Skin Lesion Classification by Ensembles of Deep Convolutional Networks and Regularly Spaced Shifting

KARL THURNHOFER-HEMSI [1,2], **EZEQUIEL LÓPEZ-RUBIO** [1,2], **ENRIQUE DOMÍNGUEZ** [1,2], **AND DAVID A. ELIZONDO** [3], (Senior Member, IEEE)

[1]Department of Computer Languages and Computer Science, Universidad de Málaga, 29071 Málaga, Spain
[2]Biomedic Research Institute of Málaga (IBIMA), 29010 Málaga, Spain
[3]School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, U.K.

Corresponding author: Karl Thurnhofer-Hemsi (karlkhader@lcc.uma.es)

ABSTRACT Skin lesions are caused due to multiple factors, like allergies, infections, exposition to the sun, etc. These skin diseases have become a challenge in medical diagnosis due to visual similarities, where image classification is an essential task to achieve an adequate diagnostic of different lesions. Melanoma is one of the best-known types of skin lesions due to the vast majority of skin cancer deaths. In this work, we propose an ensemble of improved convolutional neural networks combined with a test-time regularly spaced shifting technique for skin lesion classification. The shifting technique builds several versions of the test input image, which are shifted by displacement vectors that lie on a regular lattice in the plane of possible shifts. These shifted versions of the test image are subsequently passed on to each of the classifiers of an ensemble. Finally, all the outputs from the classifiers are combined to yield the final result. Experiment results show a significant improvement on the well-known HAM10000 dataset in terms of accuracy and F-score. In particular, it is demonstrated that our combination of ensembles with test-time regularly spaced shifting yields better performance than any of the two methods when applied alone.

INDEX TERMS Image processing, deep learning, classification, skin lesion.

I. INTRODUCTION

Skin lesions are one of the most common types of malignancies, and a considerable increase of cases is expected in the following years due to pandemics (lockdown). Apart from other factors like allergies or infections, exposition to the sun is the leading cause of skin alterations, producing an abnormal multiplication of melanocytes and, consequently, melanomas. Searching for beauty, looking for a tan of their skin with a high exposition to the sun, can have a negative effect on the appearance of skin lesions, especially after a low exposition to the sun due to the lockdown and mask-wearing.

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

Skin cancers initiate from the epidermis, which is the topmost skin layer. Therefore, it is pretty visible, and a computer-aided system might use the images of skin lesions to reveal the preliminary diagnosis without assessing any other relevant information. Melanoma is one the most lethal type of skin cancer in humans, while in the current year, thousands of new cases of melanoma are predicted to be identified and are estimated to die due to melanoma [1], [2]. However, melanoma is highly curable when detected in its earliest stages, and it is more likely than other skin cancer to spread to other parts of the body [3]. There is a high probability of treatment if the skin cancer is detected in the early stages. Nevertheless, it is difficult to analyze whether the skin lesion is malignant or benign and detect skin cancer

during these early stages since the skin lesions look similar to one another. In their initial growth phases, melanoma is very similar to other benign moles in their characteristic, which makes the diagnosis difficult between what is malignant and what is benign even for experienced dermatologists [4].

Numerous computational intelligence techniques, including genetic algorithms, artificial neural networks, support vector machines, ABCDE rule, have been proposed to assess and classify skin lesions either malignant or benign. Most of the automatic classification systems in medical imaging have suffered the lack of data availability, provoking an insufficient generalization of the prediction models. In addition to this, training datasets lack sufficient quality in the sense of homogeneity in the acquisition procedure and non-expected objects present in the image, making it necessary to carry out several preprocessing steps [5] and segment the region of interest [6], [7]. Moreover, another commonly used technique is the extraction of features that are used then to improve the classification rate [8], [9]. The use of specific features extracted from the melanoma images was widely used to develop classification models [10]–[12], although the main inconvenience of these approaches is the requirement of specific expertise to extract the adequate features and the high quantity of time necessary to select the most appropriate. Moreover, image preprocessing may introduce errors or loss of essential information that can affect the final classification rate. A simple example is the low accuracy obtained when a poor skin lesion segmentation is carried out. Until a few years, the classical workflow was the use of these traditional techniques [13], yielding not good enough accuracy. In order to overcome these limitations, deep learning models have recently been developed with success, having the ability to automatically learn the crucial features that can help differentiate among the classes that can be found in digital images.

Deep learning has been applied to resolve very complex classification and segmentation tasks [14], [15] without the use of any image preprocessing method. The architecture of these networks is mainly based on convolutional layers. These layers filter and extract essential features of the images to learn to identify different lesions. For example, Zhou *et al.* [16] used different modality images to learn the features that determine dementia cases. Commonly named Convolutional Neural Networks (CNNs), they have been applied to many areas of interest, showing exceptional performance in image and video processing [17], [18]. Nowadays, CNNs use the power of GPUs to compute a large number of operations in a few seconds, allowing them to process large datasets to create reliable models to be applied to image classification, decision support systems and object recognition and segmentation. With the increase of publicly available datasets, deep networks have shown excellent performance on medical image analysis [19]. Gao *et al.* [20] used neural networks fed with extra privileged information to carry out strain reconstruction in ultrasound elastography. Deep learning models have also been used to detect vessel borders [21], to perceive blood flow from angiographies [22], and for

segmentation of different human organs like vertebrae [23] or liver [24]. Specifically, recent research related to skin lesion classification [25], [26] have been published, although there is still margin for improvement. This research is based on a two-stage process where deep networks are used to segment and extract features and then make the prediction. Moreover, most of them focus on the binary classification problem. Often different types of skin pathologies are grouped into the same class and not classified.

Convolutional Neural Networks (CNNs) is one of the most popular deep learning techniques for image analysis. CNNs were inspired by the animal visual cortex. They are one of the first truly successful deep learning architectures, which have shown outstanding performance in processing images and videos. Nowadays, with the help of GPU-accelerated computing techniques, CNNs have been successfully applied to object recognition (e.g. handwriting, face, behavior...), decision support systems and image classification. Recent research shows that deep networks are powerful tools for medical image analysis [19], [27]. Therefore, they offer great potential for melanoma classification [28], [29]. In particular, CNN ensemble methods have proved particularly successful for this task [30].

In this work, an improved CNN model based on a test-time regularly spaced shifting technique is proposed. Other shifting techniques like random shifting have been successfully applied to increase the resolution of magnetic resonance images [31]. In this research a shifting technique with a regular displacement for the test input image is proposed. The method is aimed at improving the performance of the CNN in the classification. It should be noted that this proposal is not related to train-time data augmentation by shifting the training images. It is also different from previous approaches that apply a random shift to the input image [32] and from previously considered transformations for test time augmentation [33]. Experimental results show a significant improvement in terms of accuracy and F_1-score when an ensemble of deep networks is combined with the test-time regular shifting technique.

