IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 March 2025, accepted 26 March 2025, date of publication 1 April 2025, date of current version 9 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3556479

RESEARCH ARTICLE

CNN Issues in Skin Lesion Classification: Data Distribution and Quantity

GIULIANA RAMELLA[1], (Member, IEEE), AND LUCA SERINO[2]
[1]Institute for the Applications of Calculus "M. Picone," National Research Council (CNR), 80131 Naples, Italy
[2]Institute for High Performance Computing and Networking, National Research Council (CNR), 80131 Naples, Italy

Corresponding author: Giuliana Ramella (giuliana.ramella@cnr.it)

This research has been accomplished within Italian Mathematical Union (UMI) Research Groups: Approximation Theory and Applications (TAA) and Mathematics for Artificial Intelligence and Machine Learning (MAT AI&ML), and in part by Gruppo Nazionale per il Calcolo Scientifico (GNCS) Istituto Nazionale di Alta Matematica (INdAM).

ABSTRACT Convolutional Neural Networks (CNNs) have become indispensable tools in skin cancer classification, aiding clinical experts to achieve earlier and more accurate diagnoses, improving treatment outcomes, and driving advancements in medical research. Despite their pivotal role, the most popular CNN architecture families exhibit a critical issue related to the distribution and quantity of available data, potentially compromising their performance and generalization abilities. This challenge is commonly overlooked in most skin lesion classification papers, which mainly rely on weighted classification techniques. Directly using appropriately dataset balancing or Transfer Learning (TL) methods, as suggested in recent studies, has the potential to deliver more satisfactory results, providing a more effective approach to addressing this issue. In the effort to tackle this problem, we provide a comprehensive quantitative evaluation aimed at identifying the most critical new emerging computational aspects and the related effective techniques. Specifically, we propose twelve Computational Models (CMs) based on four prominent CNN models with increasing structural complexity. We assess their effectiveness in both pretrained and unpretrained versions, incorporating TL strategies as well. Our experiments focus on the ISIC 2018 image dataset, a benchmark widely recognized for its extensive application in skin cancer research yet challenged by significant class imbalance issues. To mitigate this, we also randomly extracted a balanced image subset from ISIC 2018 for evaluation purposes. The experimental results, regarding four different scenarios, provide valuable insights into the design and utilization of CNNs for skin lesion classification, laying the groundwork for further investigations.

INDEX TERMS Balanced image dataset, convolutional neural network, dermoscopic image, skin lesion classification, transfer learning, unbalanced image dataset.

I. INTRODUCTION

Convolutional Neural Networks (CNNs) have proven to be an indispensable tool in the fight against cancer [1], [2], [3] and in many other applicative fields [4], [5], [6], [7], [8]. In particular, CNNs excel at analyzing large image datasets to classify suspicious skin patterns and assist dermatologists at every main stage of the oncological journey, from early diagnosis to treatment. Furthermore, CNNs play a pivotal role in advancing research since they aid in a deeper understanding of the skin's biological aspects and potential therapeutic targets [9], [10], [11].

The main CNN architectures [12] used for skin cancer classification stem from VGGNet [13], InceptionNet [14], MobileNet [15], ResNet [16], EfficientNet [17] and DenseNet [18] family. These architectures, coupled with appropriate modifications and integration with other computational models, have yielded promising outcomes and driven notable progress in skin lesion classification. Nevertheless, their application remains challenging due to numerous unresolved issues that can significantly compromise the CNNs' performance.

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell.

The main challenges can be divided into two categories. The first pertains to the quality of the processed images, here abbreviated as the ''IQ issue''. The second is associated with the distribution/quantity of images to be processed, shortly indicated as the ''ID issue''. These issues also have a prominent role in other tasks related to skin lesion studies, such as color quantization [19], [20], [21], [22] and segmentation [23], [24], [25].

In ''IQ issue''category, visual fidelity [26], [27], [28] may be compromised by artifacts resulting from the removal of skin hair, imperfections in the imaging process, or inadequate resolution, all of which can either highlight or obscure crucial details. Artifacts are typically managed through preprocessing techniques aimed at their removal [29], [30], [31], [32] before feeding image data into CNNs. Alternatively, Data Augmentation (DA) methods, such as rotation, scaling, and noise addition, are employed to mitigate the impact of artifacts during training and improve the model's robustness in handling them [33], [34]. The strategy primarily used to mitigate the effects of image resolution encompasses appropriate image resizing techniques [35], [36], [37], [38], and/or specialized architectures tailored to handle diverse resolutions effectively [39], [40], [41].

In ''ID issue''category, data distribution significantly impacts the lesion classification process since an unbalanced distribution in the initial dataset can induce the classifier to be biased towards the particular class holding more samples, causing overfitting problems [9], [42], [43]. Moreover, the data quantity also influences the lesion classification process since a limited dataset can also lead to limitations in the generalization process, thus affecting the performance of the CNN [1], [9]. The consequences of an inadequate data distribution/quantity are usually overlooked in most papers, primarily focused on identifying specific architectures and providing their results on unbalanced limited datasets [44]. Typically, the performance of CNN is evaluated without considering the ID issue or simply trying to consider this disparity in number and distribution by using weighted classification criteria. Moreover, most comparative studies of the main CNN architectures for skin lesion classification disregard the ID issue.

In the alternative, when the issue is considered, a few recent papers propose computational techniques to address it. Specifically, DA methods are employed to address the ID issue by generating variations of under-represented images, thereby increasing the overall dataset size and achieving a more balanced class distribution. Transfer Learning (TL) techniques are considered well-suited for addressing data distribution challenges within the class imbalance issue, as they leverage knowledge gained from a larger source dataset to improve model performance by adapting it to the target data [45], [46], [47], [48], [49], [50]. In particular, in the context of CNNs, TL involves using a ''pretrained CNN model'' so-called since it has been trained on a large dataset, often ImageNet [51]. Then, the pretrained CNN model is fine-tuned on a smaller dataset for a different but related task [52], [53], [54]. Beyond delivering more reliable results, TL techniques also alleviate the need to train a model from scratch, significantly reducing the required computational time and resources. However, TL approaches often overlook the use of balanced datasets, which are rarely considered and employed [47], [48], [55], [56].

In this paper, we focus on the ID open issue, with new computational aspects emerging, referring back to the IQ issue for future investigations. In particular, we aim to identify the most critical new emerging ID computational aspects and the related effective techniques.

To accomplish this, we conduct a comprehensive quantitative and statistical evaluation by introducing twelve Computational Models (CMs) of increasing structural complexity. These CMS are inspired by the aforementioned leading CNN model families, which serve as base Classification Models (bCMs). For each CM, we define both pretrained and unpretrained configurations. In the pretrained configurations, we employ a TL approach, where the base Classification Models (bCMs) are frozen.

As a dataset, we use ISIC 2018 [57], which is widely used in the literature [58], despite its significant imbalance. Since one of our primary objectives is to assess the CMS' performance both on balanced and unbalanced datasets, we apply a classical balancing technique. Specifically, we extract two datasets with the same total image number from ISIC 2018: one created by randomly selecting images from the various classes, and the other constructed by applying a balancing technique to ensure each class is equally represented. Then, we quantitatively and statistically evaluate and compare the CNNs' performance across four distinct scenarios: 1. Pretrained CMs on an unbalanced dataset; 2. Pretrained CMs on a balanced dataset; 3. Unpretrained CMs on an unbalanced dataset; 4. Unpretrained CMs on a balanced dataset.

This evaluative analysis provides valuable insights regarding the favorable scenarios, the reference CMs, the practical implications and improvement strategies (see the end of Section IV). Our findings could guide future developments in the design and application of CNNs for skin lesion classification, serving as a foundation for further investigation.

The remainder of the paper is organized as follows. Section II provides a detailed description of the materials and the methodology, including the CMs employed in the experimental phase. Section III and Section IV present the quantitative and statistical evaluation of the CMs' performance across all scenarios and our interpretation of the experimental results, respectively. Finally, Section V concludes the paper.

II. MATERIALS AND METHODOLOGY

This section outlines the evaluative analysis by sequentially describing the (un)balanced image datasets (see Subsection II-A); the designed CMs employed in the experimental phase (see Subsection II-B); the adopted evaluation metrics (see Subsection II-C); the data preparation and

preprocessing (see Subsection II-D); and the training process of all CMs (see Subsection II-E).

A. UNBALANCED AND BALANCED IMAGE DATASETS

For the validation of the considered CMs, we use the multi-class dermoscopic image dataset made available by the International Skin Imaging Collaboration (ISIC) [59]. Specifically, we refer to the ISIC challenge dataset for 2018 [60], shortly indicated as ISIC 2018, which is widely used for skin lesion classification tasks in many scientific papers. The ISIC 2018 challenge includes three subtasks, focusing on Lesion Segmentation (Task1), Lesion Attribute Detection (Task2) and Disease Classification (Task3). We consider the three image datasets released after the closure of the ISIC Task3 Challenge. For clarity, we denote these datasets as ISIC-Train, ISIC-Validate, and ISIC-Test to distinguish them from the subsets defined below. All images, originating from [57] and consequently from the ISIC Task 3 Challenge, are provided in 8-bit RGB format with a resolution of 450 × 600 pixels. The images are distinct in 7 diagnostic classes (lesion types): AKIEC (Actinic Keratoses IntraEpithelial Carcinoma); BCC (Basal Cell Carcinoma); BKL (Benign Keratosis Lesion); DF (DermatoFibroma); MEL (Melanoma); NV (Melanocytic Nevi); VASC (Vascular Lesion) and distributed according to Table 1. Image data are accompanied by a metadata file containing the descriptive information for each lesion image, such as the age, the gender, the location, the lesion class, and so on. From Table 1 and the diagram reported in Figure 1 a), it is apparent that these classes are not equally distributed: most of the images belong to the class NV, while the classes AKIEC, BCC, DF, and VASC are the least populated.

