAZIO GALLO

Department of Theoretical and Applied Science, University of Insubria, 21100 Varese, Italy

Corresponding author: Mirco Gallazzi (mgallazzi2@uninsubria.it)

ABSTRACT The advent of Deep Learning methodologies has revolutionized the field of medical image analysis, particularly in skin lesion diagnosis and classification. This paper proposes an explorative approach utilizing Transformer-based deep neural networks to classify multiclass skin lesion datasets. Initially introduced for natural language processing tasks, Transformers have remarkably succeeded in capturing long-range dependencies in sequential data. However, their application to image data, especially in medical imaging, remains relatively unexplored. Our proposed framework leverages the self-attention mechanism of Transformer models to effectively capture spatial dependencies across image regions without relying on handcrafted features or extensive pre-processing. We present a comprehensive evaluation of several Deep Learning models on skin imaging reference datasets for various types of skin lesions, including melanoma. We objectively evaluate the test performance of the different trained models using a test dataset released in 2023 with ground-truth labels. Our experiments demonstrate that the Transformer-based architecture achieves high performance in lesion classification tasks. The best result was obtained using a Large Dataset, which we modeled by merging smaller datasets, achieving a test accuracy of 86.37%. This dataset can be considered a good solution to improve the generalization capabilities of the Transformer neural network. Our work establishes Transformer-based deep neural networks as a promising framework for skin-lesion classification in medical imaging and potential clinical utility. This research paves the way for further exploration and integration of advanced Deep Learning techniques into medical image analysis, ultimately contributing to a powerful initial analysis tool for clinicians. The code is publicly available at https://github.com/UnluckyMirco/A-Large-Dataset-to-Enhance-Skin-Cancer-Classification-with-Transformer-Based-DNN.

INDEX TERMS Skin cancer, skin lesion, transformer neural network, image classification.

I. INTRODUCTION

Malignant Melanoma (MM), Squamous Cell Carcinoma (SCC), and Basal Cell Carcinoma (BCC) are the most prevalent types of skin cancers, collectively accounting for approximately 95% of all cases. According to estimates from the World Health Organization, there are between two to three million new cases of skin cancer annually. Despite advancements in treatment, the 5-year survival rate for skin cancer patients starts from almost 100% survivability in the early stages and drops to approximately 25-35% in the latest stages of all skin cancer-related deaths [1], [2]. When BCC, SCC, or other skin cancers metastasize, the prognosis is generally bleak, significantly threatening patients' lives and often severely impacting their appearance [3], [4], [5].

Machine Learning (ML) technologies have emerged as a cornerstone for developing innovative and efficient solutions to support clinicians. The advent of ML in dermatology has revolutionized diagnostics, enhancing accuracy, speed, and scalability. The International Skin Imaging

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

Collaboration (ISIC) [6] provides a repository of diverse datasets, often accompanied by challenges [7] to advance nevus image classification techniques through ML and Deep Learning (DL) solutions.

Despite considerable progress, developing models that generalize across large and heterogeneous datasets remains a challenge. Many datasets in the ISIC repository contain duplicate images, both within and between datasets, introducing undesirable biases in ML models [8]. In addition, some datasets, such as ISIC2020, offer many images with binary labels (malignant or benign). Although such datasets provide a substantial number of images, their limited labeling constrains the classification potential of DL models. More granular labeling would allow the development of more specific tools to support clinicians.

The integrity of test datasets is crucial for unbiased model performance evaluation. Issues such as the internal partitioning of datasets into training and testing sets can introduce variability, complicating objective model assessment. A standardized test dataset with ground-truth labels would facilitate fairer performance comparisons across different models.

Among these challenges, DL and Transformer Models (TMs) have shown exceptional promise due to their ability to discern intricate patterns in complex datasets, surpassing the capabilities of traditional analytical methods. However, these models still face deficiencies in objective assessments of test performance.

This paper delves into the investigation and application of TMs within dermatology, focusing particularly on the classification of skin lesions—a critical aspect of early skin cancer detection and diagnosis. Our work introduces a comprehensive approach that minimizes data augmentation techniques and leverages specific neural network architectures, starting with pre-trained networks to adapt different training and testing experiments for skin disease classification.

The main contributions of this paper are as follows:

1. Investigate the Swin Transformer (ST) model [9] for skin lesion multiclass classification.
2. Use a standard test set available in the literature for a fair comparison of performances.
3. Explore how increasing training data impacts model performance.
4. Share our benchmarks and modeled dataset on our GitHub repository (check Section VII for further information).

Our paper is structured as follows: Section II provides an overview of existing works and their limitations. Section III delves into creating the Large Dataset, describing the datasets used and the modeling operations performed. Section IV outlines our work pipeline, detailing the data augmentation techniques, chosen models, and evaluation methods. Section V presents the experiments conducted and the results obtained. Section VI discusses these results, and Section VII concludes with final remarks.

II. RELATED WORK

Skin lesion classification using DL models has been widely applied in the last years. In particular, Convolutional Neural Networks (CNN) have shown notable outcomes. A deep CNN (DCNN) model for binary classification (benign vs malignant skin cancer) was presented in [11], obtaining 91.93% testing accuracy on the HAM10000 (HAM) [12] dataset. The main datasets available in the literature are presented and briefly reviewed in Section III.

Pretrained networks, specifically AlexNet [13] and VGG16 [14], were used in two separate contexts by Gulati and Bhogal [15]: as feature extractor and as Transfer Learning (TL) paradigm. VGG16 produced the best result as TL model, obtaining 97.5% accuracy and 96.87% specificity for multiclass classification tasks. A CNN model was also proposed by Naqvi et al. [16].

The authors reported 78.81% specificity and an accuracy of 84.76% in multiclass classification. Additionally, they suggested a CNN network using the Keras Sequential API and contrasted the outcomes with already trained models, including VGG16, ResNet50 [17], and DenseNet121 [18].

A CNN model, combined with Soft-Attention, was proposed by Datta et al. [19]. They evaluate the effectiveness of the Soft-Attention mechanism in the VGG16, ResNet34 [17], ResNet50, Inception ResNet v2 [20], and DenseNet201 [18] architectures for the multiclass classification of skin lesions and achieved a precision of 93.7% on the HAM dataset and 91.6% on ISIC-2017 dataset [21]. NASNetMobile, a lightweight DL architecture for skin multiclass cancer, was proposed in 2021 by Yilmaz et al. [22], and on the ISIC-2017 dataset, it scored 91.7% in testing accuracy and 81.77% in test precision.

With the advent of TMs, a shift has been observed in the methodologies employed for skin lesion analysis. Recently, the exploitation performances of Transformer networks with multiclass medical imaging were investigated by Matsoukas et al. [23], demonstrating their ability to handle sequential data within images and providing an alternative to the spatial hierarchies of CNNs. Moreover, Pedro and Oliveira [24] investigated how self-attention mechanisms can be integrated into TMs to focus more precisely on relevant image features, thereby potentially increasing model interpretability and performance.

Gulzar and Khan [25] present a comparative study on U-Net [26] and attention-based methods for skin lesion image segmentation. Their research addresses the challenges of melanoma detection due to the visual similarities between lesions. While traditional CNNs struggle with long-range spatial relations, the Vision Transformer (VIT) [27] offers a solution but requires large datasets, which are limited in medical imaging. They demonstrate that the hybrid TransUNet model, achieving an accuracy of 92.11% and a dice coefficient of 89.84%, outperforms other benchmarking methods, highlighting its potential for improving skin lesion diagnosis.

TABLE 1. The top five finishers for task 3 of the 2018 ISIC challenge [10]. The top two ranked used an external dataset for training.

Team	Approach Name	External Dataset	Acc.
MetaOptima Technology Inc.	Top 10 Models Averaged	Yes	88.5%
DAISYLab	Large Ensemble with heavy multi-cropping and loss weighting	Yes	85.6%
Medical Image Analysis Group, Sun Yat-sen University	Ensemble Of SENET and PNANET with DataAugmentation when TEST	No	84.6%
Li	densenet	No	81.5%
Ask Sina	Xception, DenseNet121 plus three CONV layers	No	81.2%

Xin et al. [28] proposed SkinTrans, an improved Transformer network. A VIT network has been established to verify that SkinTrans is a helpful tool for categorizing skin lesions. Multi-scale patch embedding is carried out after the image has been serialized using multi-scale and overlapping sliding windows. Contrastive Learning is the final strategy used to produce the maximum difference in the encoding outcomes of separate data. Two datasets were considered: HAM, which achieved a multiclass classification accuracy of 94.3%, and their own dataset, which achieved 94.1% on a three-class classification task.

Hao et al. [29] proposed ConvNetXT, a model with high multiclass classification capabilities. The proposed model uses pre-trained ConvNeXt and ST networks to extract local and global features from pictures. Attentional Feature Fusion (AFF) submodules are then utilized to fuse the extracted features. Furthermore, an Efficient Channel Attention (ECA) module is included in the ConvNeXt network to improve the model's focus on the skin lesion locations during training.