Consequently, this paper has the following contributions:
- a successful application of transfer learning for dermoscopic image classification;
- an implementation of a test-time regularly space shifting method to improve the classification;
- the proposal of an ensemble model of deep networks that takes the benefit of the knowledge learned by several classifiers.

The rest of the paper is organized as follows: Section II sum up the recent works in the field of skin lesion diagnosis. Section III presents the proposed methodology to carry out the classification of skin lesions. Section IV describes the convolutional neural networks used in the ensemble model, as well as the parameter setup. The experimental setting and the discussed results are presented in Section V. Finally, the main conclusions and further works are summarized in Section VI.

II. RELATED WORK

Several approaches have been developed for the classification of skin lesion images during the last years. For that purpose, many datasets were released to motivate researchers to find a proper solution to this task. The most famous challenges are the ISBI/ISIC, created in 2016, which comprises data for classification and segmentation. Specifically, the ISIC2018 contains the HAM10000 dataset as training data. Next, we summarize some of the published works using these datasets, focusing on the performance achieved.

Romero Lopez *et al.* [34] presented a model based on VGG16 to classify images of the ISBI 2016 challenge dataset. This method enhances the conventional CNN performances and achieved 81.33% accuracy and 78.6% sensitivity. In [35] a VGG19 combined with randomized trees is used to achieve a precision and F-score of 83%.

Yap *et al.* [36] employed a pre-trained ResNet 50 architecture using the ISIC 2017 datasets and also utilize data augmentation techniques. In this work, two parallel architectures are used to build a multimodal architecture to train dermoscopic images, and macroscopic images, achieving 72.9% mean average precision. Mahbod *et al.* [37] proposed a hybrid model to classify the skin lesion into seven classes. They combined AlexNet, VGG16, and ResNet18 pre-trained models to extract features that will further used to train the SVM classifier. The results were evaluated on ISIC 2016 and ISIC 2017 datasets using data augmentation. The model achieved an accuracy of 90.69%.

In [38] is presented a comparative analysis of state-of-art CNN models, Inceptionv3, ResNet50, DenseNet201, and Inception-ResNet-v2, trained by using transfer learning in order to classify the HAM10000 dataset. The overall accuracy of Inception-ResNet-v2 is 81.62%, and that of DenseNet is 81.43%. Also, with this dataset, Kadampur *et al.* [39] the performances of five deep learning models, including ResNet, SqueezeNet, and DenseNet, are compared. They reported precision of 98.19% on the training set.

Mahbod *et al.* [40] presented an ensemble of pre-trained CNN models to classify the ISIC 2018 skin lesion images. It incorporates a three-level fusion method. The prediction vectors of Efficient Net B0, Efficient Net B1, and SeResNext-50 are fused with models trained on images of different sizes, and finally, they are all combined together. It achieved a balanced multi-class accuracy of 86.25% in the test set of the ISIC 2018 challenge.

III. METHODOLOGY

In what follows, our proposal is explained in detail. Our aim is to combine several convolutional deep classifiers which form an ensemble, by merging the class information provided by the classifiers when they are run on shifted versions of the test image, where the shift vectors are distributed on a regular lattice. This way, the advantages of each classifier are exploited, while positional invariance is enhanced by the shifting procedure. The scope of our proposal is the classification of skin lesions, although the proposal could be extended to other types of medical images.

Let us note $\mathcal{F}_i$ the i-th deep convolutional classifier, where $i \in \{1, \ldots, N\}$ so that N stands for the number of classifiers in the ensemble. Each classifier produces a vector of class scores $\mathbf{z} \in \mathbb{R}^C$ for each possible input image $\mathbf{X}$, where C is the number of considered classes:

$$\mathbf{z} = \mathcal{F}_i(\mathbf{X}) \tag{1}$$

Now, let us consider shifted versions $\mathbf{X} \oplus \mathbf{s}$ of the input test image $\mathbf{X}$, where $\mathbf{s} \in \mathbb{Z}^2$ is a shift vector which indicates the displacement in pixels that is applied to $\mathbf{X}$. The circular shift operation is assumed here. Independent on how the image is shifted and which classifier is employed, the true class remains the same. The rationale of our approach is that, under these circumstances, one can merge the outputs of the networks for various shifts in order to obtain a more accurate estimation of the class.

We propose to employ a regular, square lattice of possible displacement vectors $\mathbf{s}$ which is characterized by two parameters: a pixel stride $R \in \mathbb{N}^+$ and a maximum Manhattan distance $\rho \in \mathbb{R}^+$. The displacements to be considered are those that fall into a square of side ρ around the null shift $\mathbf{s} = \mathbf{0}$ because too large shifts may compromise the ability of the networks to correctly recognize the class of the shifted image. Therefore, the set of considered shift vectors is given by:

$$\mathcal{S} = \left\{ \mathbf{s} \in \mathbb{Z}^2 \mid \mathbf{s} = (Rs_1, Rs_2), (s_1, s_2) \in \mathbb{Z}^2, \|\mathbf{s}\|_1 < \rho \right\} \tag{2}$$

where $\|\cdot\|_1$ stands for the Manhattan norm.

Each shifted version of the input test image $\mathbf{X}$ is tested with one of the classifiers in the ensemble to yield a set of tentative class scores. This is done by assigning each shift vector $\mathbf{s} \in \mathcal{S}$ uniformly at random to one out of N subsets $\mathcal{S}_i$, $i \in \{1, \ldots, N\}$, that form a partition of $\mathcal{S}$. The set of tentative class scores is obtained as follows:

$$\mathcal{T} = \bigcup_{i=1}^{N} \{\mathcal{F}_i(\mathbf{X} \oplus \mathbf{s}) \mid \mathbf{s} \in \mathcal{S}_i\} \tag{3}$$

Finally, the tentative class scores in $\mathcal{T}$ must be merged to yield a final class score vector $\hat{\mathbf{z}} \in \mathbb{R}^C$:

$$\hat{\mathbf{z}} = g(\mathcal{T}) \tag{4}$$

where g is a suitable combination function. In our experiments we have tested two different functions separately: $g =$ mean and $g =$ median.