Due to the imbalance of the ISIC 2018 (and ISIC Task 3 Challenge) dataset, we performed a data balancing operation. This enables us to evaluate the performance of each CM on a balanced dataset, providing insights into the differences compared to results obtained from the unbalanced dataset.

TABLE 1. Diagnostic classes with the corresponding image number for each ISIC Task 3 Challenge subset.

Class	Description	ISIC - Train	ISIC - Validation	ISIC - Test	Total for class
AKIEC	Actinic Keratoses IntraEpithelial Carcinoma	327	8	43	378
BCC	Basal Cell Carcinoma	514	15	93	622
BKL	Benign Keratosis Lesion	1099	22	217	1338
DF	DermatoFibroma	115	1	44	160
MEL	Melanoma	1113	21	171	1305
NV	Melanocytic Nevi	6705	123	909	7737
VASC	Vascular lesion	142	3	35	180
Total for subset		10015	193	1512	11720

Data balancing can be obtained by operating both at the ''data level'', i.e., changing the training set and its distribution, and at the ''method level'', i.e., by keeping the dataset unchanged and adjusting the inference criteria for classification [43]. We use a ''data level'' balancing method and avoid a method-level balancing, even partially, to prevent the evaluative analysis from depending on the type of chosen balancing method. The ''data level'' balancing can be performed through oversampling or undersampling operations. According to oversampling, randomly selected images from minority classes are suitably replicated and augmented to reach the same number of instances in each class. As opposed to oversampling, images are removed randomly from the majority class until all classes have the same number of samples when an undersampling technique is applied. There is some evidence that, in our case, undersampling is preferable to oversampling [43] even if a portion of available data is discarded. The main motivation is that the evaluation of the CMs' performance is not influenced by any preventive augmentation strategies aimed at addressing the issue of data imbalance.

Hence, we focus solely on images of ISIC Task 3 Challenge belonging to classes with the highest population indices, i.e., MEL (with 1305 images), 2. NV (with 7737 images), 3. BKL (with 1338 images) for a total of 10380 images. We indicate this dataset as sISIC 2018. This restriction results in a negligible reduction in the volume of input data, with approximately 88% of the data remaining available. Moreover, the diagnostic class distribution is such that we have to drastically reduce the number of input images to perform data balancing. Consequently, we have restricted ourselves to considering a dataset with 3000 images, reducing the dataset to about 30%. This drastic reduction in the number of images does not compromise the result's reliability. However, this reduction can affect the classification rate, potentially leading to a lower rate than using the same model on the entire dataset.

TABLE 2. Diagnostic classes with the corresponding image number for UDS subset.

Class	UDS - Training	UDS - Validation	UDS - Test	Total for class
BKL	259	84	43	386
MEL	251	74	35	360
NV	1590	442	222	2254
Total for subset	2100	600	300	3000

TABLE 3. Diagnostic classes with the corresponding image number for BDS subset.

Class	BDS - Training	BDS - Validation	BDS - Test	Total for class
BKL	700	200	100	1000
MEL	700	200	100	1000
NV	700	200	100	1000
Total for subset	2100	600	300	3000

Following the consolidated choice of 70, 20, and 10 percent, we divided the input dataset of 3000 images randomly selected from sISIC 2018 into training, validation, and test sets comprising 2100, 600, and 300 images, respectively. The training and validation sets are used during the learning phase, while the test set is reserved for evaluating the CMs' performance.

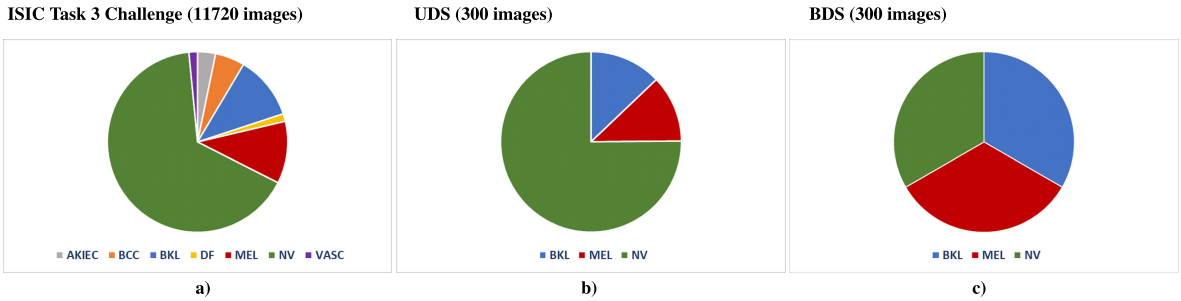

FIGURE 1. Image class distribution of each dataset.

In the unbalanced case, images for each set (training, validation, test) are randomly selected based on the above-mentioned quantities, without considering class membership. We refer to this dataset as Unbalanced DataSet, shortly indicated as UDS, with its subsets named UDS-Training, UDS-Validation, and UDS-Test, corresponding to the training, validation, and test sets. The class distribution of UDS is shown in Table 2 and Figure 1 b). As expected, UDS contains a higher proportion of images classified as NV due to the initial distribution of the sISIC 2018 dataset. Nonetheless, the less represented classes (BKL and MEL) maintain proportions consistent with those in the original dataset.

In the balanced case, images for each set (training, validation, and test) are randomly selected from sISIC 2018 based on the above quantities, ensuring equal partitioning across classes. Specifically, for each set, we randomly select $700 \times 3 = 2{,}100$ images for training, $200 \times 3 = 600$ images for validation, and $100 \times 3 = 300$ images for testing. We refer to this dataset as the Balanced DataSet (BDS), with its subsets named BDS-Training, BDS-Validation, and BDS-Test, corresponding to the training, validation, and test sets. The class distribution of BDS is shown in Table 3 and Figure 1c.

B. THE CLASSIFICATION MODELS

As already stated, CNNs are highly effective tools, especially for feature extraction and image classification within the intricate landscape of deep learning architectures [61]. Among pretrained architectures, notable models such as DenseNet121 [18], InceptionV3 [62], MobileNet [15], and VGG16 [63] have gained significant popularity due to their robust performance in skin lesion classification. DenseNet121 is part of the DenseNet family, which also includes variants like DenseNet169 and DenseNet201, offering different trade-offs between model depth and computational efficiency. Similarly, InceptionV3 belongs to the Inception family — building on earlier versions such as InceptionV1 and InceptionV2 and evolving into more advanced forms like InceptionV4 and Inception-ResNet — thereby providing a versatile framework for complex image classification tasks. Additionally, VGG16 is among several models in the VGG family, comprising variants such as VGG11, VGG13, and VGG19. These models, here already shortly indicated as bCMs, have been trained on an extensive dataset of over a million images from the ImageNet Large-Scale Visual Recognition Challenge [64] and are readily accessible as pretrained models.

The proposed CMs are generated in the following way. For each bCM, we append an ‘‘output layer’’ using the softmax activation function to produce a probability distribution and generate classifications across three classes. In this way, we generate four CMs labeled with the name of the corresponding bCM followed by an acronym indicating version V0. For instance, the CM based on DenseNet121 in version V0 is denoted as DenseNet121 V0.

To explore more complex versions, we developed other CMs using standard strategies [47], [48], starting from the bCMs to suit the specific target task. Specifically, before the already ‘‘output layer’’ inserted in V0, we add new layers, creating other two versions labeled as V1 and V2.

Indeed, the CMs in the V1 and V2 versions are obtained by a) the addition of dense layers to enable the model to learn complex relationships among the extracted features and leverage them effectively; b) the inclusion of dropout layers to improve robustness against overfitting, albeit at the cost of slower training. In particular, for the V1 versions, a 256-unit dense layer followed by a 50% dropout and a 64-unit dense layer is added to the corresponding bCMs before the softmax activation function. For the V2 versions, the corresponding bCMs are modified with a sequence of layers applied before the softmax activation function, as follows: a 256-unit dense layer followed by a 50% dropout, a 128-unit dense layer followed by a 50% dropout, a 64-unit dense layer followed by a 50% dropout, and finally, a 32-unit dense layer. Similarly, for the V1 and V2 versions, we distinguish the CMs by appending the version to the name of the corresponding bCM. For instance, DenseNet121 V1 refers to the CM based on DenseNet121 in version V1. The overall structure of the CMs, grouped by version and illustrating all layers and dropout configurations, is presented in Figure 2. The activation function ReLU is used for each CM, and both pretrained and unpretrained configurations are created, as previously stated. In particular, in the pretrained configuration, a fine-tuning technique is employed, freezing the corresponding bCM. A summary of the structure of the CMs is provided in Table 4.