By using both Transformer and CNNs, which are based on end-to-end mapping and do not require previous information, an ST model for multiclass skin lesion classification is suggested by Ayas [30]. Moreover, a weighted cross-entropy loss was used to solve the class imbalance issue. The ISIC 2019 dataset achieved a sensitivity, specificity, accuracy, and balanced accuracy value of 82.3%, 97.9%, 97.2%, and 82.3%, respectively.

He et al. [31] suggest using a Fully Transformer Network (FTN) to analyze skin lesions to learn long-range contextual information. They use the ISIC 2018 dataset to perform extensive skin lesion analysis tests to confirm the efficacy and efficiency of FTN. Because of its effective Spatial Pyramid Transformer (SPT) and hierarchical network topology, FTN routinely beats other cutting-edge CNNs in terms of computing efficiency and the number of tunable parameters, according to their experimental results.

The collection and analysis of large datasets have been critical to the progress of the field. Methodologies for creating and curating the dataset and the implications for ML applications in dermatology have been investigated by [32] and [33]. These studies emphasize the importance of data variety and volume in improving the generalization capabilities of ML models.

However, in all the work reviewed, we always noticed that the datasets considered were not highly populated. Wen et al. [34] have also discussed this issue, highlighting significant variability in dataset characteristics and metadata reporting. Additionally, the under-representation of darker skin types further limits the generalizability and applicability of these datasets to real-life clinical settings. Thus, we did not take full advantage of the ability to train a large number of parameters, which is characteristic of Transformer networks.

Furthermore, we noticed that no external dataset has been predisposed for an objective comparison methodology in all the datasets proposed and used in the different works. All of the papers analyzed have modeled the datasets in such a way as to achieve a division of the dataset by training and testing. We know this split approach is correct, although it allows the model to be biased during model evaluation on the test dataset.

Particularly interesting is the ISIC 2018 challenge [10], where the HAM dataset is proposed as training, and a separate test set is provided with the corresponding ground-truth labels available since 2023 [35]. In Table 1, we report the results of this challenge (more information in Section III). Although these performances do not exploit the TMs, they offer a solid basis for objectively comparing the models' performance on the HAM dataset.

III. PROPOSED DATASET

As already noticed, the HAM dataset is widely used in the literature. It consists of ten thousand images divided into seven classes with an associated metafile with truth labels. The seven classes identified in the dataset are:

- Melanoma (MEL): Melanoma is a malignant neoplasm originating from melanocytes, the cells responsible for skin pigmentation.
- Melanocytic Nevi (NV): Melanocytic nevi are benign lesions comprised of localized accumulations of melanocytes.
- Basal Cell Carcinoma (BCC): Basal cell carcinoma is the most common form of skin cancer, originating from the basal cells of the epidermal layer.
- Actinic Keratoses (AKIEC): Actinic keratoses are precancerous, scaly skin lesions induced by chronic exposure to ultraviolet rays.

- Benign Keratosis-like Lesions (BKL): Benign keratosis-like lesions encompass a variety of benign conditions, such as seborrheic keratosis, sebaceous hyperplasia, and clear cell acanthoma.
- Dermatofibroma (DF): Dermatofibroma is a benign cutaneous nodule commonly resulting from a reaction to minor injuries or insect bites.
- Vascular Lesions (VASC): Vascular lesions include a range of conditions characterized by abnormal proliferation or dilation of blood or lymphatic vessels.

Moreover, the ISIC 2018 challenge provides a second dataset, totally independent from HAM, with a metadata file used for testing. This dataset is composed of the same seven classes presented in HAM. Before 2023, it was possible to download the test dataset without ground truth labels, but since 2023, it has been possible to download it with metadata files and ground truth [35]. This dataset is composed of 1511 images divided into seven classes as follows: 171 in MEL, 908 in NV, 93 in BCC, 43 in AKIEC, 217 in BKL, 44 in DF, and 35 in VASC.

The clear separation of the two datasets (training and external test) and the release of the ground truth facilitates unbiased evaluation and validation of the models. Furthermore, the usage of this dataset lays the foundations for a solid and fair comparison of the models' performances.

Considering several datasets already available in the literature [34], and starting from the HAM dataset, we here propose a Large Dataset (LD). This dataset is obtained from a thorough selection and integration of public datasets. We focus on datasets that include dermatoscopic images and select those images that correspond to one of the seven classes mentioned above.

Table 4 details the selected datasets and their cardinalities, image size, and number of images present for each class. The last column displays the total number of images corresponding to the seven selected classes. It is crucial to mention that some of these datasets possess a greater cardinality due to additional skin disease classes that were not considered in this work.

Our LD proposal consists of 41.975 images, and Figure 2 shows some example images from the several single datasets considered. We observe a wide variety of images, like, for example, zooming factors or features related to image chromaticity.

Table 3 shows the original cardinality and the operations performed on each single dataset to obtain the values in the table's last column. Below are the motivations that prompted us to perform these operations.

- *HAM10000 (HAM) [12]:* No image removal was performed, as its seven classes were considered reference datasets.
- *Consecutive biopsies for melanoma [36]:* Images not belonging to one of the seven selected classes were removed; images whose precise diagnosis was missing were also removed.
- *MSK1-4 [37]:* Images not belonging to one of the seven classes listed above were removed; images whose precise diagnosis was missing were also removed.
- *SKINL2 [38]:* Removal of images not belonging to one of the seven classes listed above was carried out.
- *UDA-1 [37]:* The removal of images whose precise diagnosis was missing was done.
- *7-point criteria evaluation database [39]:* Since our analysis is aimed at using only dermoscopic images, macroscopic type images and those not belonging to one of the seven classes listed above were removed from this dataset.
- *ISIC 2020 Challenge Training Set [40]:* Removal of duplicate images contained in a list provided by the ISIC institute itself was performed; furthermore, since this dataset was created for a binary (malignant, benign) rather than multiclass analysis, a large number of images whose detailed diagnosis was not specified were removed. Finally, images not belonging to one of the seven classes listed above were removed.
- *ISIC Challenge 2018: Task 1-2 Test [21]:* Upon cross-checking with the other datasets belonging to the ISIC archive, duplicate images were found, i.e., whose IDs were present in both this dataset and the 4 MSK1-4 datasets.
- *ISIC Challenge 2018: Task 1-2 Validation [21]:* Upon cross-checking with the other datasets belonging to the ISIC archive, duplicate images were found, i.e., whose IDs were present in both this dataset and the 4 MSK1-3 datasets.
- *MSK5 [37]*, *Hospital Italiano de Buenos Aires [41]*, *BCN20000 (BCN) [42]*, *UDA-2 [37]*, *PH2 [43]:* No image removal was performed.

IV. PROPOSED APPROACH

We have seen that Transformer-based models offer great performance scalability depending on the cardinality of the training dataset and the number of model parameters. In this section, we propose an approach that investigates the potential of the ST model [9] on skin disease classification. We have performed several experiments, where we investigate the behavior of the model's performance when increasing the number of training images (see Table 2), and we evaluate its classification accuracy on the external test dataset (see Table 7). Figure 1 provides a comprehensive understanding of our methodology.

In the first group of experiments, we considered the HAM dataset as training data. In the second group of experiments, we use the LD proposal as training. In both cases, the external test dataset was used for a fair comparison of the seven-class classification accuracies.

For both groups of experiments, data augmentation techniques are applied. In subsection IV-A, we explain

TABLE 2. Distribution of datasets used in the various experiments broken down class by class.

ACRONYM REFERENCE	MEL	NV	BCC	AKIEC	BKL	DF	VASC	TOT
HAM_noDuplicates	614	5403	327	228	727	73	98	7470
HAM_Duplicates	1113	6705	514	327	1099	115	142	10 015
HAM_NV_Downsampled	1113	5403	514	327	1099	115	142	8713
BCN_noDuplicates	524	1281	983	358	353	40	37	3576
BCN_Duplicates	2857	4206	2809	1168	1138	124	111	12 413
HAM_BCN_Duplicates	3970	10 911	3323	1495	2237	239	253	22 428
HAM_Duplicates_BCN_noDuplicates	1637	7986	1497	685	1452	155	179	13 591
HAM_BCN_noDuplicates	1138	6684	1310	586	1080	113	135	11 046
LARGE_DATASET_Derm_Duplicates	7167	22 498	4854	2646	3937	400	473	41 975
LARGE_DATASET_Derm_NV_Downsampled	7167	13 306	4854	2646	3937	400	473	32 783
LARGE_DATASET_Derm_NV_30Balanced	7167	9314	4854	2646	3937	400	473	28 791
LARGE_DATASET_Derm_NV_20Balanced	7167	10 644	4854	2646	3937	400	473	30 121
LARGE_DATASET_Unified_Duplicates	8413	24 929	6936	5034	4811	637	1373	52 133
LARGE_DATASET_Unified_NV_Downsampled	8413	13 521	6936	5034	4811	637	1374	40 726

TABLE 3. Datasets used to create the LD proposal. For each of the single datasets, we report the original number of classes, the total cardinality (IC), the number of classes (lesions) removed (DL), number of Missing Lesions (ML), number of Macroscopic Images (MI), number of duplicate images (DI) and Final Cardinality (FC).