For sake of clarity, Fig. 1 depicts a summary of the operation of the proposed model,[1] using the ensemble of two networks ($N = 2$). First, the square lattice of displacements is created and divided into two equally sized subsets. Each of these subsets are applied to the image and tested through their

---

[1]The source code of the proposed method in available online in: https://github.com/icai-uma/SLC_ShiftingEnsemble
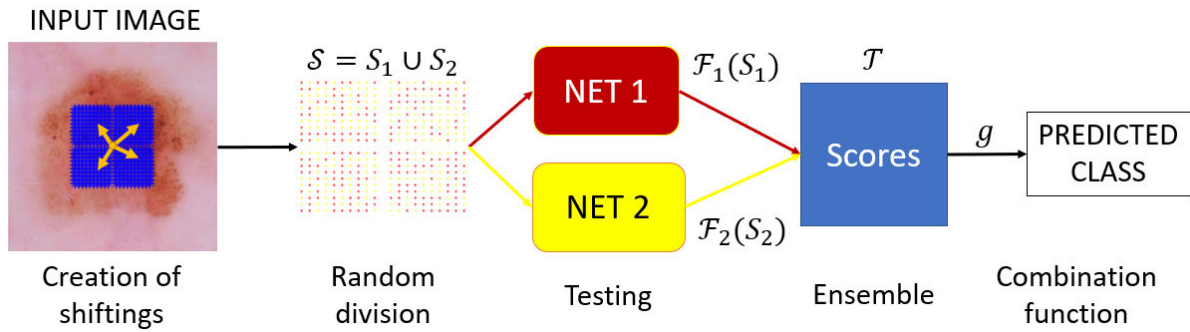
FIGURE 1. Scheme of the operation of the proposed method. The input image is shifted by defining a lattice of displacement vectors around the center of the image. Then, two subsets of shifted images are passed through the CNNs and the scores are fused using and ensemble function.

respective network. Then, the outputted scores are grouped and the combination function is applied. The predicted class is the one which corresponds to the maximum value of the computed score vector.

IV. ENSEMBLE MODEL
Here the concrete definition of the ensemble model components are described, as well as the estimation of the model parameters.
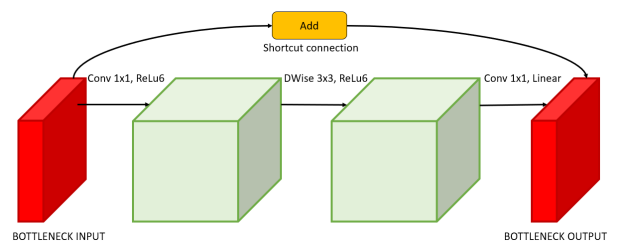
A. DEEP NETWORKS
The proposed ensemble of neural networks can be configured with any number of classification networks. The more classifiers, the lower is the computational efficiency. Thus, reaching equilibrium is very important. In our proposal, we found that with $N = 2$, using MobileNetV2 and GoogLeNet neural networks, the results are satisfactory.

The first deep network is MobileNetV2 [42]. This network is composed of an initial full convolutional layer followed by 19 residual bottleneck blocks. As shown in Fig. 2a, the latter are connected by shortcut connections in order to eliminate the non-linearity and maintain the representation of the data. This model was created as a light neural network suitable for its use in mobile devices. It has been trained on the ImageNet dataset and tested with several well-known datasets (ImageNet, COCO, VOC), demonstrating more efficiency with fewer parameters and achieving the same accuracy as its predecessor (MobileNetV1).

We also used a deeper neural network, GoogLeNet [41]. This network has an architecture based on Inception modules and it is one of the best state-of-art deep networks. The fundamentals of GoogLeNet is both increasing the depth of the network and the number of neurons at each layer. As shown in Fig. 2b, it contains 22 layers. Most of the layers correspond to Inception modules. Here the convolutional layers use a ReLU activation function. Experiments with several datasets showed that the accuracy increases substantially. Nevertheless, the required processing time is larger than others networks and of course, more than of MobileNetV2.

The input image size of both networks is 224×224 taking RGB color channels with mean subtraction. The

(a) Residual bottleneck blocks of the MobileNetV2 neural network
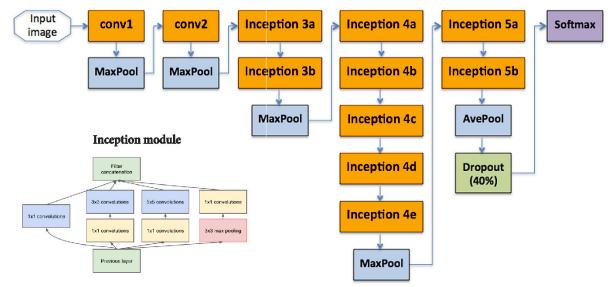
(b) GoogLeNet (image extracted from [41])

FIGURE 2. Structure of the deep networks used for the proposed ensemble model.

deep networks were fine-tuned for the melanoma classification problem in Matlab R2019b, with the following hyper-parameter values:

- Batch size $=16$.
- Learning rate $= 0.0001$.
- Validation frequency $= 10$.
- Max. epochs $= 10$.

B. MODEL PARAMETERS SETUP
The definition of the square lattice depends on the type of deep networks and the dataset used. First, the input layer of the networks restricts the maximum values of the displacement vectors $\mathbf{s} \in \mathbb{Z}^2$. Both MobileNetV2 and GoogLeNet require an input of size 224×224 pixels, so the images need to be resized. Furthermore, the maximum Manhattan distance allowed is $\rho_{max} = 224$. Taking into consideration the image features is essential since oversized shiftings may distort the original shape of the moles.
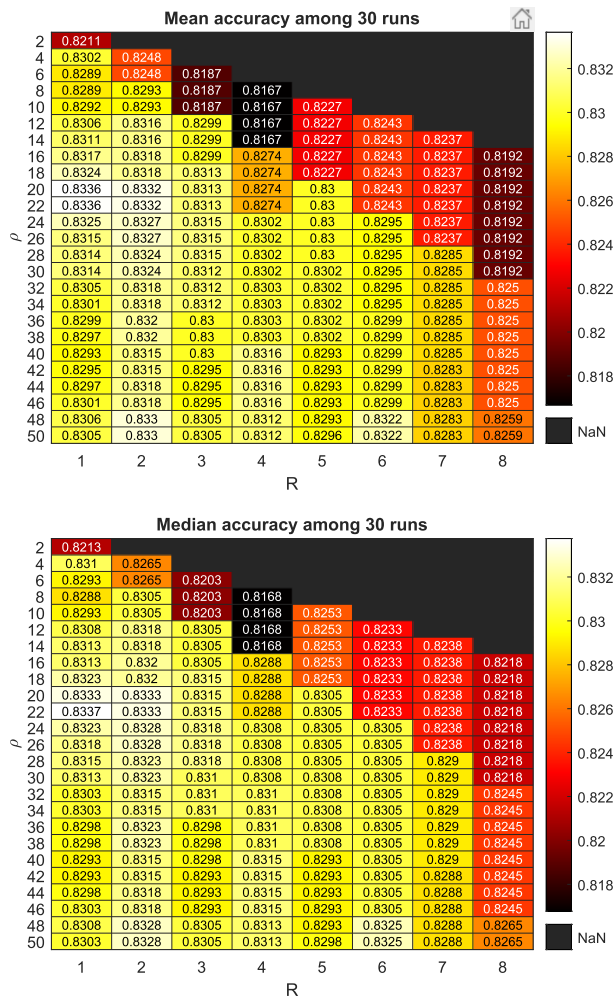
Mean accuracy among 30 runs