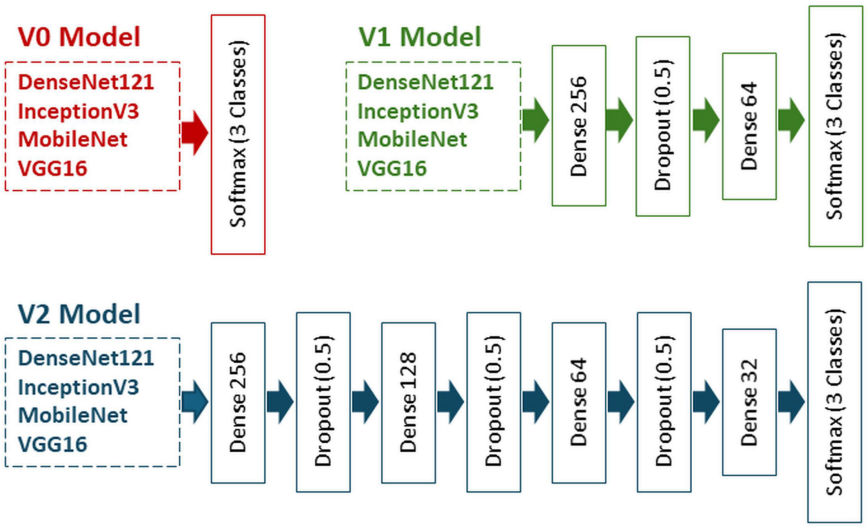

FIGURE 2. The Classification Models (CMs).

Table 5 summarizes the parameters of the CMs, where TP denotes the total parameters, TTP the total trainable parameters, and NTP the total non-trainable parameters, which remain constant for both pretrained and unpretrained CMs. Additionally, PTL (UTL) indicates the total trainable layers for the pretrained (unpretrained) CMs. The table also includes the number of Giga Floating Point Operations per Second (GFLOPs) for each CM.

From these tables, it can be inferred that DenseNet121, InceptionV3, MobileNet, and VGG16 exhibit a progressively increasing total number of trainable parameters. This number approximately doubles (or triples) when transitioning from the V0 to the V1 version and remains essentially unchanged when moving from V1 to V2.

The CMs are implemented using the Keras deep learning framework (version 2.4.3), with Tensorflow (version 2.3.0) serving as the backend. All experiments have been conducted on a single workstation with an Intel Xeon Gold Core 6134 3.20 GHz CPU, 128 GB of RAM, and an NVIDIA Quadro 6000 with 24 GB of installed memory.

C. EVALUATION METRICS

The CMs' performance is assessed in the following four scenarios (defined also in I).

- ⇒ **Scenario 1 (S1) - Pretrained CMs on UDS**
- ✎ **Scenario 2 (S2) - Pretrained CMs on BDS**
- ◇ **Scenario 3 (S3) - Unpretrained CMs on UDS**
- ➢ **Scenario 4 (S4) - Unpretrained CMs on BDS**

For each scenario, we quantitatively assess the CMs' performance by metrics widely adopted in the literature. Precisely, during the training phase, are employed the loss, computed as categorical cross-entropy, and the accuracy A, defined as:

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

During the testing step, in the four scenarios, besides the Accuracy (1), we compute the classical metrics: Precision (P), True Positive Rate (TPR), also known as Recall (R), F1 Score ($F1$) and Mattews Correlation Coefficient (MCC).

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$TPR = R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

We also consider the confusion matrices (cms), as they provide detailed insights into model performance by identifying which classes are accurately distinguished or often misclassified. This analysis is particularly valuable for understanding performance on the unbalanced dataset.

Lastly, to comprehensively evaluate the model's discriminative capability, we also utilize the widely recognized measure AUC (Area Under the Curve) - ROC (Receiver Operating Characteristic). This measure, in addition to the True Positive Rate (TPR), incorporates the computation of the False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

Since AUC-ROC was originally defined for binary classification, we use its multiclass classification extension, sensitive

TABLE 4. Summary of the structure of the CMs.

	Layer(type)	Output shape	#Params
DenseNet121	**densenet121**	**(None, 7, 7, 1.024)**	**7.037.504**
V0	Flatten	(None, 50.176)	0
	Dense	(None, 3)	150.531
V1	Flatten	(None, 50.176)	0
	Dense	(None, 256)	12.845.312
	Droput	(None, 256)	0
	Dense	(None, 64)	16.448
	Dense	(None,3)	195
V2	Flatten	(None, 50.176)	0
	Dense	(None, 256)	12.845.312
	Droput	(None, 256)	0
	Dense	(None, 128)	32.896
	Droput	(None, 128)	0
	Dense	(None, 64)	8.256
	Droput	(None, 64)	0
	Dense	(None,32)	2.080
	Dense	(None,3)	99
InceptionV3	**inception_v3**	**(None, 5, 5, 2.048)**	**21.802.784**
V0	Flatten	(None, 51.200)	0
	Dense	(None, 3)	153.603
V1	Flatten	(None, 51.200)	0
	Dense	(None, 256)	13.107.456
	Droput	(None, 256)	0
	Dense	(None, 64)	16.448
	Dense	(None,3)	195
V2	Flatten	(None, 51.200)	0
	Dense	(None, 256)	13.107.456
	Droput	(None, 256)	0
	Dense	(None, 128)	32.896
	Droput	(None, 128)	0
	Dense	(None, 64)	8.256
	Droput	(None, 64)	0
	Dense	(None,32)	2.080
	Dense	(None,3)	99
MobileNet	**mobilenet**	**(None, 7, 7, 1.024)**	**3.228.864**
V0	Flatten	(None, 50.176)	0
	Dense	(None, 3)	150.531
V1	Flatten	(None, 50.176)	0
	Dense	(None, 256)	12.845.312
	Droput	(None, 256)	0
	Dense	(None, 64)	16.448
	Dense	(None,3)	195
V2	Flatten	(None, 50.176)	0
	Dense	(None, 256)	12.845.312
	Droput	(None, 256)	0
	Dense	(None, 128)	32.896
	Droput	(None, 128)	0
	Dense	(None, 64)	8.256
	Droput	(None, 64)	0
	Dense	(None,32)	2.080
	Dense	(None,3)	99
VGG16	**vgg16**	**(None, 7, 7, 512)**	**14.714.688**
V0	Flatten	(None, 25.088)	0
	Dense	(None, 3)	75.267
V1	Flatten	(None, 25.088)	0
	Dense	(None, 256)	6.422.784
	Droput	(None, 256)	0
	Dense	(None, 64)	16.448
	Dense	(None,3)	195
V2	Flatten	(None, 50.176)	0
	Dense	(None, 256)	6.422.784
	Droput	(None, 256)	0
	Dense	(None, 128)	32.896
	Droput	(None, 128)	0
	Dense	(None, 64)	8.256
	Droput	(None, 64)	0
	Dense	(None,32)	2.080
	Dense	(None,3)	99

to class imbalance, by computing it of each class against all others.

We remark that all experimental measures (see Section III) in this paper have been computed as average values using the robust Keras libraries without sample weighting.

TABLE 5. Parameters, layers and GFLOP of the CMs model.

	TP	TTP	NTP	PTL	UTL	GFLOP
DenseNet121						
V0	7.188.035	7.104.387	83.648	2	364	2,9
V1	19.899.459	19.815.811	83.648	6	368	2,9
V2	19.926.147	19.842.499	83.648	10	372	2,9
InceptionV3						
V0	21.956.387	21.921.955	34.432	2	190	5,70
V1	34.926.883	34.892.451	34.432	6	194	5,73
V2	34.953.571	34.919.139	34.432	10	198	5,74
MobileNet						
V0	3.379.395	3.357.507	21.888	2	83	0,57
V1	16.090.819	16.068.931	21.888	6	87	0,59
V2	16.117.507	16.095.619	21.888	10	91	0,61
VGG16						
V0	14.789.955	14.789.955	0	2	28	15,30
V1	21.154.115	21.154.115	0	6	32	15,31
V2	21.180.803	21.180.803	0	10	36	15,31

D. DATA PREPARATION AND PREPROCESSING

Due to the architecture of the CMs, every RGB image is resized to 224 × 224 utilizing the bicubic interpolation method. This ensures that the resulting performance meets that is resulted to be an acceptable standard [38]. We also adopt a common preprocessing step that normalizes the value of each component pixel in the interval [0,1] before feeding them into a CNN. This preprocessing helps in improving the computation and stability of the whole analysis.

Unlike existing papers that undergo a preprocessing primarily aimed at improving classification, such as a sharpening filter for enhancing the image contrast [44], we prefer to work with raw data. This decision reflects our focus on evaluating model performance without introducing alterations caused by preprocessing. Consequently, this approach may result in lower classification values compared to values reported in papers adopting such models with adequate preprocessing. On the contrary, to remain consistent with the standard use, we employ classical standard augmentation techniques, such as scaling rotation, etc., in all CMs.

E. TRAINING PROCESS

Common challenges occurring during the training of a CM include overfitting, limited data availability, high computational demands, and difficulties in optimizing the network architecture. As mentioned above, TL is a technique that can help to alleviate or resolve some of these issues.

Thus, we apply TL in the scenarios 1 and 2 in the following way. We consider the pretrained bCM on the corresponding (un)balanced dataset and use the Imagenet weights as the initial weights to the training process of each CM in version V0, V1, and V2, freezing the layers of bCM to prevent them from being updated during training. In scenarios 3 and 4, we train the CMs on the corresponding (un)balanced dataset using initial weights randomly selected. The hyperparameters used for training each CM are the same in all scenarios. Their values are given in Table 6. We remark that during the training trials of the CMs, even after 100 epochs, there was no noteworthy improvement in accuracy, i.e., the accuracy fluctuations remained within a consistent range. This trend was also reflected in the loss values. Consequently, it became evident that extending the training duration beyond 100 epochs did not yield any

TABLE 6. Training parameters of each CM.