Dataset	N. Classes	IC	DL	ML	MI	DI	FC
HAM10000 [12]	7	10 015					10 015
Consecutive biopsies for melanoma (2020) [36]	21	1295	157	22			1116
MSK1 [37]	17	1678	50	248			1380
MSK2 [37]	22	4880	50	50			4374
MSK3 [37]	25	466	32				434
MSK4 [37]	24	2050	234	10			1806
MSK5 [37]	17	111					111
Hospital Italiano de Buenos Aires Dataset [41]	10	1616					1616
SKINL2 [38]	51	437	40				397
BCN20000 [42]	8	12 413					12 413
UDA-1 [37]	7	557	2				555
UDA-2 [37]	7	60					60
7-point criteria evaluation database (dermatoscopic only) [39]	20	2013		48	1002		963
ISIC 2020 Challenge Training Set [40]	8	33 126	46	26 706		425	5949
ISIC Challenge 2018: Task 1-2 Test [21]	3	1000				461	539
ISIC Challenge 2018: Task 1-2 Validation [21]	3	100				53	47
PH2 [43]	3	200					200

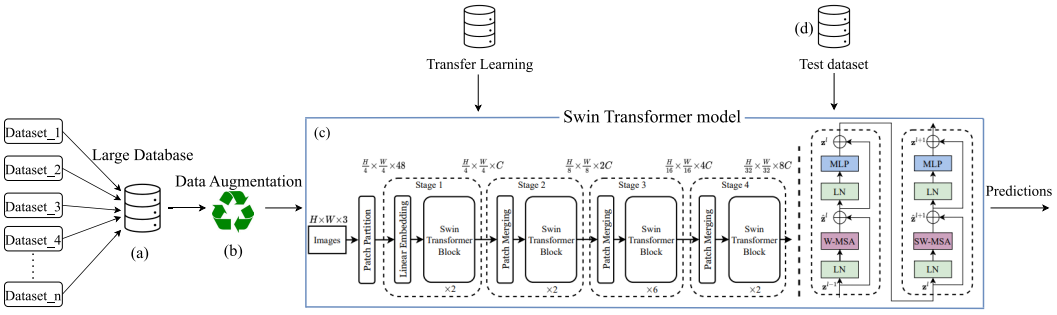


FIGURE 1. Overview of the proposed model and dataset for skin lesion classification with the key components: (a) creation of an LD assembled from multiple existing datasets; (b) application of various data augmentation techniques to improve the robustness of the model; (c) the architecture of the pre-trained ST model used [9]; (d) use of a standardized test set.

how the datasets are partitioned into training and validation sets and the relative augmentation techniques. Section III already discussed the methods used to choose datasets, their compositions, and the LD creation.

Most of the experiments were conducted using the pre-trained weights of the chosen TMs, while some experiments involved training from scratch and TL (see subsection IV-B and IV-C).

TABLE 4. Datasets used to create the LD proposal. For each single dataset, we report image size, number of images present for each of the seven classes here considered, and the total number of images.

DATASETS	Size of images	MEL	NV	BCC	AKIEC	BKL	DF	VASC	TOT
HAM10000 [12]	600 x 450	1113	6705	514	327	1099	115	142	10 015
Consecutive biopsies [36] for melanoma (2020)	3264 x 2448	117	691	12	60	223	6	7	1116
MSK-1 [37]	variable	368	760	71	17	116	9	39	1380
MSK-2 [37]	variable	937	1861	672	480	392	9	23	4374
MSK-3 [37]	variable	27	148	61	56	124	11	7	434
MSK-4 [37]	variable	247	595	278	303	343	26	14	1806
MSK-5 [37]	variable	0	0	0	0	109	0	2	111
Hospital Italiano de Buenos Aires Dataset [41]	variable	253	602	340	221	88	61	51	1616
SKINL2 [38]	1920 x 1080	53	151	52	14	64	17	46	397
BCN20000 [42]	1024 x 1024	2857	4206	2809	1168	1138	124	111	12 413
UDA-1 [37]	variable	159	396	0	0	0	0	0	555
UDA-2 [37]	variable	34	12	3	0	7	2	2	60
7-point criteria evaluation database (dermatoscopic only) [39]	768 x 512	252	575	42	0	45	20	29	963
ISIC 2020 Challenge Training Set [40]	variable	581	5191	0	0	177	0	0	5949
ISIC Challenge 2018: Task 1-2 Test [21]	variable	118	410	0	0	11	0	0	539
ISIC Challenge 2018: Task 1-2 Validation [21]	variable	11	35	0	0	1	0	0	47
PH2 [43]	765 x 575	40	160	0	0	0	0	0	200
TOT		7167	22 498	4854	2646	3937	400	473	41 975

FIGURE 2. Examples of skin disease images from different datasets for each of the seven classes here considered. In order, from top to bottom, we have selected (1) HAM10000, (2) MSK 1-5, (3) Consecutive biopsies for melanoma (2020), (4) Hospital Italiano de Buenos Aires Dataset, (5) SKINL2, (6) BCN20000, and (T) is the test dataset.

A. DATASET PREPROCESSING AND AUGMENTATION

An important aspect of this work concerns the pre-processing of data. Splitting the datasets in training and validation was the first key aspect. From Table 4, we notice that the datasets are highly imbalanced, and this led us to perform different types of experiments to find the optimal split choice. Regarding the training datasets, we have considered HAM, BCN, and LD datasets, and each time, 80% images were

selected, class by class and randomly, to create the training dataset and the remaining 20% to create the validation dataset. Data Augmentation (DA) allows us to increase the data by producing different images using the original ones as a base. DA techniques were investigated to mitigate the problem of unbalanced data.

The primary purpose is to increase the amount of data by leading to randomness and increasing the model's generalization. To implement this, we decided to implement dynamic DA by randomly selecting the techniques to provide the model with new information all the time.

However, it is essential to differentiate the techniques and the datasets to which they were applied. Specifically, applying the same techniques to the training and the validation or test datasets is not interesting because it would make the model inefficient in a general classification problem.

The techniques analyzed include:

- *Resize:* which allows an image to be resized according to the size required by the selected architecture;
- *Crop:* which allows an image to be cropped in a specific way (in our work, Centered and Random were investigated);
- *Horizontal Flip:* which allows images to be randomly flipped horizontally (i.e., along the vertical axis). It can also be defined as a "p" value, the value that determines the probability of application;
- *Rotation:* which rotates the image by a certain value of degrees. It can also allow the definition of a "p" value as the probability value of the application;
- *Normalization:* process that adjusts the pixel values of an image to have a mean and standard deviation that match specified values. The specific numbers utilized are commonly used for models pre-trained on the ImageNet dataset (mean 0.485, 0.456, 0.406; std 0.229, 0.224, 0.225).

As a result, different combinations of these techniques are proposed. The difference between the various proposals arises from the need to adapt the resize according to the input required by the model and the type of dataset on which they are applied. On the training dataset, several combinations are proposed where the main difference lies in the crop type applied. For the validation datasets, the only operations proposed are to resize the image to fit the required model input and the type of image crop. On the other hand, for the test dataset, the only difference concerns the crop type applied. This decision comes from the fact that we want to make the model evaluation as objective as possible without changing the image type of the test dataset.

B. TRANSFORMER MODELS

Classic deep learning methods, such as CNNs, excel at capturing local patterns through hierarchical feature extraction but struggle with long-range dependencies. In contrast, Transformers leverage self-attention mechanisms to capture global context and relationships between all input elements, allowing for more flexible and powerful modeling of complex data structures. One of the best models proposed in recent literature is the ST [9]. It enhances the representation power while maintaining efficiency by incorporating locally computed self-attention in non-overlapping windows and shifting these windows between successive Transformer layers. The main innovation of ST is its ability to adjust the processing scale dynamically, seamlessly moving from local to more global representations. The model focuses on fine-grained details within small windows in the initial layers. As information progresses through the stages, the feature maps are downsampled, and the model begins to attend to broader areas, integrating more context into its representations. This progression from local to global processing is crucial for capturing complex visual patterns and relationships in images [9]. The primary components of the ST architecture are as follows:

Hierarchical Structure: Similar to CNNs, ST hierarchically processes inputs. Let H and W stand for the input image's initial width and height. The number of channels C_s is doubled, but the feature map dimensions $H_s = H/2^s$ and $W_s = W/2^s$ are cut in half at each step s, which consists of several ST blocks.

Shifted Window-Based Self-Attention (ST Block): Swin calculates self-attention within each window, significantly reducing computational complexity. The attention mechanism is further refined by limiting the calculation of attention weights to a small set of neighboring pixels within a window, enhancing efficiency. Layer Normalisation (LN), a Multi-head Self-Attention (MSA) module with shifted windows, and a 2-layer MLP with GELU nonlinearity include each ST block.