ρ \ R	1	2	3	4	5	6	7	8
2	0.8211							
4	0.8302	0.8248						
6	0.8289	0.8248	0.8187					
8	0.8289	0.8293	0.8187	0.8167				
10	0.8292	0.8293	0.8187	0.8167	0.8227			
12	0.8306	0.8316	0.8299	0.8167	0.8227	0.8243		
14	0.8311	0.8316	0.8299	0.8167	0.8227	0.8243	0.8237	
16	0.8317	0.8318	0.8299	0.8274	0.8227	0.8243	0.8237	0.8192
18	0.8324	0.8318	0.8313	0.8274	0.8227	0.8243	0.8237	0.8192
20	0.8336	0.8332	0.8313	0.8274	0.83	0.8243	0.8237	0.8192
22	0.8336	0.8332	0.8313	0.8274	0.83	0.8243	0.8237	0.8192
24	0.8325	0.8327	0.8315	0.8302	0.83	0.8295	0.8237	0.8192
26	0.8315	0.8327	0.8315	0.8302	0.83	0.8295	0.8237	0.8192
28	0.8314	0.8324	0.8315	0.8302	0.83	0.8295	0.8285	0.8192
30	0.8314	0.8324	0.8312	0.8302	0.8302	0.8295	0.8285	0.8192
32	0.8305	0.8318	0.8312	0.8303	0.8302	0.8295	0.8285	0.825
34	0.8301	0.8318	0.8312	0.8303	0.8302	0.8295	0.8285	0.825
36	0.8299	0.832	0.83	0.8303	0.8302	0.8299	0.8285	0.825
38	0.8297	0.832	0.83	0.8303	0.8302	0.8299	0.8285	0.825
40	0.8293	0.8315	0.83	0.8316	0.8293	0.8299	0.8285	0.825
42	0.8295	0.8315	0.8295	0.8316	0.8293	0.8299	0.8283	0.825
44	0.8297	0.8318	0.8295	0.8316	0.8293	0.8299	0.8283	0.825
46	0.8301	0.8318	0.8295	0.8316	0.8293	0.8299	0.8283	0.825
48	0.8306	0.833	0.8305	0.8312	0.8293	0.8322	0.8283	0.8259
50	0.8305	0.833	0.8305	0.8312	0.8296	0.8322	0.8283	0.8259

Median accuracy among 30 runs

ρ \ R	1	2	3	4	5	6	7	8
2	0.8213							
4	0.831	0.8265						
6	0.8293	0.8265	0.8203					
8	0.8288	0.8305	0.8203	0.8168				
10	0.8293	0.8305	0.8203	0.8168	0.8253			
12	0.8308	0.8318	0.8305	0.8168	0.8253	0.8233		
14	0.8313	0.8318	0.8305	0.8168	0.8253	0.8233	0.8238	
16	0.8313	0.832	0.8305	0.8288	0.8253	0.8233	0.8238	0.8218
18	0.8323	0.832	0.8315	0.8288	0.8253	0.8233	0.8238	0.8218
20	0.8333	0.8333	0.8315	0.8288	0.8305	0.8233	0.8238	0.8218
22	0.8337	0.8333	0.8315	0.8288	0.8305	0.8233	0.8238	0.8218
24	0.8323	0.8328	0.8318	0.8308	0.8305	0.8305	0.8238	0.8218
26	0.8318	0.8328	0.8318	0.8308	0.8305	0.8305	0.8238	0.8218
28	0.8315	0.8323	0.8318	0.8308	0.8305	0.8305	0.829	0.8218
30	0.8313	0.8323	0.831	0.8308	0.8308	0.8305	0.829	0.8218
32	0.8303	0.8315	0.831	0.831	0.8308	0.8305	0.829	0.8245
34	0.8303	0.8315	0.831	0.831	0.8308	0.8305	0.829	0.8245
36	0.8298	0.8323	0.8298	0.831	0.8308	0.8305	0.829	0.8245
38	0.8298	0.8323	0.8298	0.831	0.8308	0.8305	0.829	0.8245
40	0.8293	0.8315	0.8298	0.8315	0.8293	0.8305	0.829	0.8245
42	0.8293	0.8315	0.8293	0.8315	0.8293	0.8305	0.8288	0.8245
44	0.8298	0.8318	0.8293	0.8315	0.8293	0.8305	0.8288	0.8245
46	0.8303	0.8318	0.8293	0.8315	0.8293	0.8305	0.8288	0.8245
48	0.8308	0.8328	0.8305	0.8313	0.8293	0.8325	0.8288	0.8265
50	0.8303	0.8328	0.8305	0.8313	0.8298	0.8325	0.8288	0.8265

FIGURE 3. Study of optimal hyperparameters. Mean and median accuracy varying the stride and the maximum shifting. The X and Y axes represent the stride and the side of the square lattice. 30 random divisions of the lattice were computed.

Therefore, analyzing the data visually, we found that most of the moles are placed in the center of the image with a margin of 40 pixels around each side. Thus, we need to define $\rho \ll 224$. In addition, the square grid can be generated with different stride sizes. A tiny stride would create a very dense lattice, augmenting the computational cost. However, a large stride would produce few score vectors, and the combination function g would not be precise enough.

In order to study which are the optimal parameters of the ensemble model, we carried out a parameter optimization, running a batch of 30 random executions (referred to the $\mathcal{S}_i$ sets) varying the values of the square side ρ and the stride R. These runs were done using a tenth part of the dataset (details can be found in Section V). The results are shown in Fig. 3. The mean was used as the combination function g (similar results were obtained with the median). In this analysis, where the mean and median among the 30 runs were computed, it can be observed that an intermediate value of ρ provides a better overall performance. Besides, using a big stride R does not provide good performance. In both analysis,

we found that the best configurations are obtained for $\rho = 20$ or $\rho = 22$, using $R = 1$, and in second place, $R = 2$.