Parameter	Value
Batch size (training)	20
Batch size (validation)	10
Activation function	ReLu
Optimizer	RMSprop (learning rate=$1e^{-4}$)
Number of epochs	100

significant performance enhancements. Therefore, limiting the training process to 100 epochs was deemed appropriate.

To choose the optimizer, we conducted controlled experiments on both balanced and unbalanced datasets, evaluating the accuracy, the stability of the training process, and the convergence speed. Based on the comprehensive evaluation of these factors, we opted for the RMSprop optimizer since it does not penalize the performance across all the considered CMs without penalizing their effectiveness.

Each CM was trained multiple times for each scenario, yielding diverse sets of result data. Pretrained CMs showed minimal variation between runs, while unpretrained CMs exhibited greater differences.

The average training times over 5 runs are reported as 1 hour plus minutes:seconds for the four scenarios in Table 7. This table reveals that MobileNet consistently stands out as the fastest model across all scenarios, particularly in Scenario 2, while VGG16 is the slowest. DenseNet121 and Inception V3 exhibit comparable training times, with slight variations depending on the scenario. Notably, Scenario 2 emerges as the most efficient scenario in most cases, especially for MobileNet and VGG16. In contrast, Scenario 3 tends to demand more training time compared to the other scenarios.

TABLE 7. Training times reported as 1 hour + minutes:seconds for the four scenarios.

	S1	S2	S3	S4
DenseNet121				
V0	48:58.744	49:10.711	50:30.557	48:43.355
V1	48:42.095	49:18.921	49:50.614	48:49.159
V2	48:27.169	49:58.102	49:13.957	48:51.407
InceptionV3				
V0	49:27.023	49:42.916	49:08.521	48:22.640
V1	48:44.093	48:34.591	48:43.353	49:33.543
V2	48:42.979	48:44.982	49:03.834	49:26.953
MobileNet				
V0	48:41.402	45:37.933	49:08.650	49:01.960
V1	48:19.573	45:18.229	49:05.402	48:42.406
V2	48:44.041	45:09.495	48:37.584	48:05.267
textbfVGG16				
V0	49:52.793	46:42.991	48:07.334	47:36.566
V1	49:11.756	46:29.485	47:33.044	48:32.079
V2	49:23.804	46:18.094	47:36.899	48:09.781

Further details about the training process can be found in the supplementary material where Accuracy and loss change graphs for each CM during one of the performed training and validation steps for all scenarios are available. From these, as expected, apart from a few cases representing anomalies where convergence is achieved after a few epochs, we observe the decrease in loss values and the increase in accuracy values during the training and validation phases for each CM.

III. EXPERIMENTAL RESULTS

This section presents the experimental results, offering a comprehensive quantitative and statistical analysis of the CMs' performance. It explores the model outcomes from multiple perspectives, identifying gaps and opportunities to enhance model effectiveness in challenging contexts. The analysis focuses on comparing the results of pretrained and unpretrained CMs on both UDS and BDS, according to the four scenarios outlined above.

A. QUANTITATIVE PERFORMANCE ANALYSIS

In Tables 8- 11, for the testing phase, the Mean percentage (%) and Standard Deviation of *A*, *P*, *R*, *F1* and *MCC* for all scenarios after 5 runs are reported. The best values in these tables are highlighted in bold. The corresponding bar charts are displayed in Figures 3-4. Additional details are provided in the supplementary material file, which includes the *cms* and the AUC-ROCs of each CM obtained from a randomly selected run among the 5 considered runs during the testing phase.

We underline our decision to share all the aforementioned types of experimental data, as they aid in understanding how effectively the model differentiates among various input classes. Indeed, additional metrics are crucial for evaluation, particularly in unbalanced scenarios. At the same time, Accuracy - defined as the proportion of correct predictions relative to the total number of predictions made and often used alone in many studies - may result to be not sufficient from a quantitative point of view (see also Subsection III-B). The confusion matrix provides detailed insights into both correct and incorrect classifications. The AUC-ROC metric comprehensively assesses the discriminative capabilities of CMS across all output classes.

As noted in Sections II-A and II-D, the employed evaluation metrics may yield values that are lower than those produced for existing models adopting a reliable number of input images and preprocessing. Furthermore, since our focus is on examining the effects of data balance and the impact of TL techniques, these values should be taken as they are and cannot be considered isolated from their context.

TABLE 8. Mean percentage (%) and Standard Deviation of Accuracy *A*, Precision (*P*), Recall (*R*), F1 Score (*F1*) and Matthews Correlation Coefficient (*MCC*) for Scenario 1 (Pretrained CMs on UDS-Test) after 5 runs.

	A	*P*	*R*	*F1*	*MCC*
DenseNet121					
V0	79,47 ± 2,41	66,39± 3,93	62,99±2,30	63,52±1,99	50,21± 1,48
V1	**80,80 ± 0,65**	**67,80±2,26**	**67,24 ±0,74**	**66,76±0,83**	54,63±1,16
V2	77,40 ± 1,19	45,66 ±3,50	50,01±1,49	47,39±1,34	40,01±2,64
InceptionV3					
V0	78,80± 1,17	64,08± 2,81	56,20±2,33	58,55± 2,25	44,34± 1,93
V1	78,87± 1,41	65,10± 5,09	55,96±2,59	57,88± 1,96	44,60± 2,93
V2	75,40±1,28	35,38± 9,78	41,59 ± 7,56	38,12± 8,92	20,61± 18,83
MobileNet					
V0	77,00 ± 2,37	63,55±1,77	62,56± 2,17	62,20±0,87	46,35±1,96
V1	75,74±6,00	66,33± 2,00	65,36±1,37	61,99±5,33	**57,85±18,32**
V2	77,47± 1,02	45,97±5,87	48,47± 1,85	46,18±1,62	38,48±4,05
VGG16					
V0	76,93± 1,23	60,77± 2,31	53,66±3,15	55,67±1,76	39,61±1,74
V1	76,87 ± 1,52	61,16 ±3,00	55,49± 4,05	55,99±2,75	41,54±2,66
V2	74,00 ± 0,01	24,67 ±0,01	33,33± 0,01	28,35±0,01	0,01± 0,01

TABLE 9. Mean percentage (%) and Standard Deviation of Accuracy *A*, Precision (*P*), Recall (*R*), F1 Score (*F1*) and Matthews Correlation Coefficient (*MCC*) for Scenario 2 (Pretrained CMs on BDS-Test) after 5 runs.

	A	*P*	*R*	*F1*	*MCC*
DenseNet121					
V0	69,40 ± 2,53	**72,29± 0,85**	69,40±2,53	68,98±3,14	55,64± 2,56
V1	70,00 ± 0,91	71,25±1,15	70,00 ±0,91	70,12±1,21	55,45±1,26
V2	60,20± 5,28	61,57 ±6,96	60,20±5,28	58,41±8,06	42,28±6,22
InceptionV3					
V0	65,13± 3,48	66,94± 0,84	65,13±3,48	64,17± 4,13	49,12± 3,58
V1	67,20± 2,02	67,75± 1,84	67,20±2,02	67,29± 2,04	50,94± 2,96
V2	59,20±4,89	62,49±6,43	59,20 ± 4,89	57,01± 6,95	41,44± 6,57
MobileNet					
V0	69,07 ± 1,23	69,70±1,52	69,07± 1,23	68,91±1,60	53,98±1,67
V1	**71,60±1,38**	72,08± 2,09	**71,60±1,38**	**71,62±1,38**	**57,59±2,38**
V2	66,67± 2,84	69,90±2,12	66,67± 2,84	66,97±2,83	51,13±3,84
VGG16					
V0	66,20± 1,98	68,10± 1,61	66,20±1,98	65,90±2,65	50,16±1,71
V1	67,53 ± 1,07	68,05 ±1,14	67,53± 1,07	67,44±1,13	51,56±1,57
V2	60,00 ± 7,23	60,80 ±8,85	60,00± 7,23	58,04±10,26	42,31± 8,45

TABLE 10. Mean percentage (%) and Standard Deviation of Accuracy *A*, Precision (*P*), Recall (*R*), F1 Score (*F1*) and Matthews Correlation Coefficient (*MCC*) for Scenario 3 (Unpretrained CMs on UDS-Test) after 5 runs.