Layer Normalization (LN): Applied before self-attention and MLP layers;

Multi-head Self-Attention (MSA): Utilizing shifted windows, defined by
$\mathrm{MSA}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O$ where each head head_i is computed as:
$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

The following is the expression of the key equation for window-based self-attention:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (1)$$

where

- Q, K, and V are the query, key, and value matrices, respectively;
- softmax represents the softmax function;
- $(\frac{QK^T}{\sqrt{d}})$ denotes the matrix multiplication of Q, K transposed, scaled by the square root of d, where d is the dimensionality of the query and key matrices;
- B is the bias term, which can be a matrix for relative positional biases in the context of Transformers;
- The entire expression inside the softmax function is then multiplied by V (value matrices).

Cyclic Shift and Window Partitioning: The window partitioning is shifted by half the window size in horizontal and vertical directions after every ST block, enabling cross-window connections and enhancing modeling power. To perform self-attention within local windows and allow for cross-window connection, ST use: Cyclic Shift$(x) = \sigma(x)$ where σ denotes the shifting operation applied to the input tensor x, facilitating cross-window communication.

Downsampling Layers: Between stages, a patch merging layer is used for downsampling, which can be described by: Patch Merging$(x) = \text{LN}(x)W + b$ Here, W and b are the learnable parameters for the linear transformation applied post-normalization.

Each layer in the ST contains parameters derived from the dimensions of Q, K, and V matrices and the number of heads in the multi-head attention mechanism. The total number of parameters and layers varies depending on the specific configuration of the Swin model (e.g., Swin-Tiny, Swin-Small, etc.). Table 5 shows, for each Swin network model, the number of layers per stage, the total number of layers, and the number of trainable parameters. The larger the network model becomes, the more trainable parameters there are. Swin's architecture is inherently flexible and scalable. Its performance scales favorably with model size and image resolution.

TABLE 5. Swin transformer parameters.

Model	Layers per Stage	#Layers	Param.
Swin-Tiny	2, 2, 6, 2	12	28M
Swin-Small	2, 2, 18, 2	24	50M
Swin-Base	2, 2, 18, 2	24	88M
Swin-Large	2, 2, 18, 2	24	197M
SwinV2-Large	2,2,18,18	48	197M

1) SWIN V2: ENHANCEMENTS AND DIFFERENCES

We also investigate whether further improvements in classification can be achieved by considering SwinV2 Transformer (SwinV2) [44]. It introduces some improvements over the original Swin:

Enhanced Window Attention: Incorporation of larger window sizes to capture more global context. This modification helps the model better understand long-range dependencies and contextual information.

Layer Normalization Adjustments: Placement of layer normalization inside the residual paths, rather than outside, which aligns more closely with the original Transformer architecture.

Increased Model Sizes: Larger variants with more parameters, aiming to further boost performance on various vision tasks.

While both Swin Large (SwL) and SwinV2-Large (SwV2L) models have similar parameter counts, SwV2L significantly increases the depth by incorporating 48 layers compared to SwL's 24 layers, as shown in Table 5.

C. MODEL EVALUATION AND FINE-TUNING

We evaluate the performance of the classification tasks in terms of accuracy, precision, recall, and F1-Score:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where "TP" indicates the "True Positives", "TN" the "True Negatives", "FP" "FN" indicates the "False Positives" and the "False Negatives", respectively.

Accurate model evaluation and fine-tuning are paramount for enhancing ML models' predictive performance and reliability. Model evaluation takes place in two different stages. The first occurs during model training. In each experiment, a large number of training epochs is set. At each epoch, the images in the training dataset are used to train the model, and the images in the validation dataset are used to evaluate the model's learning on images never seen during the training. At the end of the process, accuracy values for training and validation and loss values for training and validation are printed. Each time a training epoch ends, the loss function value is checked. If this value is lower than the previous epoch, the model is saved; otherwise, a new epoch starts immediately. This process is repeated until a lower loss value is found for 20 or 30 consecutive times (depending on the type of the experiment) than the best saved one. At the end of the training process, the confusion matrix and the accuracy, precision, recall, and f1-score metrics of the best epochs were selected through the process described before being printed.

The second, on the other hand, is performed at the end of model training. In this case, the utilized network is reloaded with the weights of the best epoch obtained during the training. At this point, the test dataset is used to evaluate the model's behavior on a completely independent dataset. In this way, through the printing of the confusion matrix and the accuracy, precision, recall, and f1-score metrics, we can verify the reliability of the model.

Whenever the fine-tuning approach is used, the explained procedure remains unchanged, the only difference being that it is repeated twice. In ML, fine-tuning refers to incrementally adjusting the parameters of an existing model, already trained on a general task, to improve its performance. Consequently, after applying the first training and test cycle on a specific dataset and evaluating its performance, a second dataset is chosen to refine the weights of the already trained model. The training and test methods are identical to the one described.

V. EXPERIMENTS

In this section we present the results of the different experiments performed. In the first group of experiments, the Swin model is trained on the HAM dataset; in the second

TABLE 6. Augmentation techniques used in the different experiments. For each strategy, the resize, crop size, horizontal flip, rotation, and normalization parameters are indicated. The parameter "p" represents the probability with which that technique is applied.

IDENTIFIER	Resize	Crop	Horizontal Flip	Rotation	Normalization
DA_Train	224*280	RandomCrop (224,224)	RandomHorizontalFlip (p=0.5)	RandomRotation (-180,180), p=0.99)	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225
DA_Train2	(600*0.65),(450*0.65)	RandomCrop (224,224)	RandomHorizontalFlip (p=0.5)	RandomRotation (-180,180), p=0.99)	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225
DA_Train3	(600*0.56),(450*0.56)	RandomCrop (224,224)	RandomHorizontalFlip (p=0.5)	RandomRotation (-180,180), p=0.99)	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225
DA_Train4	256*280	RandomCrop (256,256)	RandomHorizontalFlip (p=0.5)	RandomRotation (-180,180), p=0.99)	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225
DA_Val	224*280	RandomCrop (224,224)	-	-	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225
DA_Val2	256*280	RandomCrop (256,256)	-	-	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225
DA_Test	224*280	RandomCrop (224,224)	-	-	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225
DA_Test2	256*280	RandomCrop (256,256)	-	-	mean 0.485, 0.456, 0.406; std= 0.229, 0.224, 0.225

group, both HAM and BCN are considered, and in the third group, the LD is taken into account.

Table 2 shows the cardinality of each of the classes used for training during the different experiments. As pointed out by Cassidy et al. [8], some datasets contain duplicated images that can influence the performance of the classifiers.

To investigate the dependence of the model's performance on both duplicated images and imbalance data, we consider the following training strategies:

Removing all duplicate images (*noDuplicates*), downsampling the class NV since it has the most imbalance compared to the other classes (*Downsampled*), and keeping the duplicates (*Duplicates*). Images were considered duplicates if they shared the same ''lesion_id'' in the metadata file (each dataset also offers the possibility of downloading a metadata file in addition to the images). For each lesion, only the first image was kept, while the others, which were essentially the same image taken from different angles or with different zoom levels, were removed. Despite this, the inclusion of data augmentation ensures that there will still be sufficient variation in the images. The first column, ''REFERENCE ACRONYM,'' shows the name of the dataset used with the acronym of the chosen balancing technique just shown.

The other columns show, class by class, the corresponding number of images. The last column shows the total number of images considered in each experiment.

Table 6 summarizes the various types of augmentation applied. In the first column, we report the identifier associated with each of these experiments. The details of the data augmentation strategies are reported.

We kept standard parameters such as normalization, rotation, flip, and crop type. The resize adapts to the type of input the network requests except for DA_Train2 and DA_Train3, where we wanted to investigate different resize possibilities.

In table 7, we collected all the experiments in three groups. The ID of the experiments will be used to recall the experiments in their discussion, as well as the model selected, the dataset type, the data augmentation type, and the metrics to evaluate the test performances. Other experiments were conducted, but we decided to reproduce only the most significant ones.

All computations were performed on a containerized environment utilizing 1 Nvidia A100 80GB GPU, 16 cores of an AMD Epyc 7742 64-core CPU, and 64GB of DDR4-3200 RAM, all connected to a 76TB RAID6 storage server via a 25Gbps low-latency network. The environment used in our experiments uses PyTorch version 2.1.0 with CUDA version 11.8 and Python version 3.11.6 (more information can be found on the GitHub page, see VII). A batch size of 128 or 64 (depending on the type of experiment) was used for both training and validation datasets to ensure a balance between computational efficiency and model convergence. All experiments were conducted with a batch size of 128 except for MEL7, where a batch size of 64 returned the best result. The Adam optimizer with a learning rate of 1×10^{-4} was utilized to adaptively adjust the learning rate during training, which is effective in improving model performance. CrossEntropyLoss was employed as the loss function to handle the multi-class classification task. To prevent overfitting, early stopping was incorporated, terminating training after 30 epochs without improvement in validation loss. The model's performance was monitored and the best model parameters were saved upon observing the lowest validation loss. This approach ensures robust training and effective generalization of the model.