Therefore, we have adopted an intermediate position in our work, taking $\rho = 22$ and a stride of $R = 1$ pixels. This defines a lattice $\mathcal{S}$ of 484 displacement vectors. Depending on the time and precision requirements, these values can be chosen in a different way. For example, a larger stride ($R = 2$ or $R = 3$) can provide similar results with less shifts.

V. EXPERIMENTS

This section describes the dataset, the evaluation metrics, the experimental setup, and the discussion of the results obtained from the set of experiments.

A. DATASET

The evaluation of the proposed method is carried out by using a well-known dataset of labeled dermoscopic images, called HAM10000 [43]. This dataset contains 10,050 images divided into seven classes (see examples in Fig. 4):

1) Actinic keratosis (akiec): usually not pigmented.
2) Basal cell carcinoma (bcc): flat, nodular, pigmented, cystic.
3) Benign keratosis (bkl): often biopsied or excised and similar to melanoma.
4) Dermatofibroma (df): reticular lines at the periphery with a white center.
5) Melanoma (mel): malign neoplasms which are chaotic
6) Nevi (nv): benign neoplasms.
7) Vascular skin (vasc): red or purple circle.

This data was collected from the Medical University of Viena and Cliff Rosendahl in Queensland. These are two prestigious institutions in Austria and Australia, respectively. The International Skin Imaging Collaboration (ISIC) 2018 challenge and posterior editions have included this dataset within their competition. This has become a benchmark for testing new dermatological classification and segmentation techniques.

However, the main disadvantage of this dataset is the irregular distribution of the number of diseases. Most of the images, exceeding 70% of the total number of images, correspond to the nevi class. This sets up an extreme dataset imbalance. This fact severely affects the training, provoking an extreme specialization in the nevi class. In the second level is situated the bkl class, with around 13% of the images. The rest of classes represent a large minority of the samples. In particular, the df class, with less than 2%, will be the most difficult class to predict. Therefore, one may think about using data augmentation to balance the data and make the learning procedure more robust. In this work, we show that this is not an essential requirement for the proposed ensemble model.

We also carried out a set of experiments with data augmentation using different reflections and rotations of the original images. The specific type of transformations used include:

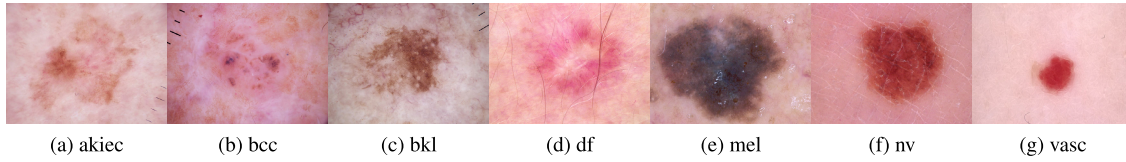- Horizontal and vertical flipping with a probability of 0.5.

| (a) akiec | (b) bcc | (c) bkl | (d) df | (e) mel | (f) nv | (g) vasc |

FIGURE 4. Example images of the seven classes of HAM10000 dataset.

- Random image rotations between $-90°$ and $90°$, with a probability of 0.75.

It should be considered that this data augmentation may smooth but not solve the great ratio between the nevi class and the others.

B. EVALUATION METRICS

Typical classification measures were used to analyze the performance of the shifting model. Since we are dealing with a seven-class problem, these measures will be computed, binarizing the results depending on the analyzed class. Thus, the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are computed as well as the following measures:

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

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \tag{9}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

which are the accuracy, sensitivity (True Positive Rate), specificity (True Negative Rate), precision (Positive Predictive Value), F_1-score and Matthews Correlation Coefficient, respectively. The metrics range from 0 to 1, where higher measures indicate better performance.

The *Sensitivity* and the *Specificity* provide a measure of how well the method is classifying the relevant instances. The *Acc* and F_1 provide a general overview of the performance, taking into account the positive and negative samples. The latter gives equal importance to precision and recall. The *MCC* takes into account the TP, TN, FP, and FN, being a balanced measure that can be used even if the classes are of very different sizes.

C. EXPERIMENTAL SETTING

Our ensemble model method, which we called *Shifted MobileNetV2 + GoogLeNet* or *Shifted 2-Nets*, is compared with the following methods:
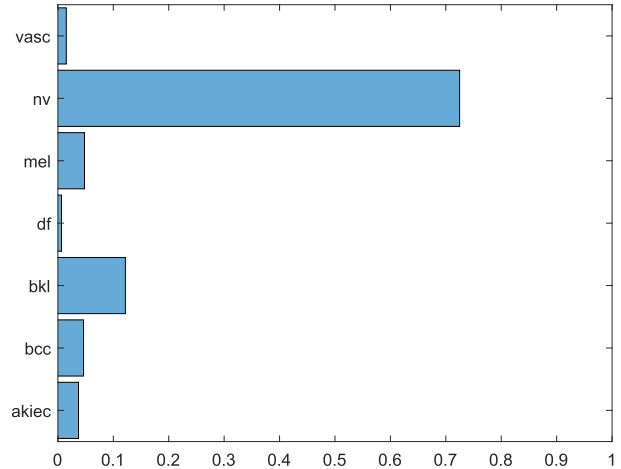
FIGURE 5. Probability histogram of the HAM10000 dataset.

- *Raw MobileNetV2 and Raw GoogLeNet:* The deep network directly tests the original image once, without any modifications.
- *Shifted MobileNetV2 and Shifted GoogLeNet:* The proposed shifting model, using the same configuration explained in Subsection IV-B, is used with the deep network to test the image.

We carried out 10-fold cross-validation for all the executions. The following division of the dataset was used: 70% of samples for training, 10% for validation, and 20% for testing. The data was distributed in a balanced manner, that is, the proportion of images belonging to each class shown in Fig. 5 is respected within each set. Thus, we should find the same distribution of classes within each of the ten folds. This way, we directly visualize the benefit or deficit of the ensemble model compared with the original networks.

In addition to this, the proposed method *Shifted MobileNetV2 + GoogLeNet* entails a random division of the set of displacement vectors $\mathcal{S}$. Thus, to have a realistic statistically significant comparison. A total of 30 random divisions were computed and tested for each execution. The mean and standard deviation values were calculated as the final performance of the ensemble method.