	A	*P*	*R*	*F1*	*MCC*
DenseNet121					
V0	66,93 ± 13,47	48,66± 13,64	46,10±12,13	43,64±14,08	22,18± 20,04
V1	73,27 ± 5,33	59,43±4,94	57,35 ±13,37	53,51±14,20	37,06±16,22
V2	70,67± 7,06	50,31 ±12,94	44,12±5,72	39,92±2,62	26,83±8,20
InceptionV3					
V0	**79,80± 2,77**	**67,61± 5,96**	59,97±3,58	62,24± 4,10	48,42± 6,40
V1	79,00± 2,44	64,38± 4,38	61,78±4,19	62,07± 4,53	48,31± 6,19
V2	76,07±1,23	47,84±10,20	44,67 ± 9,82	43,02± 9,21	30,04± 11,97
MobileNet					
V0	78,80 ± 2,14	65,56±3,33	**65,29± 3,69**	**64,84±2,74**	**50,10±3,88**
V1	78,07±0,72	65,11± 3,43	58,53±3,00	59,92±2,23	43,71±1,80
V2	75,93± 1,80	37,00±11,34	43,58± 9,72	39,57±10,28	23,41±21,57
VGG16					
V0	76,93± 1,69	63,77± 5,82	53,46±10,67	54,14±11,10	37,37±12,91
V1	76,80 ± 1,45	65,17 ±7,57	51,88± 6,18	53,99±5,63	37,24±5,00
V2	73,20 ± 1,79	27,84 ±7,10	37,10± 8,42	31,34±6,69	7,44± 16,65

TABLE 11. Mean percentage (%) and Standard Deviation of Accuracy *A*, Precision (*P*), Recall (*R*), F1 Score (*F1*) and Matthews Correlation Coefficient (*MCC*) for Scenario 4 (Unpretrained CMs on BDS-Test) after 5 runs.

	A	*P*	*R*	*F1*	*MCC*
DenseNet121					
V0	60,13 ± 7,35	64,41± 3,37	60,13±7,35	58,60±8,76	42,20± 9,79
V1	59,20 ± 8,69	65,05±3,51	59,20±8,69	57,06±11,98	40,50±12,46
V2	47,07± 9,24	45,54 ±12,04	47,07±9,24	39,24±13,77	25,46±13,04
InceptionV3					
V0	69,60± 3,77	**71,30± 2,47**	69,60±3,77	69,50± 4,02	55,17± 4,88
V1	**70,00± 3,19**	70,97± 3,38	**70,00±3,19**	69,60± 3,36	**55,70± 4,79**
V2	55,53±16,01	52,79±25,52	55,53 ± 16,01	51,81± 22,36	34,18± 24,25
MobileNet					
V0	66,87 ± 1,56	69,28±1,29	66,87± 1,56	66,82±1,49	51,21±1,89
V1	65,40±5,32	69,05± 5,71	65,40±5,32	64,90±5,88	49,80±7,86
V2	44,33± 10,77	31,53±20,36	44,33± 10,77	34,06±17,36	19,16±17,84
VGG16					
V0	69,60± 1,28	71,13± 1,24	69,60±1,28	**69,63±1,31**	54,93±1,79
V1	68,33 ± 1,13	69,70 ±0,26	68,33± 1,13	68,32±1,18	53,01±1,30
V2	66,40 ± 2,90	71,49 ±2,56	66,40± 2,90	65,66±3,75	51,75± 3,41

The analysis of the experimental data yields the following results.

⇒ **Scenario 1 - Pretrained CMs on UDS-Test** - (cfr. the Table 8, the bar charts in Figure 3 Top, the corresponding *cms* and AUC-ROCs in the supplementary material file). DenseNet121 V1 dominates in this scenario, achieving the highest values for most evaluation metrics: Accuracy ∼ 80%, Precision ∼ 67%, Recall ∼ 67%, and F1 Score ∼ 66%. For this CM, the examination of the corresponding confusion matrices shows that there are no classes the CM struggles to predict. Moreover, this CM also exhibits low variability, indicating consistent performance.
The V2 version of all CMs in this scenario is ineffective since they have F1 Score dropping below ∼ 50%, and the metrics show even lower values than those of the corresponding versions V0 and V1. Moreover, some classes are not classified.

✎ **Scenario 2 - Pretrained CMs on BDS-Test** - (cfr. the Table 9, the bar charts in Figure 4 Top, the corresponding *cms* and AUC-ROCs in the supplementary material file). The CM with the better performance in classifying all classes is MobileNet V1. This model achieves the highest scores in four evaluation metrics: Accuracy ∼ 71%, Recall ∼ 71%, F1 Score ∼ 71%, MCC ∼ 57%. Furthermore, an analysis of the confusion matrix reveals that this model has no difficulty predicting any particular class. Similar to Scenario 1, V2 versions perform poorly. DenseNet121 V2, Inception V3 V2, and VGG16 V2 show F1 Score values below 60%.

◇ **Scenario 3 - Unpretrained CMs on UDS-Test** - (cfr. the Table 10, the bar charts in Figure 3 Bottom, the corresponding *cms* and AUC-ROCs in the supplementary material file).
The CM with the better performance in classifying all classes is MobileNet V0. This CM has the highest values of three evaluation metrics with Recall ∼ 65%, F1 Score ∼ 64%, and MCC ∼ 50%. Moreover, for this CM, the corresponding confusion matrix examination shows that there are no classes the CM struggles to predict. InceptionV3 follows in performance since it has the highest values of two evaluation metrics, particularly Accuracy ∼ 79%, Precision ∼ 67%. Similar to scenarios 1 and 2, the V2 version of all CMs performs poorly, often failing to classify some classes entirely (e.g., F1 Score values below 40% for most V2 models).

➢ **Scenario 4 - Unpretrained CMs on BDS-Test** - (cfr. the Table 11, the bar charts 4 bottom, the corresponding *cms* and AUC-ROCs in the supplementary material file). The CM with the better performance in classifying all classes is InceptionV3 V1, achieving the highest values in three metrics with Accuracy ∼ 70%, Recall ∼ 70%, MCC ∼ 55%. Moreover, for this CM, the examination of the corresponding confusion matrix shows that there are no classes that the CM struggles to predict. Similar to other scenarios, V2 versions fail to deliver competitive performance. For example, MobileNet V2 has F1 Score values below 40%, and DenseNet121 V2 performs similarly poorly. Note that some V2 versions exhibit the highest standard deviations, suggesting more variability and inconsistency than other scenarios.

B. STATISTICAL PERFORMANCE ANALYSIS

To assess the performance differences among the CMs, we conduct a statistical analysis [65], which allows a deeper understanding of the quantitative results.

Based on the insights about Accuracy discussed in Subsection III-A from a quantitative perspective,

FIGURE 3. Bar charts for pretrained and unpretrained DenseNet12, InceptionV3, MobileNet and VGG16 in V0, V1, V2 versions on UDS-Test (scenario 1 and 3) .

we highlight that, from a statistical viewpoint, this performance analysis should prioritize the F1 Score over Accuracy. The main compelling reasons for this preference are the following.

1) Accuracy provides a general indication of how well a CM performs in classifying benign and malignant lesions. This metric is especially relevant for balanced datasets, where the counts of benign and malignant cases are comparable. However, its significance diminishes in unbalanced datasets, where one class overshadows the other, potentially distorting the outcomes. In such situations, relying solely on Accuracy can be misleading, as it does not capture the proficiency of the CM in identifying minority classes. Furthermore, it fails to account for the differing clinical implications of false positives and false negatives, which can vary significantly in importance.
2) F1 Score is the harmonic mean of Precision and Recall, providing a balance between these metrics and highlighting configurations where both are important. It is particularly valuable for identifying false positives and analyzing errors from missed correct classifications. By accounting for false positives and false negatives, the F1 Score comprehensively evaluates the overall CMs' performance across the analyzed scenarios. This holistic and balanced perspective makes it an essential

FIGURE 4. Bar charts for pretrained and unpretrained DenseNet12, InceptionV3, MobileNet and VGG16 in V0, V1, V2 versions on BDS-Test (scenario 2 and 4) .

metric for assessing prediction effectiveness. In particular, the F1 Score is crucial in skin lesion classification, where balancing Precision and Recall is necessary since it is important to avoid overdiagnosis (Precision) and detect all malignant lesions (Recall). Indeed, a high F1 Score ensures that the model maintains a balance between detecting all malignant lesions and minimizing incorrect diagnoses. Moreover, it is particularly relevant in unbalanced datasets, such as those common in skin lesion analysis, where prioritizing Recall should not come at the expense of Precision (or vice versa).

Thus, we generate F1 Score boxplots for each scenario, shown in Figure 5, highlighting the average percentage of F1 Score for each CM and its respective version V0, V1, V2.

Since we are considering 12 different CMs and a potential non-normal distribution of the data, to determine whether the medians are significantly different, we apply the non-parametric Kruskal-Wallis test [66]. The resulting p-value is compared against a significance level of $\alpha = 0.05$. A p-value less than α indicates the presence of significant differences between the medians of at least one pair of CMs.