A. HAM EXPERIMENTS

This group of experiments includes only HAM as a training dataset. As discussed, it is one of the most widely used datasets. Consequently, starting a comprehensive bench of this dataset on a standard test dataset was fundamental. The experiments were conducted by dividing the selected datasets into training and validation 80-20 class by class. All the experiments performed involve the use of pre-trained models. Consequently, the weights of the original selected trained model will be imported and modified with a new training cycle. The best model is then saved as explained in IV-C and will be used to test the performance of the external test dataset. The augmentations used for the experiments changed

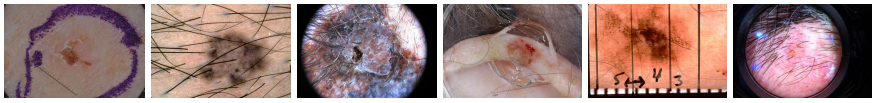


FIGURE 3. Examples of dermatoscopic images containing various types of noise that can affect the performance of classification algorithms.

due to the training phase, in which different combinations were utilized to verify the behavior of the selected networks. In the validation and testing phases, the techniques used remain unchanged, except for some cases where the network required a different input size.

In **MEL1**, we use HAM with duplicates. We specify that duplicate images represent the same lesion with different zoom or rotation. We selected the model SwL and used the DA_Train as augmentation. We obtained 83.06% on the test dataset. This represents our starting point. Since we noticed a great disparity in the distribution of classes, especially between the class NV and all others, and supported by Tschandl et al. [45], who demonstrated that class imbalance has a significant impact on model performance, we decided to conduct an initial experiment by eliminating all duplicate images in the classes. This led us to have a total number of images of 7470. **MEL2** and **MEL3** are experiments involving the HAM dataset without duplicates. The same DA strategy is applied while we use the SwL network for **MEL2** and SwinV2 Base network for **MEL3**. The accuracies obtained are respectively 83.85% and 82.13%.

In **MEL4, MEL5**, we investigate the influence of different DA techniques, especially the resize factor. In **MEL4**, we use a resize factor of 65% (DA_Train2); instead, in **MEL5**, we apply a resize of 56% (DA_Train3). These experiments obtained respectively 84.51% and 83.52% performance accuracy. Let us recall that we are dealing with dermatoscopic images, and the zoom quality can be considered quite high. However, the images can still contain many artifacts that may introduce noise into our classification task or obscure the lesions themselves. Such noise includes elements such as skin hair, a substantial amount of surrounding skin, marks around the lesions, and other characteristic artifacts commonly associated with dermatoscopic imaging. Figure 3 illustrates examples of these types of noise, highlighting their potential impact on the image quality and subsequent classification.

The large imbalance between the classes NV and all the others remains visible in these first experiments (a difference of 4,676 images between the most populated class, NV, and the second most populated class, BKL). From the confusion matrix of **MEL4**, reported in Figure 4 at the top left, we observe that MEL and BKL are strongly influenced by the high number of images in NV.

In the **MEL6** experiment, We keep all the classes with duplicates except the NV one to reduce the disparity of images between the various classes. We utilized the same network used in previous experiments and the same type of DA. It achieves a test accuracy of 84.32%, which is the second highest in this first group of experiments.

We also investigate the performance accuracy of the SwV2L network. We kept the same conditions of MEL6 experiment, except for the image resize, as this network model accepts an input size of 256*256 (DA_Train4). This change was also applied to the validation and test sets (DA_Val2 and DA_Test2, respectively). The result was the highest test accuracy value for this experiment group, equal to 84.64% in **MEL7**. The corresponding confusion matrix is found in Figure 4 at the top-mid and confirms the same behavior seen for MEL4.

Other tests were also conducted, mixing the various combinations of models and DA strategies. Still, none of them produced better or more significant results than the ones proposed here. Consequently, the highest classification value in tests we obtained is produced in the experiment **MEL7**.

B. HAM AND BCN EXPERIMENTS

Given the results of previous experiments and being aware of the potential of TMs, we investigated the Swin model's behavior on a more populated dataset. We use fine-tuning between HAM and BCN datasets. HAM was chosen for the same reasons that led us to use it previously, and BCN was selected because it was the most similar dataset to HAM in terms of the number of classes (BCN has 8 classes, but we eliminated the different one) and populousness among the proposed datasets.

In the first experiment, **MEL8**, we investigate the union of the two datasets into a single one. This dataset consisted of 10015 HAM images and 12413 BCN images. All duplicates were eliminated. We use DA_Train3 as an augmentation to verify its behavior in a different context. Using the external test set, the accuracy obtained was equal to 85.70%. In Figure 4, at the top-right, we report the corresponding confusion matrix.

In experiments from **MEL9** to **MEL15**, the models are trained from scratch on HAM or BCN and then fine-tuned on the other dataset.

In **MEL9** (**MEL10**), HAM (BCN) is used to train the network from scratch and BCN (HAM) for fine-tuning, in both cases, without duplicated images. A test accuracy of 79.81% (83.45%) is obtained.

In **MEL11**, the SwL model is trained from scratch using HAM with duplicated images and then fine-tuned with BCN without duplicates. This experiment achieved a test accuracy of 80.68%.

The last experiment of the second group is **MEL12**. The model is trained from scratch on HAM without duplicates and fine-tuned on BCN without duplicates. The accuracy achieved is 83.65%.

TABLE 7. This table collects all the experiments in this paper. The lines divide the three groups of experiments in detail: MEL1-7 are the experiments where HAM was used, MEL8-12 the experiments where HAM and BCN were used for fine-tuning, and MEL13-20 are the experiments involving the use of the proposed LD. Values in bold are the best values obtained for each type of experiment. The acronyms TA, TP, TR, and TF1 represent Test Accuracy, Test Precision, Test Recall, and Test F1 Score in percentage and they are calculated as a weighted average, namely taking into account the attendance of each class.

ID	MODEL	DATASET	DA	TA	TP	TR	TF1
MEL1	Swin Large	HAM_Duplicates	DA_Train	83.06	83.32	83.06	83.00
MEL2	Swin Large	HAM_noDuplicates	DA_Train	83.85	83.47	83.85	83.53
MEL3	SwinV2 Base	HAM_noDuplicates	DA_Train	82.13	82.18	82.13	81.65
MEL4	Swin Large	HAM_noDuplicates	DA_Train2	84.51	84.35	84.51	84.00
MEL5	Swin Large	HAM_noDuplicates	DA_Train3	83.52	83.44	83.52	83.36
MEL6	Swin Large	HAM_NV_Downsampled	DA_Train	84.32	84.71	84.32	84.37
MEL7	SwinV2 Large	HAM_NV_Downsampled	DA_Train4	**84.64**	84.77	84.64	84.57
MEL8	Swin Large	HAM_BCN_noDuplicates	DA_Train3	**85.70**	85.50	85.70	85.27
MEL9	Swin Large	HAM_BCN_noDuplicates	DA_Train	79.81	80.79	79.81	78.81
MEL10	Swin Large	HAM_BCN_Duplicates	DA_Train	83.45	83.54	83.45	82.76
MEL11	Swin Large	HAM_Duplicates_BCN_noDuplicates	DA_Train	80.68	80.34	80.73	80.19
MEL12	Swin Large	HAM_BCN_noDuplicates	DA_Train	83.65	83.62	83.65	83.53
MEL13	Swin Large	LARGE_DATASET_Derm_Duplicates	DA_Train	85.84	85.70	85.84	85.66
MEL14	Swin Large	LARGE_DATASET_Derm_NV_Downsampled	DA_Train	**86.37**	86.84	86.37	86.44
MEL15	Swin Large	LARGE_DATASET_Derm_NV_30Balanced	DA_Train	84.71	86.10	84.71	85.02
MEL16	Swin Large	LARGE_DATASET_Derm_NV_20Balanced	DA_Train	83.98	84.85	83.98	84.19
MEL17	SwinV2 Large	LARGE_DATASET_Derm_NV_Downsampled	DA_Train4	83.65	83.55	83.65	83.48
MEL18	Swin Large	LARGE_DATASET_Unified_Duplicates	DA_Train	74.39	74.20	74.39	73.88
MEL19	Swin Large	LARGE_DATASET_Unified_NV_Downsampled	DA_Train	84.58	86.07	84.58	84.98
MEL20	Swin Large	LARGE_DATASET_Unified_NV_Downsampled	DA_Train	70.68	73.73	70.68	71.51

MEL8 shows the best accuracy of this second group of experiments.

The third group of experiments considers the LD here proposed as training data.

C. LARGE DATASET EXPERIMENTS

In this section, we analyze how the performance of the models depends on the cardinality of the training data.

The DA utilized in this group is always the same, except for one experiment where we used a different version of Swin. The performances are always evaluated on the external test dataset.

In **MEL13**, we investigate the use of SwL on the LD with duplicates that contain a total number of 41975 images.