D. RESULTS AND DISCUSSIONS

The first batch of experiments are reported in Figs. 6-9. The result of each of the ten folds of the cross-validation procedure is reported for each tested method. Besides, the plotted bar of our proposed model contains a small error bar in
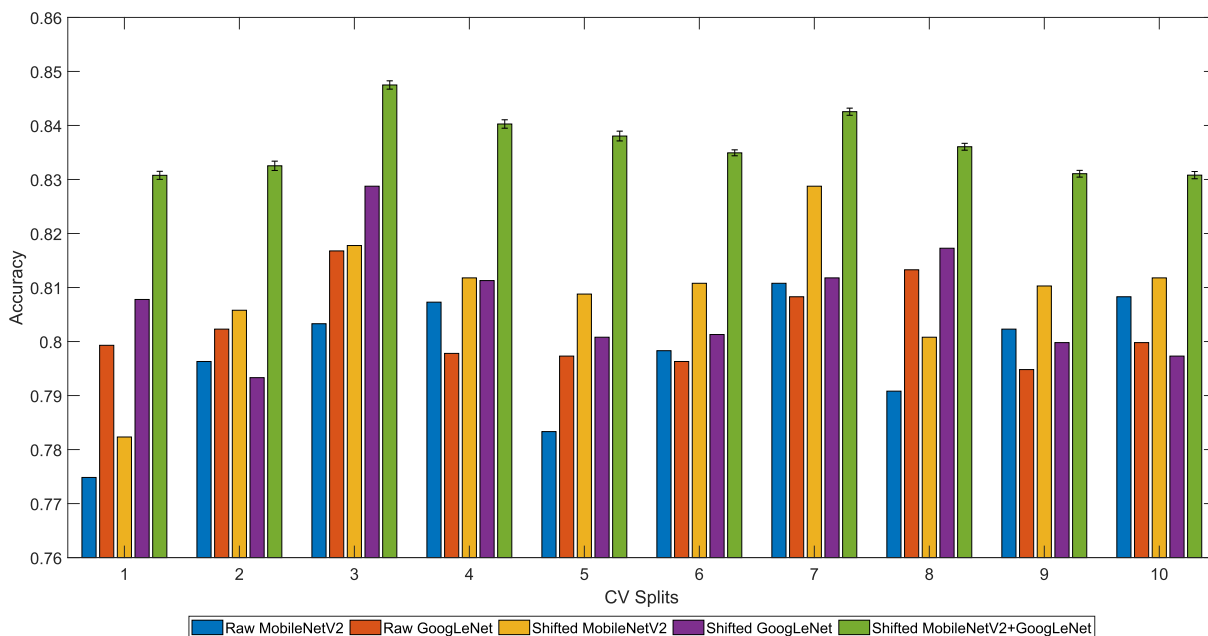
FIGURE 6. Comparison of the proposed models (without data augmentation) using the mean as the combination function. Accuracy of each cross-validation split is presented. The error bar represents the variability of the random divisions of the lattice.
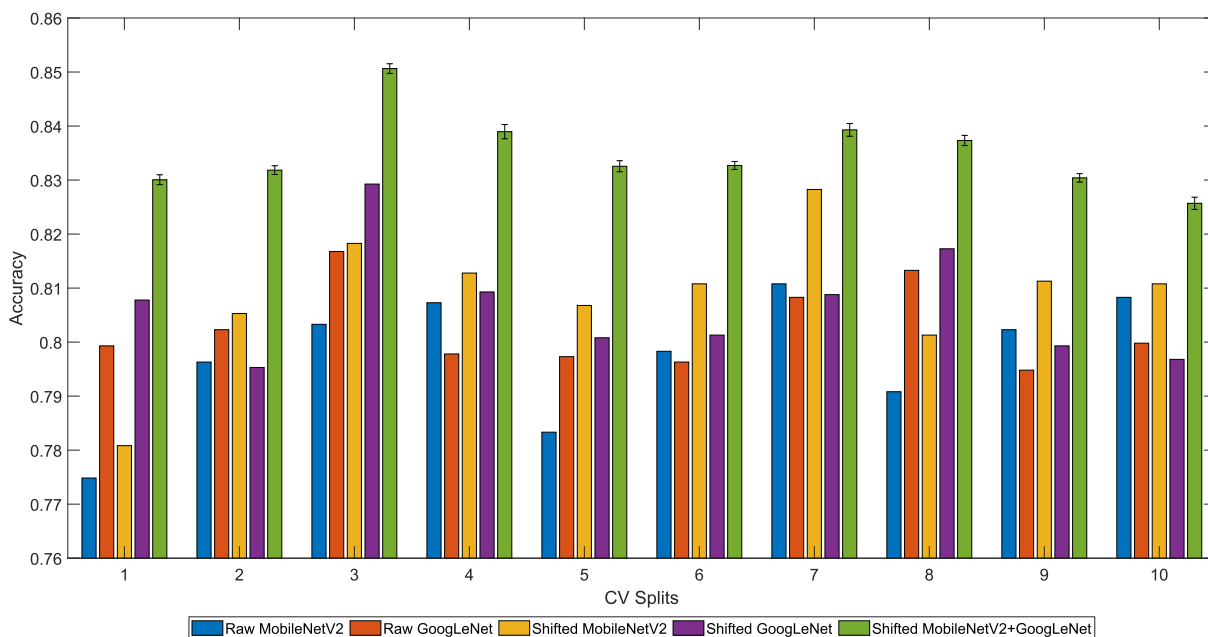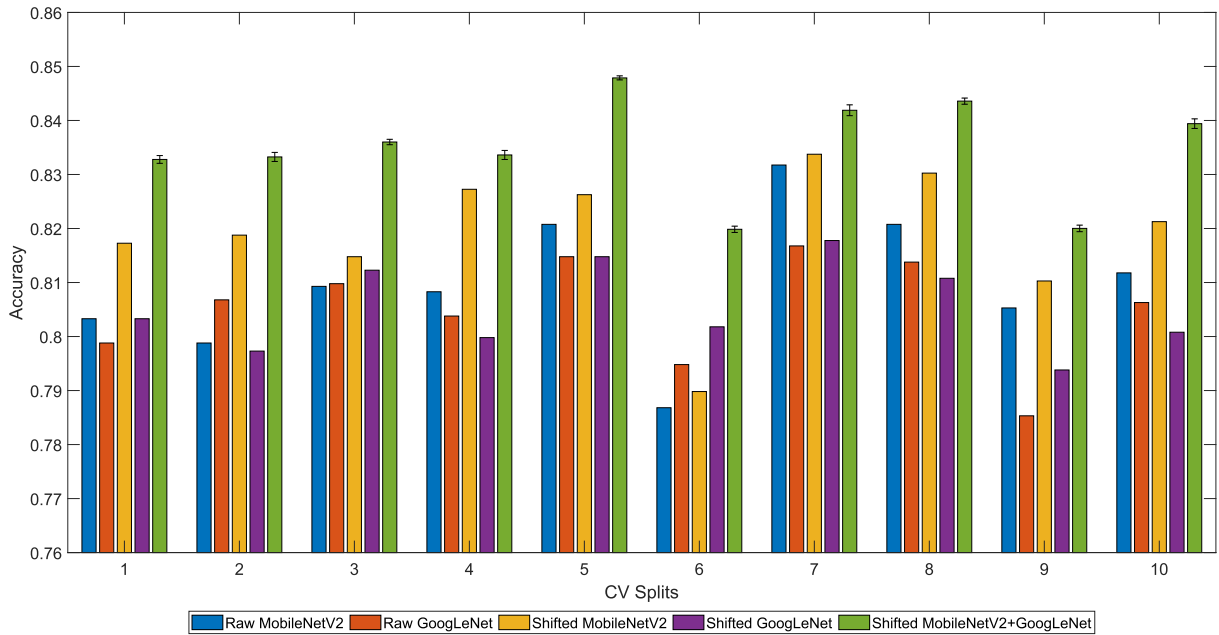