In each scenario, we found a statistically significant difference between at least one pair of the CMs, as evidenced by a

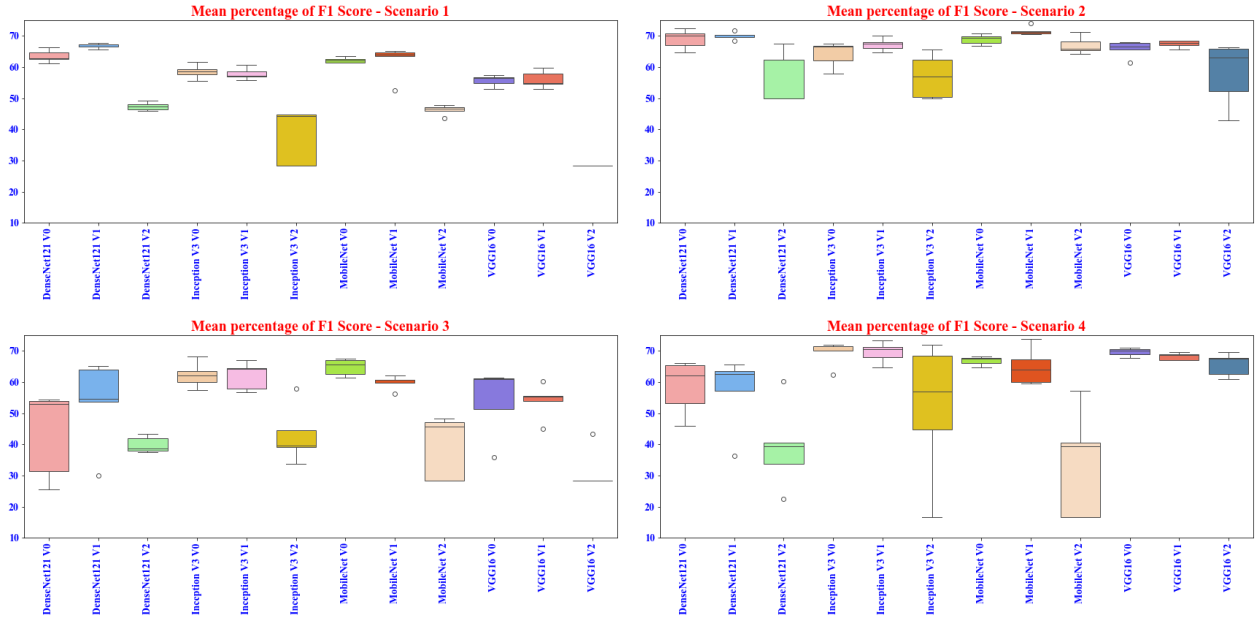

FIGURE 5. Boxplots of mean percentage of F1 Score for all scenarios.

p-value from the Kruskal-Wallis test lower than α. However, this test does not specify which CMs differ from each other. To pinpoint these significant pairs, we conduct a post-hoc Dunn test [67], implementing the Benjamini-Hochberg correction to control for type I [68]. This correction is deemed particularly suitable given the nature and distribution of the data, especially when comparing the accuracy and F1 Score of the 12 CMs.

Following the post-hoc test, we identified the pairs of the CMs with significant differences. Figures 6 present the F1 Score heatmaps (also simply as F1 Heatmaps). In these heatmaps, the CMs are labeled from "a" to "n" as indicated in the legend. Warm colors between two CMs denote high correlation, meaning that the CMs produce very similar results. Conversely, cool colors (shades of blues) indicate a low correlation, reflecting significant differences in their outcomes.

Comparing the obtained boxplots and heatmaps, the following considerations emerge.

⇒ **Scenario 1 - Pretrained CMs on UDS-Test**
DenseNet121 V1 leads in achieving a balanced trade-off between Precision and Recall, with medians close to 70%, and narrow interquartile range (IQR) indicating low variability. In contrast, InceptionV3 V2 performs the worst, with a notably low median ($\sim$ 40%) and high variability.
Analysis of the heatmap reveals strong correlations within the same model families. Indeed, models like VGG16 or MobileNet exhibit strong correlations between their respective versions, e.g., between VGG16 V1 and V2 and MobileNet V0 and V1, indicating similar performance patterns. Different model families, such as InceptionV3 and MobileNet, show low correlations, e.g., InceptionV3 V2 and MobileNet V1. There are some exceptions, e.g., MobileNet V0 and V1 and DenseNet121 V0 display higher-than-expected correlations, suggesting similar behavior between these CMs.

✎ **Scenario 2 - Pretrained CMs on BDS-Test**
MobileNet V1 stands out as the best-performing model due to its combination of high mean F1 Score ($\sim$ 70%), comparable to DenseNet121 V0 and V1 where the version V0 delivers a similar F1 Score but with slightly higher variance, reducing its overall stability.
The heatmap analysis reveals that CMs within the same family tend to be uncorrelated strongly. For example, MobileNet V0 and MobileNet V1 demonstrate a very low correlation. Similarly, VGG16 V1 and VGG16 V2 also exhibit a weak correlation. In contrast, CMs from different families, such as DenseNet121 and MobileNet, generally show higher correlations; e.g., MobileNet V0 shows a high correlation.

◇ **Scenario 3 - Unpretrained CMs on UDS-Test**
MobileNet V0 and InceptionV3 V1 are the top performers in Scenario 3, with MobileNet V0 demonstrating consistent performance across all classes. In contrast, the poor performance of the V2 models is clearly visible. This analysis aligns closely with the visual representation.
The heatmap reveals strong correlations for the family DenseNet121: DenseNet121 V0 with InceptionV3 V2, DenseNet121 V1 with VGG16 V0, and DenseNet121 V2 with MobileNet V2. In contrast, correlations between CMs from different families, such as DenseNet121 and MobileNet or VGG16

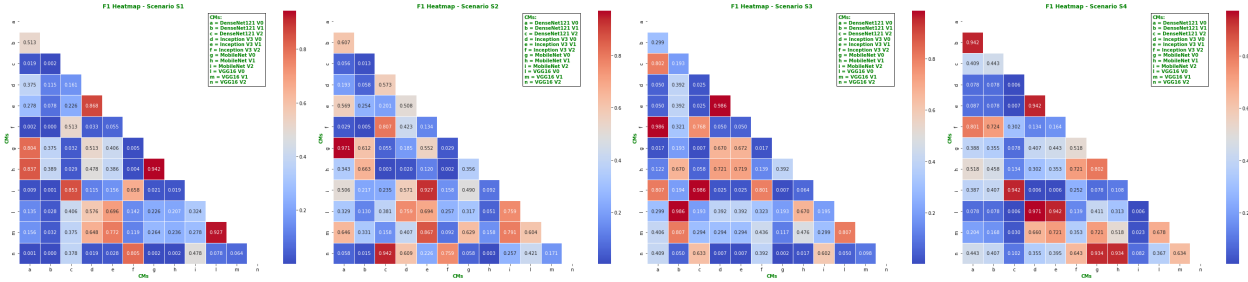

FIGURE 6. F1 heatmaps for all scenarios.

and MobileNet, tend to have lower correlations. For instance, MobileNet V0 and V1 display a moderate correlation with VGG16 V2.

➢ **(Scenario 4 - Unpretrained CMs on BDS-Test**
InceptionV3 V0 and V1, followed by VGG16 V0, stand out as reliable and high-performing models in this scenario.

The heatmap analysis reveals strong intra-family correlations, with versions of the same CM family-such as DenseNet121 and InceptionV3 - showing high correlation values, indicative of similar performance. In contrast, cross-family correlations are present. For instance, the VGG16 family shows a strong correlation with InceptionV3 and MobileNet families.

The statistical analysis results align closely with those of the quantitative analysis. The presence of only a few isolated outliers across all boxplots further highlights the robustness and consistency of the performance evaluation for the CMs. Additionally, the poor performance of the V2 versions of the CMs across all scenarios, as identified in the quantitative analysis, is corroborated by the statistical evidence. Furthermore, the statistical analysis reinforces the insights derived from the quantitative analysis when examining each scenarios individually.

IV. DISCUSSION

In this section, we interpret the experimental data of each scenario given in Section III, trying to offer a comprehensive framework based on the quantitative and statistical assessments of the performance of each CM, seeking to glean useful insights for CM design and employment.

To this purpose, we perform a comparative analysis focusing on two key computational aspects: the first pertains to data balancement, namely ''UDS-Test versus BDS-Test'', see subsection IV-A, and the second concerns TL strategies, namely ''Pretrained versus Unpretrained'', see subsection IV-B. Next, we draw comparisons across all four scenarios to highlight overarching trends and differences; see subsections IV-C.

A. UDS-TEST VERSUS BDS-TEST

To assess the impact of data balancing, we analyze the experimental results comparing Scenario 1 versus Scenario 2 (Scenario 3 versus Scenario 4) for the pretrained (unpretrained) CMs. Both quantitative and statistical analyses reveal a consistent decline in performance on UDS compared to BDS, regardless of the employed training approach. Indeed, Recall, which measures the correctness of classification for the minority class, and F1 Score, which simultaneously evaluates correctness and coverage for the minority class, improve significantly for all CMs after balancing, regardless of whether they are pretrained or unpretrained. This observation is substantiated by the comparison of R and F1 Score values in Table 8 and Table 9 for pretrained CMs, as well as in Table 10 and Table 11 for unpretrained CMs. Additionally, an analysis of the confusion matrices (see supplementary material) confirms that high Accuracy values may sometimes coincide with incorrect classifications on UDS-Test.

It is worth emphasizing that this improvement of R and F1 after balancing is particularly crucial in the context of skin lesion classification, where the minority class often holds the greatest importance: missing a cancer diagnosis (false negative) can lead to a delayed treatment, while a false diagnosis may lead to unnecessary interventions. Consequently, these experimental results confirm the hypothesis that data balancing should be prioritized whenever feasible, regardless of the training type (pretrained or unpretrained).

From a methodological perspective, we remark that these experimental data underscore the importance of evaluating CMs using additional metrics beyond Accuracy in the quantitative analysis and highlight the relevance of employing the F1 Score in the statistical analysis, as previously discussed (see Sections III-A and III-B).

B. PRETRAINED VERSUS UNPRETRAINED

To evaluate the impact of pretraining, we compare the experimental results of Scenario 1 versus Scenario 3 (Scenario 2 versus Scenario 4) for the unbalanced (balanced) dataset. Quantitative and statistical analyses reveal a general decline in performance for unpretrained CMs compared to pretrained ones. Indeed, with few exceptions, all metrics improve significantly for pretrained CMs, regardless of the dataset balancing. This metrics improvement can be argued by comparing the metrics in Table 8 and Table 10 for UDS-Test and in Table 9 and Table 11 for BDS-Test. These comparisons suggest that the learned features during pretraining are well-suited for skin lesion classification, enabling the CMs to generalize effectively when a TL technique is employed.