As data augmentation, DA_Train is used. The accuracy obtained for the test dataset is 85.84%. In Figure 4, at the bottom-mid, we report the corresponding confusion matrix. A SwL model has 197M of trainable parameters. We observe that by increasing the number of training images, the model learns to better generalize.

However, we have seen in previous experiments that by removing only the duplicated images of the class NV (reducing the disparity between the most populated class and the remaining classes), the model was more accurate in classification. Therefore, we decided to conduct a series of experiments where we varied the number of images within the class NV. If, in MEL16, the disparity between the two most populated classes was 15311 images, in the experiments **MEL14, MEL15** and **MEL16**, we reduced this disparity in three different ways.

In **MEL14**, we removed all the duplicates from the class NV. The model and DA remain the same as in the previous experiment. The accuracy achieved in the test was 86.37%, the highest value so far. In this case, we have reduced the number of images in NV from 22498 to 13306, leading to a disparity between the two most populated classes, which is 6139. As the dataset remains very unbalanced, the classification performance turns out to be the highest. In **MEL15** and **MEL16**, instead of removing all duplicates from NV, we decided to remove 30% and 20% of the images, respectively, randomly from NV. **MEL15** achieved an 84.71% of test accuracy, and **MEL16** achieved a value of 83.98%. Figure 4 has at the bottom-left the confusion matrix of **MEL16** and at the bottom-right the confusion matrix of **MEL15**. In both cases, the number of images within the class NV was lower than the value obtained in MEL14 (9314 in MEL15 and 10644 in MEL16).

In the last experiment, **MEL17**, we investigate the classification ability of the Swin V2 model with the combination of the dataset that gave the best result. We chose SwV2L as the model to evaluate, given its good performance on the HAM dataset. The DA used is DA_Train4, which is necessary to fit the size of the images to the input required by the model. The result obtained in accuracy on the test dataset is 83.65%.

Figure 4 shows the confusion matrices of the experiments **MEL4, MEL7, MEL8, MEL15, MEL13** and **MEL14**. From all these confusion matrices, we observe quite similar behavior; some classes are classified well, and others are affected by the numerosity of the class NV. In particular, the MEL classes carry many misclassifications under the class NV; the BKL class, on the other hand, misclassifies many images under the classes MEL and NV. In addition, the number of images belonging to the class NV also affects performance, as can be seen from the results. In light of this, we therefore decided to investigate the classification behavior in testing better.

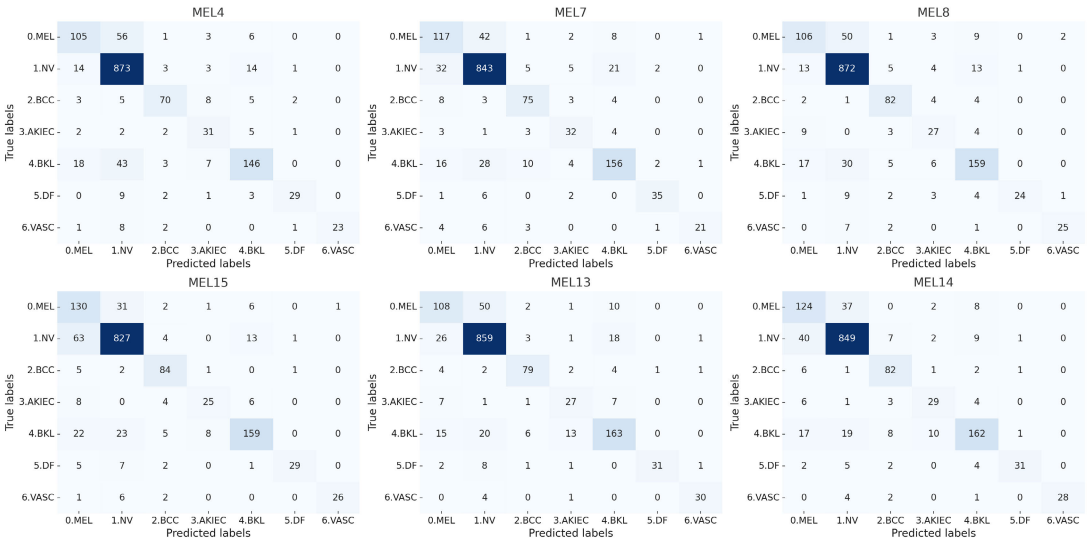


FIGURE 4. Confusion matrices of some of the experiments described in Table 7. From left to right, first raw: MEL4, MEL7, and MEL8 and second raw: MEL15, MEL13, and MEL14.

D. CLASSIFICATION OPTIMIZATION

A twofold evaluation strategy was adopted to optimize model classification performance on the test dataset and ensure accurate image classification. The first involved testing the experiment with higher accuracy for each group of experiments in the table 7. The training was repeated five times so that a statistical evaluation of the mean and standard deviation over the five training cycles could be made. Conversely, the second involved the random rotation of images during the inference process for each train to check the incidence of images in the classification. The rotation procedure consists of the following steps:

- Evaluation through n rotation cycles are iteratively performed. In each cycle, all images in the test dataset are randomly rotated before being submitted to the model for classification;
- After each rotation cycle, the model predictions for each image are recorded. At the end of all cycles, the list of predictions obtained for each image is analyzed, and the most frequent classification using the mode is calculated;
- For each image, the classification that obtained the highest number of occurrences during the model run over all rotation cycles is selected. This approach aims to improve classification accuracy by considering the variability introduced by image rotation.

Since it is impossible to decide a priori the optimal number of rotation cycles for this experiment, a random number of 500 was set to monitor the rotation behavior. After a series of tests, it was set at 100. This choice was motivated by the observation that the accuracy value tends to decrease beyond this threshold, as shown in Figure 5. This value represents an optimal compromise between the variation introduced by the rotations and the model's overall accuracy. Therefore, using evaluation cycles with random rotations aims to improve the robustness and accuracy of the classification model, allowing better generalization to images not seen during training.

The experiments considered for this additional experiment are **MEL7, MEL8** and **MEL14**. The model and the DA remain the same as in the previous experiment (see Table 7). Four new training sessions were conducted for each of these group experiments. In addition, an experiment on rotations was conducted for each result obtained. Table 8 shows the results of all experiments. We can see that almost all experiments conducted using the rotation technique on test images resulted in increased test performance. In particular, this technique demonstrated greater performance increases when used on the model trained on the LD, where we obtained an increase of 0.53%. In the same group of experiments, however, we also obtained a result where the final value was lower than the initial value by -0.26%. This confirms how images and the way they are used affect test performance.

TABLE 8. Additional experiments where the best models are towed five times, tested on the normal test dataset, and using rotations. EXP represents the Experiment's name, TA stays for Test Accuracy, TAwR is Test Accuracy with Rotations, and RV stays for Result Variation between TA and TAwR.

EXP	TA	TAwR	RV
MEL7_1	84.64	84.78	+0.14
MEL7_2	84.25	84.65	+0.4
MEL7_3	83.59	83.12	-0.47
MEL7_4	83.45	84.58	+1.13
MEL7_5	83.06	83.72	+0.66
MEL8_1	85.70	85.70	+0.0
MEL8_2	86.69	86.76	+0.07
MEL8_3	85.84	85.84	+0.0
MEL8_4	85.84	85.90	+0.06
MEL8_5	85.64	85.77	+0.13
MEL14_1	86.37	86.11	-0.26
MEL14_2	85.77	86.03	+0.26
MEL14_3	84.38	85.51	+0.13
MEL14_4	85.51	86.04	+0.53
MEL14_5	85.11	85.57	+0.46

This additional experiment also allows us to calculate the mean (6) and standard deviation(std 7) for each group of experiments, both in a more classical situation where evaluation is done on the test dataset and in a different condition where rotations are used as an additional method of evaluation. The mean is calculated as:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6}$$

where:

- x_i are the given values
- n is the number of values

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \tag{7}$$

Table 9 shows these values, and as one can see, the group of experiments with the highest result is *MEL8(1-5) Rotation*, obtaining 86% as the final value. However, it is important to note the obtained standard deviation value of 0.39%. This means that in our five training cycles, the values obtained are very close to each other and, therefore, little dispersed from the average. This low dispersion reflects good consistency among the measurements. In contrast, in *MEL7* and *MEL14*, we obtained a mean value of 83.80% and 85.43% respectively, with a fairly high std compared to the other experiments of 0.57% and 0.66%. Instead, this implies a more significant variability among the data, suggesting the presence of more heterogeneity in the data. In *MEL14(1-5)Rotations*, we get an average accuracy of 85.85% with a std of 0.25%. These results and the previous *MEL14(1-5)* highlight how the training and the test dataset strongly influence the test results.

TABLE 9. Additional experiments. EXP represents the name of the group of Experiments, and MEAN and STD are the statistical values calculated from the previous Table 8.