FIGURE 7. Comparison of the proposed models (without data augmentation) using the median as the combination function. Accuracy of each cross-validation split is presented. The error bar represents the variability of the random divisions of the lattice.

its peak. This represents the variability among the 30 random divisions of the shifting lattice.

In the first instance, the results obtained using the models trained without data augmentation are presented. In Fig. 6 the used combination function was the mean. One can observe that there is a great difference between the ensemble of 2 nets and the simple nets. The former is more than 3% more accurate in most of the splits. If we focus only on the shifting

models, there are splits where the MobileNetV2 worked better than GoogLeNet, and vice versa, but there is no clear pattern.

The results of the median combination function are depicted in Fig. 7. The tendency is quite similar to the mean function. The ensemble method yielded larger percentages of accuracy for all splits. Moreover, in both cases, the error among the 30 independent runs of the proposed

FIGURE 8. Comparison of the proposed models (with data augmentation) using the mean as the combination function. Accuracy of each cross-validation split is presented. The error bar represents the variability of the random divisions of the lattice.
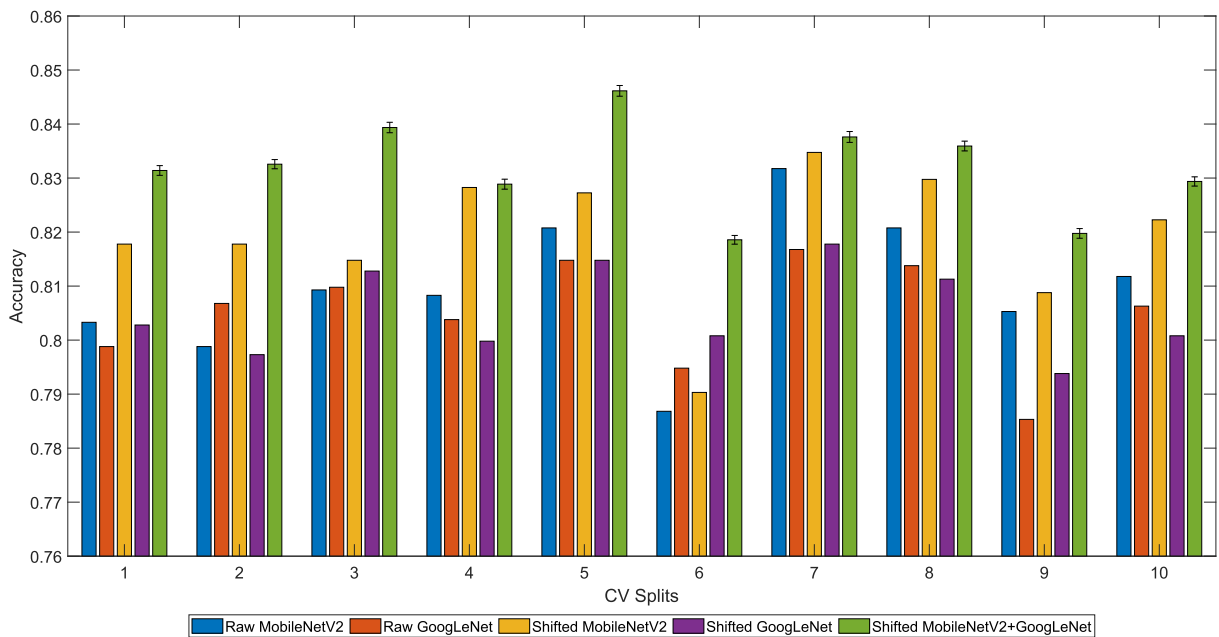
FIGURE 9. Comparison of the proposed models (with data augmentation) using the median as the combination function. Accuracy of each cross-validation split is presented. The error bar represents the variability of the random divisions of the lattice.

model is minimal. This indicates that the method is quite robust regardless of the type of data tested.

The performance increase using the proposed method is remarkable. In some cases, like in split 1, there is an improvement of almost 6%, being both raw models used in the ensemble method. Alternatively, the shifting model alone is not enough to significantly improve the classification

performance (only around 1%) and sometimes accuracy tends to decrease.