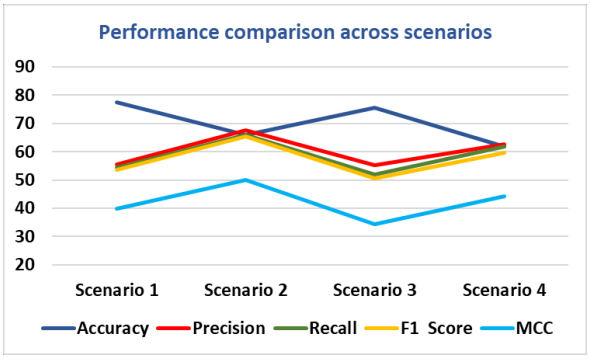

FIGURE 7. Trend of average performance in terms of Accuracy, F1 Score, Precision, Recall, and MCC across scenarios. Each line represents the average value of one metric.

Moreover, comparing all metric values, accuracy excluded, reported in Table 8 and Table 9 respectively relative to Scenario 1 versus Scenario 2, e. g. metrics values for pretrained CMs on UDS-Test and BDS-Test respectively, it can be observed that pretraining is more effective on balanced datasets, albeit with slightly lower accuracy. This highlights that utilizing TL techniques can improve the performance of a CM, particularly when applied to balanced datasets.

C. SCENARIOS COMPARISON

Figure 7 illustrates the trend described in subsections IV-A and IV-B. Accuracy (blue line) remains the highest-performing metric across all scenarios, starting around 80% in Scenario 1, dipping slightly in Scenario 2, peaking in Scenario 3, and then dropping in Scenario 4. Precision (red line) peaks slightly in Scenario 2 at around 67%, showing improved performance in that scenario and a slight dip in Scenario 3 before recovering in Scenario 4. Recall (green line) and F1 Score (orange line) show similar trends, both declining in Scenario 3 and increasing slightly in Scenario 4, mirroring Precision's trend. MCC (light blue line) starts relatively lower compared to other metrics in Scenario 1, and it follows the trend of Precision, Recall, and F1 Score, with a noticeable dip in Scenario 3 and partial recovery in Scenario 4.

A similar trend is observable for P, R, F1, and MCC, suggesting that Scenario 2 delivers superior performance compared to the other scenarios, whereas Scenario 3 exhibits the weakest performance across these metrics. The opposite trend exhibited by Accuracy confirms that this metric alone cannot reliably assess the CMs' performance despite being frequently used for this purpose. Scenario 1 performs better than Scenario 4 and Scenario 3, with Scenario 3 being the worst-performing.

Overall, based on the experimental data, we infer the following.

- ✓ **Favorable Scenarios** - Scenario 2 represents ideal conditions where CMs deliver better performance. Conversely, Scenario 3 underscores the need for more robust strategies to manage increased complexity, significant data unbalancing, and the lack of pretraining.
- ✓ **Reference CMs** - The quantitative and statistical analyses allowed us to identify the most reliable CMs demonstrating high and stable values in specific scenarios. The experimental results further suggest that implementing an optimized version of baseline models can effectively handle complex scenarios more efficiently.
- ✓ **Practical Implications** - Using multiple metrics, such as P, R, F1, and MCC, is crucial for identifying weaknesses that may be overlooked when relying solely on Accuracy. This is particularly important in skin lesion classification, where minimizing false negatives in malignant cases and ensuring accurate diagnoses are critical. Consequently, we recommend prioritizing P, R, and F1 for evaluation. Furthermore, our analysis of the experimental data demonstrates that combining quantitative and statistical analyses with class-specific assessments, such as confusion matrices and AUC-ROCs, is essential for identifying limitations and ensuring a comprehensive evaluation of dermatopathology categorization performance.
- ✓ **Improvement Strategies** - The experimental data indicate that balancing an unbalanced dataset is beneficial for improving performance since, both in pretrained and unpretrained scenarios, better performance is achieved when the balanced dataset is used. Similarly, the experimental data indicate that pretraining of the (un)balanced dataset is advantageous for improving performance as, both on UDS and BDS, better performance is observed when a TL strategy is employed.
  To address the CMs' limitations, in complex and unbalanced cases (e.g., Scenario 3), we recommend applying data balancing techniques to ensure equal class representation. In contrast, in complex and unpretrained CMs (e.g., Scenario 4), we suggest employing TL methods to enhance the training and the CMs' performance.

V. CONCLUSION

In this paper, we address the open issue related to the distribution/quantity of images in CNN-based skin lesion classification, which is commonly overlooked in most papers. A few papers have proposed using balanced datasets or TL techniques, but not their combination, to overcome the hurdles related to this challenge.

We investigate this issue by proposing twelve computational models with increasing structural complexity based on four prominent CNN models and providing a comprehensive quantitative evaluation of their effectiveness. The computational models are tested on a balanced and unbalanced dataset extracted from the ISIC 2018 dataset. Each computational model undergoes training using the same hyperparameters and loss function, either through a TL technique or a standard procedure, originating four different scenarios that consider various combinations of classification modes.

The performed comprehensive quantitative and statistical evaluation has highlighted the advantages and computational limitations of considered CMs in different scenarios. This analysis provides crucial insights, laying a basis for further investigations into the effective design and use of CNNs for skin lesion classification.

One of the main limitations of our study, which focuses on the specific application of skin lesion classification, is that the validity of our analytical findings is restricted to this domain. Whether our conclusions can be generalized to other medical imaging tasks or broader computer vision applications remains uncertain. Future research could investigate the applicability of our approach to different classification problems beyond dermatology.

Furthermore, in scenarios where pretrained models are used, our analysis is constrained by the fact that fine-tuning has been performed with ImageNet weights. Although leveraging ImageNet pretrained models is a common practice, the significant differences between ImageNet images and dermatological images suggest that using weights obtained from large dermoscopic datasets could potentially improve the evaluation process. However, this approach is currently not feasible due to the unavailability of such pretrained weights.

APPENDIX
SUPPLEMENTARY MATERIAL FILE

In the supplementary material file, as additional figures, for all scenarios are available:

- the accuracy and loss change graphs of each CM for the training and validation steps;
- the *cms* and AUC-ROCs of each CM for the test step.

ACKNOWLEDGMENT

The authors express their gratitude to the anonymous reviewers for their valuable comments, which significantly contributed to enhancing the quality of the article. They also thank Dr. Claudia Angelini, Institute for the Applications of Calculus of the Italian National Research Council, for her insightful guidance on the statistical analysis of the experimental data.

REFERENCES

[1] S. Maurya, S. Tiwari, M. C. Mothukuri, C. M. Tangeda, R. N. S. Nandigam, and D. C. Addagiri, "A review on recent developments in cancer detection using machine learning and deep learning models," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104398.

[2] R. Wu, K. Qin, Y. Fang, Y. Xu, H. Zhang, W. Li, X. Luo, Z. Han, S. Liu, and Q. Li, "Application of the convolution neural network in determining the depth of invasion of gastrointestinal cancer: A systematic review and meta-analysis," *J. Gastrointestinal Surg.*, vol. 28, no. 4, pp. 538–547, Apr. 2024.

[3] J. Chaki and A. Ucar, *Current Applications of Deep Learning in Cancer Diagnostics*. Boca Raton, FL, USA: CRC Press, 2023.

[4] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018.

[5] S. Rionero and I. Torcicollo, "On an ILL-posed problem in nonlinear heat conduction," *Transp. Theory Stat. Phys.*, vol. 29, nos. 1–2, pp. 173–186, Jan. 2000.

[6] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100134.

[7] A. Kumar, A. Vishwakarma, V. Bajaj, and S. Mishra, "Novel mixed domain hand-crafted features for skin disease recognition using multi-headed CNN," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[8] F. Ahmed, A. Fatima, M. Mamoon, and S. Khan, "Identification of the diabetic retinopathy using ResNet-18," in *Proc. 2nd Int. Conf. Cyber Resilience (ICCR)*, Feb. 2024, pp. 1–6.

[9] M. K. Hasan, M. A. Ahamad, C. H. Yap, and G. Yang, "A survey, review, and future trends of skin lesion segmentation and classification," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106624.

[10] S. M. Sivalingam, P. Patil, and S. Ramachandran, "Melanoma prediction from skin lesions using convolutional neural network against the support vector machine learning algorithm to achieve maximum accuracy and minimum loss," in *Research Advances in Intelligent Computing*. Boca Raton, FL, USA: CRC Press, 2023.

[11] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. C. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, and H. Kittler, "Human–computer collaboration for skin cancer recognition," *Nature Med.*, vol. 26, no. 8, pp. 1229–1234, Jun. 2020.

[12] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Jan. 2015, pp. 1–14.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jun. 2015, pp. 1–9.

[15] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[17] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Jan. 2019, pp. 6105–6114.

[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)*, Los Alamitos, CA, USA, Jul. 2017, pp. 2261–2269.

[19] G. Ramella and G. S. D. Baja, "Color quantization via spatial resolution reduction," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, vol. 1. SciTePress, 2013, pp. 78–83.