EXP	MEAN	STD
MEL7(1-5)	83.80	0.57
MEL7(1-5)Rotations	84.17	0.64
MEL8(1-5)	85.94	0.38
MEL8(1-5)Rotations	86.00	0.39
MEL14(1-5)	85.43	0.66
MEL14(1-5)Rotations	85.85	0.25

Intrigued by this behavior, we also decided to try slightly changing the number of images in the test dataset to see how much its changed might affect test performance.

Being aware of what was stated by Cassidy et al. [8] regarding the issues surrounding duplicate images, we decided to investigate the outcome. We applied the same duplicate image removal policy to the training dataset. We obtained a test dataset of 1222 images instead of 1511. We then tried to test the model that obtained the best performance in Table 7, namely MEL14, obtaining an accuracy value of 87.88%. Testing this approach on all the experiments in Table 7, we noticed an increase in performance of approximately 1-1.5%. This result confirms what Cassidy et al. [8] found to duplicate' great influence on calculating model performance. Moreover, during our experiments, we measured inference times for both the training and validation phases on the top three experiments (MEL7, MEL8, MEL14) using their respective datasets. Table 10 shows, in summary, the results of this measurement. When using the HAM dataset in **MEL7**, the training phase times ranged from 160 to 165 seconds, while the validation phase times were notably brief, around 13 to 14 seconds. When we combined the HAM and BCN datasets in **MEL8**, the training phase inference times ranged from 150 to 155 seconds, demonstrating the model's ability to handle combined data efficiently. For the validation phase, the inference times were slightly higher, ranging from 40 to 45 seconds. In contrast, the inference time in **MEL14** ranges between 830 and 860 seconds. The validation phase for the LD dataset consistently showed inference times between 185 and 215 seconds. These observations underline the model's computational demands and efficiency across different datasets, providing valuable insights into its performance and scalability.

TABLE 10. This table shows the inference times for the top three experiments (MEL7, MEL8, MEL14) for the train and validation datasets. EXP represents the name of the experiment, ITT refers to the Inference Time in training, and ITV refers to the Inference Time in Validation. Time values in ITT and ITV are expressed in seconds.

EXP	ITT	ITV
MEL7	160-165	13-14
MEL8	150-155	40-45
MEL14	860-830	185-215

VI. DISCUSSION

From previous experiments, it has emerged that the large disparity between classes negatively affects the learning of the models used. The experiments that obtained better results were those where the dataset was manipulated to reduce the disparity of images between the classes.

As we described at the outset, this work aimed to explore the classification capabilities of TMs and to lay the groundwork for an objective and fair comparison of the classification performance of DL models. The highest value obtained for the seven-class accuracy was 86.37%, which is a solid starting point. Even though we did not exceed the performance of the first place in the ISIC 2018 challenge (see 1), it is important to note that the winners used an external dataset to train their proposed model. Consequently, our results are not directly comparable without access to their dataset. This study considers only the datasets available in the literature. However, we can directly compare our results with the third runner-up in the challenge. From the experiments in which only HAM was used, the Swin network performed as well as the third runner-up and still showed a solid ranking. Since Swin networks, as we have seen, offer nontrivial potential, having many trainable parameters, modeling a dataset containing many more images led to a better and more accurate result.

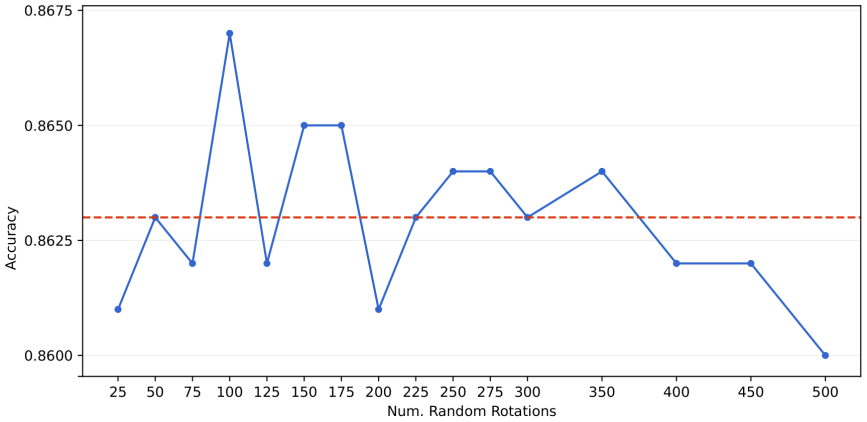


FIGURE 5. Classification accuracy as a function of the number of image rotations. The red line represents the test accuracy, while the blue line represents the test accuracy obtained after applying the rotations.

In addition, the proposed confusion matrices revealed a trend that raises an important issue: the misclassification among some classes. Analyzing Figure 2 in more detail shows that some images have similarities. The image classes carrying vascular problems (DF and VASC) have different visible characteristics than the other classes, and in fact, the classes that encapsulate vascular issues have very little misclassification with those that do not represent vascular issues (MEL, NV, BKL).

Moreover, the methodology proposed on rotations to test classification performance confirmed what was already suspected. A certain number of random rotations (see Figure 5) favored the model to classify diseases correctly. This suggests many aspects that could be worked on. The most relevant one concerns the heterogeneity of the data. Having a highly populated dataset would help the TMs perform better and learn to recognize more classification patterns. Given many trainable parameters, increasing the cardinality of the modeled LD would help the network increase its classification ability in testing.

In conclusion, while the network performs better in recognizing certain lesions, there are misclassification issues. These can be improved by increasing the number of images or by using different techniques to make the network learn new patterns for recognition and classification. All the experiments conducted from **MEL1** to **MEL17** consider datasets of dermatoscopic images only. In future work, we aim to investigate the influence of integrating macroscopic images of skin lesions for the classes used in this work. With this goal in mind, we have conducted three more experiments. **MEL18-MEL20** are preliminary results that integrate both dermatoscopic (LD) and macroscopic images during the training phase. This larger dataset (in progress) is composed at the moment of 52133 images: 41975 dermatoscopic ones plus 10158 macroscopic images taken from different datasets available in the literature [39], [46], [47], [48]. Only images from the seven classes used in the previous experiments were selected from each of these datasets. The experiments using the combined dataset are indicated with the subscript "Unified" in Table 7. In particular, **MEL19** applies SwL pre-trained and DA_Train with downsampled on the class NV to obtain a value of 84.58%. **MEL18** and **MEL20**, on the other hand, use SwL, not pre-trained. This choice stems from the fact that the number of images in this new dataset increases even more, making it also interesting to investigate the training from scratch of a Transformer model in future work.

VII. CONCLUSION

This paper explored using Transformer-based deep neural networks, specifically the ST model, for multiclass skin lesion classification. Our approach aimed to harness the self-attention mechanism intrinsic to TMs to capture intricate spatial dependencies within skin lesion images, bypassing the need for handcrafted features and extensive pre-processing steps. The performance of our proposed model was rigorously evaluated using a benchmark test set released in 2023, which includes ground-truth labels for various types of skin lesions, including melanoma.

Our experimental results demonstrate that the Transformer-based architecture achieves state-of-the-art performance in skin lesion classification, outperforming traditional CNNs and other DL models previously employed for similar tasks. The superior performance can be attributed to the model's ability to effectively manage long-range dependencies and spatial relationships in the image data, which are crucial for accurate medical image analysis.

A significant aspect of our study was exploring the impact of increased training data on model performance. By merging several smaller datasets to create a large comprehensive dataset, we enhanced the generalization capabilities of the Transformer network. This LD facilitated robust training, improving accuracy and reliability in skin lesion classification.

Furthermore, to promote transparency and reproducibility in research, we have made our benchmarks and the guide with

all the links of the dataset used to model the LD available on our GitHub repository. This contribution aims to provide a valuable resource for the research community, enabling further exploration and validation of Transformer-based approaches in medical imaging.

Our work underscores the potential of Transformer-based deep neural networks in advancing skin lesion classification, highlighting their clinical utility in aiding early and accurate skin cancer diagnosis. This research opens avenues for future studies to delve deeper into integrating advanced DL techniques in medical image analysis, ultimately contributing to developing powerful diagnostic tools for clinicians.

DATA AVAILABILITY

Notebooks are freely available at the link in the footnote present in the abstract. For the datasets, we have included references to all the datasets used, enabling them to be downloaded (https://github.com/UnluckyMirco/A-Large-Dataset-to-Enhance-Skin-Cancer-Classification-with-Transformer-Based-DNN). More information regarding the libraries used and replication of the proposed dataset can be found at the link shown.

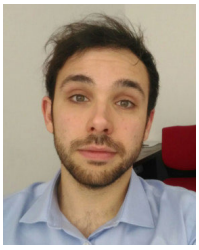
REFERENCES

[1] S. Waseh and J. B. Lee, "Advances in melanoma: Epidemiology, diagnosis, and prognosis," *Frontiers Med.*, vol. 10, Nov. 2023, Art. no. 1268479.
[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.
[3] Y. R. Woo, S. H. Cho, J. D. Lee, and H. S. Kim, "The human microbiota and skin cancer," *Int. J. Mol. Sci.*, vol. 23, no. 3, p. 1813, Feb. 2022.
[4] S. Chen, C. Han, X. Miao, X. Li, C. Yin, J. Zou, M. Liu, S. Li, L. Stawski, B. Zhu, Q. Shi, Z.-X. Xu, C. Li, C. R. Goding, J. Zhou, and R. Cui, "Targeting MC1R depalmitoylation to prevent melanomagenesis in redheads," *Nature Commun.*, vol. 10, no. 1, p. 877, Feb. 2019.
[5] D. Schadendorf, A. C. Van Akkooi, C. Berking, K. G. Griewank, R. Gutzmer, A. Hauschild, A. Stang, A. Roesch, and S. Ugurel, "Melanoma," *Lancet*, vol. 392, no. 10151, pp. 971–984, 2018.
[6] *The International Skin Imaging Collaboration*. Accessed: May 20, 2024. [Online]. Available: https://www.isic-archive.com/
[7] *Isic Challenge Webpage*. Accessed: Jul, 4, 2024. [Online]. Available: https://challenge.isic-archive.com
[8] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102305.
[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[10] *ISIC Challenge Leaderboards*. Accessed: May 20, 2024. [Online]. Available: https://challenge.isic-archive.com/leaderboards/2018/
[11] M. S. Ali, M. S. Miah, J. Haque, M. M. Rahman, and M. K. Islam, "An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models," *Mach. Learn. Appl.*, vol. 5, Sep. 2021, Art. no. 100036.
[12] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, Aug. 2018.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–12.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[15] S. Gulati and R. K. Bhogal, "Detection of malignant melanoma using deep learning," in *Proc. Int. Conf. Adv. Comput. Data Sci.* Singapore: Springer, 2019, pp. 312–325.
[16] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H.-C. Kim, "Skin cancer detection using deep learning—A review," *Diagnostics*, vol. 13, no. 11, p. 1911, 2023.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Dec. 2016, pp. 770–778.
[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[19] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*. Strasbourg, France: Springer, 2021, pp. 13–23.
[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4278–4284.
[21] 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, and 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 *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.
[22] A. Yilmaz, M. Kalebasi, Y. Samoylenko, M. E. Guvenilir, and H. Uvet, "Benchmarking of lightweight deep learning architectures for skin cancer classification using ISIC 2017 dataset," 2021, *arXiv:2110.12270*.
[23] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it time to replace CNNs with transformers for medical images?" 2021, *arXiv:2108.09038*.
[24] R. Pedro and A. L. Oliveira, "Assessing the impact of attention and self-attention mechanisms on the classification of skin lesions," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.
[25] Y. Gulzar and S. A. Khan, "Skin lesion segmentation based on vision transformers and convolutional neural networks—A comparative study," *Appl. Sci.*, vol. 12, no. 12, p. 5990, Jun. 2022.
[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.
[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[28] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang, S. Xu, and H. Chen, "An improved transformer network for skin cancer classification," *Comput. Biol. Med.*, vol. 149, Oct. 2022, Art. no. 105939.
[29] S. Hao, L. Zhang, Y. Jiang, J. Wang, Z. Ji, L. Zhao, and I. Ganchev, "ConvNeXt-ST-AFF: A novel skin disease classification model based on fusion of ConvNeXt and Swin transformer," *IEEE Access*, vol. 11, pp. 117460–117473, 2023.
[30] S. Ayas, "Multiclass skin lesion classification in dermoscopic images using swin transformer model," *Neural Comput. Appl.*, vol. 35, no. 9, pp. 6713–6722, Mar. 2023.
[31] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102357.
[32] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imag.*, vol. 5, no. 3, p. 1, Jul. 2018.
[33] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.
[34] D. Wen, S. M. Khan, A. Ji Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C. de Blas Perez, A. K. Denniston, X. Liu, and R. N. Matin, "Characteristics of publicly available skin cancer image datasets: A systematic review," *Lancet Digit. Health*, vol. 4, no. 1, pp. e64–e74, Jan. 2022.
[35] *HAM10000 Test Set Release*. Accessed: May 2024. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T
[36] (2020). *Consecutive Biopsies for Melanoma Across Year 2020*. [Online]. Available: https://doi.org/10.34970/151324

[37] *International Skin Imaging Collaborations. ISIC Archive*. Accessed: Jul. 4, 2024. [Online]. Available: https://api.isic-archive.com/images/?query=&collections=289
[38] S. M. de Faria, J. N. Filipe, P. M. Pereira, L. M. Tavora, P. A. Assuncao, M. O. Santos, R. Fonseca-Pinto, F. Santiago, V. Dominguez, and M. Henrique, "Light field image dataset of skin lesions," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2019, pp. 3905–3908.
[39] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 538–546, Mar. 2019.
[40] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, and D. Gutman, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, no. 1, p. 34, Jan. 2021.
[41] M. A. R. Lara, M. V. R. Kowalczuk, M. L. Eliceche, M. G. Ferraresso, D. R. Luna, S. E. Benitez, and L. D. Mazzuoccolo, "A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population," *Sci. Data*, vol. 10, no. 1, p. 712, Oct. 2023.
[42] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," 2019, *arXiv:1908.02288*.
[43] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "PH 2-a dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2013, pp. 5437–5440.
[44] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, and L. Dong, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12009–12019.
[45] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, and R. Hofmann-Wellenhof, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, Web-based, international, diagnostic study," *Lancet Oncol.*, vol. 20, no. 7, pp. 938–947, Jul. 2019.
[46] A. G. C. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," *Comput. Biol. Med.*, vol. 116, Jan. 2020, Art. no. 103545.
[47] I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov, "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6578–6585, Nov. 2015.
[48] S. S. Han. (2019). *SNU Dataset + Quiz*. [Online]. Available: https://figshare.com/articles/dataset/SNU_SNU_MELANOMA_and_Reddit_dataset_Quiz/6454973

MIRCO GALLAZZI received the bachelor's and master's degrees in computer science from the University of Insubria, Varese, Italy, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree in computer science. He conducts research on skin lesion classification with deep learning models. He is interested in the classification of EEG signals in moto imagery using deep learning. He is also interested in machine learning applications in the medical field.

SARA BIAVASCHI is currently pursuing the bachelor's degree majoring in computer science with the University of Insubria, Varese, Italy. She collaborates with Prof. Gallo and Prof. Corchs on a research project focused on the classification of skin tumors. She has a strong interest in deep learning algorithms. She is also passionate about developing a mobile application that leverages artificial intelligence to make early skin cancer detection technology available to everyone for free.



ALESSANDRO BULGHERONI is currently pursuing the bachelor's degree majoring in computer science with the University of Insubria, Varese, Italy. He collaborates with Prof. Gallo and Prof. Corchs on a research project focused on the classification of skin cancers. He was a Data Analyst with consulting company, Milan, Italy, for more than five years. He has a strong interest in deep learning algorithms. He is very passionate about data visualization.

TOMMASO M. GATTI is currently pursuing the bachelor's degree majoring in computer science with the University of Insubria, Varese, Italy. He has collaborated with Prof. Gallo and Prof. Corchs on a research project focused on recognition and classification of skin tumor through deep learning applications.

SILVIA CORCHS received the master's and first Ph.D. degrees in physics from the University of Rosario, Argentina, and the second Ph.D. degree in informatics from the University of Milano Bicocca, Italy, in 2014. She has been a Research Scientist with the National Council for Research, Argentina, working in theoretical physics. From 2000 to 2003, she was a Guest Scientist with the Computational Neuroscience Group, Research Department of Siemens AG, Munich, Germany, focusing on neurodynamical modeling of the visual attention mechanism. From 2005 to 2007, she was a Postdoctoral Researcher and a Teaching Assistant with the Vision and Perception Science Laboratory, University of Ulm, Germany. From 2008 to 2021, she was a Research Scientist, working in the field of image processing; and a Teaching Assistant with the Department of Informatics, Systems and Communication, University of Milano Bicocca. Since February 2022, she has been a Senior Research Scientist with the Department of Theoretical and Applied Sciences, University of Insubria, Varese, Italy. Her current research interests include artificial intelligence, machine learning, and multimodal signal processing.

IGNAZIO GALLO received the degree in computer science from the University of Milan, Italy, in 1998. From 1998 to 2002, he was with the Artificial Intelligence and Soft Computing Laboratory, National Research Council (CNR), Milan. He has defined and developed neural models for classifying and recognizing remote-sensing images. From 2002 to 2003, he was with the "Informatica e Comunicazione" Department, University of Insubria, Varese, dealing with simulators in a distributed environment. Since 2004, he has been an Assistant Professor with the University of Insubria. He is interested in stereoscopic reconstruction analysis of 3-D images produced by a scanning electron microscope and in feature selection and classification methods applied to hyperspectral data. He conducts research in deep learning, image processing, pattern recognition, neural computing, computer vision, and natural language processing. His current research interests include the design and application of deep models for object detection and recognition from digital images.

Open Access funding provided by 'Università degli Studi dell'Insubria' within the CRUI CARE Agreement