[20] V. Bruni, G. Ramella, and D. Vitulano, "Automatic perceptual color quantization of dermoscopic images," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.* SciTePress, 2015, pp. 323–330.

[21] M. E. Celebi, "Forty years of color quantization: A modern, algorithmic survey," *Artif. Intell. Rev.*, vol. 56, no. 12, pp. 13953–14034, Dec. 2023.

[22] G. Ramella and G. S. di Baja, "A new method for color quantization," in *Proc. 12th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2016, pp. 1–6.

[23] G. Ramella and G. S. D. Baja, "Color histogram-based image segmentation," in *Computer Analysis of Images and Patterns*, P. Real, D. Diaz-Pernil, vol. 6854, H. Molina-Abril, A. Berciano, and W. Kropatsch, Eds., Berlin, Germany: Springer, 2011, pp. 76–83.

[24] G. Ramella and G. S. D. Baja, "Image segmentation based on representative colors detection and region merging," in *Pattern Recognition* (Lecture Notes in Computer Science), vol. 7914, J. A. Carrasco-Ochoa, J. F. Martinez-Trinidad, J. S. Rodriguez, and G. Sanniti di Baja, Eds., Berlin, Germany: Springer, 2013, pp. 175–184.

[25] Z. Mirikharaji, K. Abhishek, A. Bissoto, C. Barata, S. Avila, E. Valle, M. E. Celebi, and G. Hamarneh, "A survey on deep learning for skin lesion segmentation," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102863.

[26] M. Frackiewicz, G. Szolc, and H. Palus, "An improved SPSIM index for image quality assessment," *Symmetry*, vol. 13, no. 3, p. 518, Mar. 2021.

[27] M. Frackiewicz and H. Palus, "K-means color image quantization with deterministic initialization: New image quality metrics," in *Image Analysis and Recognition* (Lecture Notes in Computer Science), vol. 10882, A. Campilho, F. Karray, and T. H. Romeny, Eds., Springer, 2018, pp. 56–61.

[28] G. Ramella, "Evaluation of quality measures for color quantization," *Multimedia Tools Appl.*, vol. 80, pp. 32975–33009, Sep. 2021.
[29] S. W. Pewton, B. Cassidy, C. Kendrick, and M. H. Yap, "Dermoscopic dark corner artifacts removal: Friend or foe?" *Comput. Methods Programs Biomed.*, vol. 244, Feb. 2024, Art. no. 107986.
[30] G. Ramella, "Hair removal combining saliency, shape and color," *Appl. Sci.*, vol. 11, no. 1, p. 447, Jan. 2021.
[31] R. Kaur, H. GholamHosseini, and R. Sinha, "Hairlines removal and low contrast enhancement of melanoma skin images using convolutional neural network with aggregation of contextual information," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103653.
[32] S. Barin and G. E. Güraksin, "An improved hair removal algorithm for dermoscopy images," *Multimedia Tools Appl.*, vol. 83, no. 3, pp. 8931–8953, Jan. 2024.
[33] A. Mikołajczyk, S. Majchrowska, and S. C. Limeros, "The (de)biasing effect of GAN-based augmentation methods on skin lesion images," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 13438, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Cham, Switzerland: Springer, 2022, pp. 437–447.
[34] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, May 2018, pp. 117–122.
[35] C. Arcelli, N. Brancati, M. Frucci, G. Ramella, and G. Sanniti di Baja, "A fully automatic one-scan adaptive zooming algorithm for color images," *Signal Process.*, vol. 91, no. 1, pp. 61–71, Jan. 2011.
[36] D. Occorsio, G. Ramella, and W. Themistoclakis, "Filtered polynomial interpolation for scaling 3D images," *ETNA - Electron. Trans. Numer. Anal.*, vol. 59, pp. 295–318, 2023.
[37] S. Ghosh and A. Garai, "Image downscaling via co-occurrence learning," *J. Vis. Commun. Image Represent.*, vol. 91, Mar. 2023, Art. no. 103766.
[38] D. Occorsio, G. Ramella, and W. Themistoclakis, "An open image resizing framework for remote sensing applications and beyond," *Remote Sens.*, vol. 15, no. 16, Aug. 2023, Art. no. 4039.
[39] H. Li, P. Zeng, C. Bai, W. Wang, Y. Yu, and P. Yu, "PMJAF-net: Pyramidal multi-scale joint attention and adaptive fusion network for explainable skin lesion segmentation," *Comput. Biol. Med.*, vol. 165, Oct. 2023, Art. no. 107454.
[40] 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.
[41] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Comput. Methods Programs Biomed.*, vol. 193, Sep. 2020, Art. no. 105475.
[42] A. Esposito, *The Importance of Data for Training Intelligent Devices*. Berlin, Germany: Springer, 2002, pp. 229–250.
[43] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.
[44] O. Sevli, "A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation," *Neural Comput. Appl.*, vol. 33, no. 18, pp. 12039–12050, Sep. 2021.
[45] S. Jain, U. Singhania, B. Tripathy, E. A. Nasr, M. K. Aboudaif, and A. K. Kamrani, "Deep learning-based transfer learning for classification of skin cancer," *Sensors*, vol. 21, no. 23, p. 8142, Dec. 2021.
[46] S. P. G. Jasil and V. Ulagamuthalvi, "Deep learning architecture using transfer learning for classification of skin lesions," *J. Ambient Intell. Humanized Comput.*, Mar. 2021.
[47] W. Dai, R. Liu, T. Wu, M. Wang, J. Yin, and J. Liu, "Deeply supervised skin lesions diagnosis with stage and branch attention," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 2, pp. 719–729, Feb. 2024.
[48] S. S. Chaturvedi, J. V. Tembhurne, and T. Diwan, "A multi-class skin cancer classification using deep convolutional neural networks," *Multimedia Tools Appl.*, vol. 79, nos. 39–40, pp. 28477–28498, Oct. 2020.
[49] F. Ahmed, S. Abbas, A. Athar, T. Shahzad, W. A. Khan, M. Alharbi, M. A. Khan, and A. Ahmed, "Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence," *Sci. Rep.*, vol. 14, no. 1, p. 6173, Mar. 2024.
[50] F. Ahmed, W. A. Khan, M. Iqbal, A. R. A. Abazeed, H. Alrababah, and M. F. Khan, "Rock-Paper-Scissors image classification using transfer learning," in *Proc. Int. Conf. Bus. Analytics Technol. Secur. (ICBATS)*, Mar. 2023, pp. 1–6.
[51] S. M. Jaisakthi, P. Mirunalini, C. Aravindan, and R. Appavu, "Classification of skin cancer from dermoscopic images using deep neural network architectures," *Multimedia Tools Appl.*, vol. 82, no. 10, pp. 15763–15778, Apr. 2023.
[52] A. Gong, X. Yao, and W. Lin, "Dermoscopy image classification based on StyleGANs and decision fusion," *IEEE Access*, vol. 8, pp. 70640–70650, 2020.
[53] A. Gong, X. Yao, and W. Lin, "Classification for dermoscopy images using convolutional neural networks based on the ensemble of individual advantage and group decision," *IEEE Access*, vol. 8, pp. 155337–155351, 2020.
[54] S. Abbas, F. Ahmed, W. A. Khan, M. Ahmad, M. A. Khan, and T. M. Ghazal, "Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence," *Sci. Rep.*, vol. 15, no. 1, p. 1746, Jan. 2025.
[55] G. Arora, A. K. Dubey, Z. A. Jaffery, and A. Rocha, "A comparative study of fourteen deep learning networks for multi skin lesion classification (MSLC) on unbalanced data," *Neural Comput. Appl.*, vol. 35, no. 11, pp. 7989–8015, Apr. 2023.
[56] A. Kwasigroch, A. Mikolajczyk, and M. Grochowski, "Deep neural networks approach to skin lesions classification—A comparative analysis," in *Proc. 22nd Int. Conf. Methods Models Autom. Robot. (MMAR)*, Aug. 2017, pp. 1069–1074.
[57] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Aug. 2018.
[58] 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.
[59] *The International Skin Imaging Collaboration*. Accessed: Mar. 19, 2024. [Online]. Available: https://www.isic-archive.com/
[60] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.
[61] Y LeCun, Y Bengio, and G Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[64] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[65] A. Field, *Discovering Statistics Using SPSS*, 4th ed., London, U.K.: SAGE, 2013.
[66] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, p. 583, Dec. 1952.
[67] O. J. Dunn, "Multiple comparisons among means," *J. Amer. Stat. Assoc.*, vol. 56, no. 293, pp. 52–64, Mar. 1961.
[68] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 57, no. 1, pp. 289–300, Jan. 1995.

GIULIANA RAMELLA (Member, IEEE) received the Ph.D. degree in physics, specialized in cybernetics from the University of Napoli Federico II, Italy. She is currently a Researcher with the Institute for the Applications of Calculus "M. Picone" (IAC), Italian National Research Council (CNR). Her research is centered on human perception theory and oriented toward real-world applications in fields ranging from biomedicine, cultural heritage, and the environment to security.

LUCA SERINO received the Ph.D. degree in computer science from the University of Salerno, Italy. He is currently a Researcher with the Institute of High-Performance Computing and Networking (ICAR), Italian National Research Council (CNR). His area of research is mainly in image processing and pattern recognition, focusing on 2-D and 3-D shape representation and analysis and biometrics. His current main interests include deep learning and reinforcement learning.