The next two figures show the outcomes for the models trained with data augmentation. Fig. 8 represents the results of the mean combination function. Now there are differences in the split's performance since the training was varied with more data. The ensemble model is still the best
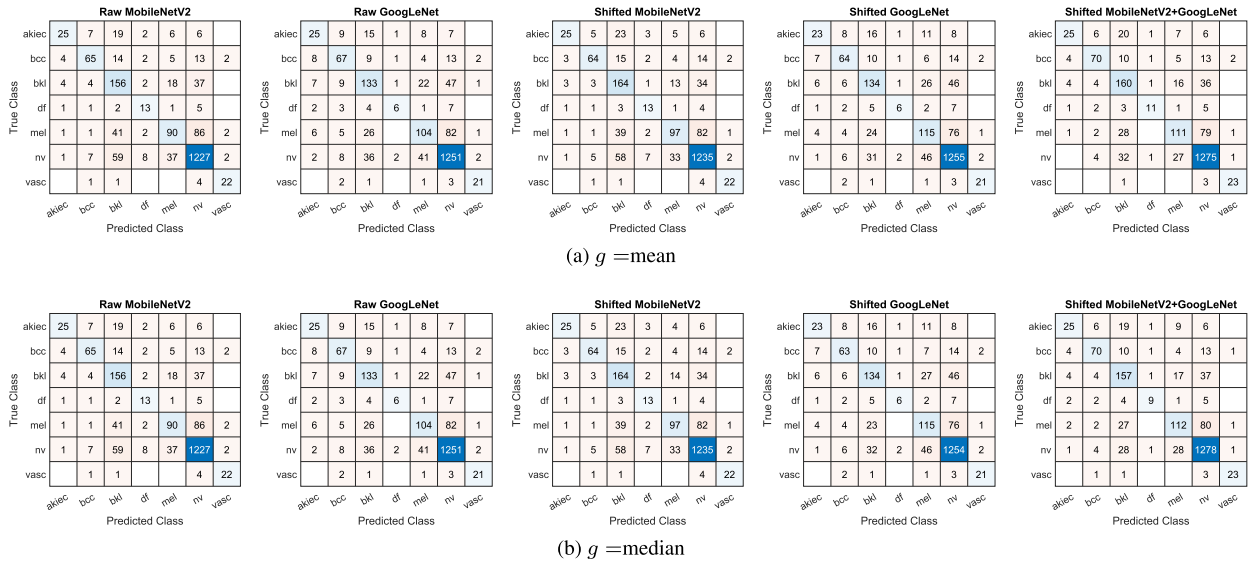
FIGURE 10. Comparison of the confusion matrices for the five tested models trained without data augmentation. This statistics were generated rounding the average of the confusion matrices of the cross-validation procedure.
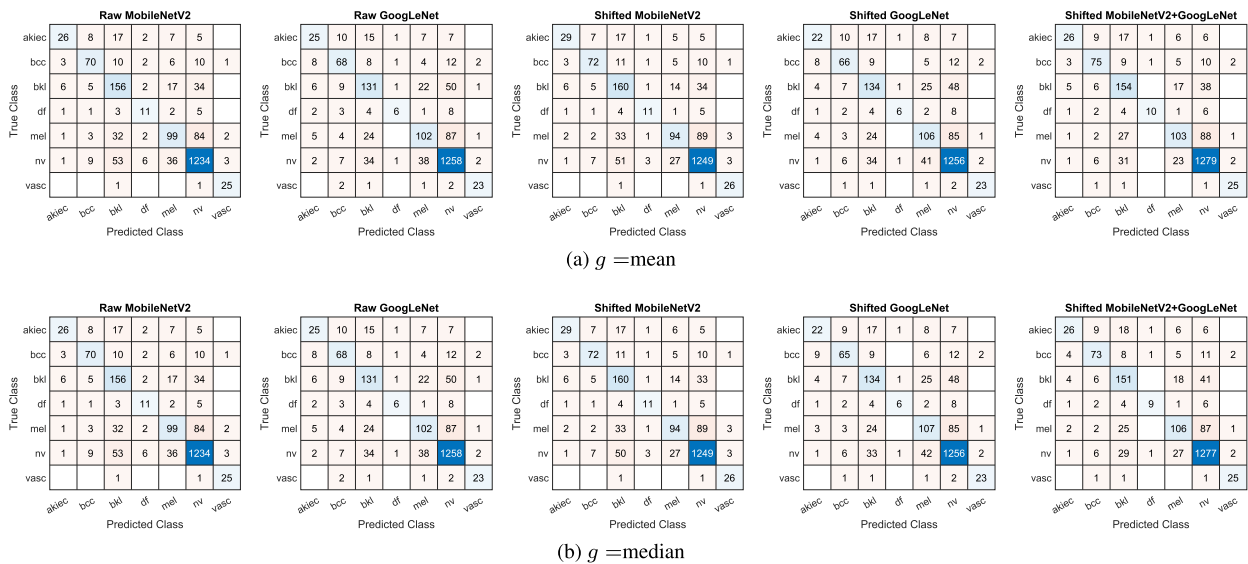

FIGURE 11. Comparison of the confusion matrices for the five tested models trained with data augmentation. This statistics were generated rounding the average of the confusion matrices of the cross-validation procedure.

classification method. This is closely followed by the *Shifted MobileNetV2*. There are some splits where the *Raw MobileNetV2* overcomes the *Shifted GoogLeNet*. This indicates that a deeper network is not always suitable for its application on specific tasks.

Results of the median combination function (Fig. 9) present even closer results. Although the simple models worked well, our proposal is still the best. In split 7, a similar behavior is noted. The raw model overcomes the shifted one. The median function yields slightly results than the mean function. This may indicate that among the 484 shiftings, there are not outlying classifications.

In general terms, we can state that data augmentation is vital for the raw and shifting models. This not however the

case for the ensemble model since the overall performance of the splits (the maximum accuracy reached) is closely similar. This can be useful if we need to develop an online method where the model needs to be re-trained again. If no data augmentation is used, the training time will be reduced.

Next, the confusion matrices are analyzed. Figs. 10 and 11 show the average confusion matrix of the cross-validation tests. Without loss of generality, since this calculation can generate fractional numbers, we have rounded the computed average values. In figeure Fig. 10 the results of the models without data augmentation are presented. The first observation is that the number of wrong classifications has been reduced in the *Shifted MobileNetV2 + GoogLeNet* model. Here the number of blank squares is higher (nine instead

TABLE 1. Evaluation metrics for the five compared models trained without data augmentation. Mean and standard deviation values of the cross-validation procedure are reported.

Function g	mean				
Measure	Raw MobileNetV2	Raw GoogLeNet	Shifted MobileNetV2	Shifted GoogLeNet	**Shifted 2-Nets**
Accuracy	0.798 ± 0.012	0.803 ± 0.008	0.809 ± 0.012	0.807 ± 0.011	0.836 ± 0.006
Sensitivity	0.626 ± 0.040	0.580 ± 0.025	0.636 ± 0.049	0.576 ± 0.032	0.649 ± 0.033
Specificity	0.947 ± 0.004	0.947 ± 0.003	0.950 ± 0.004	0.948 ± 0.004	0.955 ± 0.003
Precision	0.677 ± 0.029	0.667 ± 0.033	0.703 ± 0.030	0.678 ± 0.039	0.760 ± 0.025
F_1	0.634 ± 0.020	0.606 ± 0.020	0.648 ± 0.030	0.607 ± 0.030	0.687 ± 0.022
MCC	0.593 ± 0.021	0.566 ± 0.020	0.613 ± 0.030	0.569 ± 0.030	0.656 ± 0.021
Function g	median				
Accuracy	0.798 ± 0.012	0.803 ± 0.008	0.809 ± 0.012	0.807 ± 0.010	0.835 ± 0.007
Sensitivity	0.626 ± 0.040	0.580 ± 0.025	0.637 ± 0.047	0.575 ± 0.031	0.638 ± 0.025
Specificity	0.947 ± 0.004	0.947 ± 0.003	0.950 ± 0.004	0.948 ± 0.004	0.954 ± 0.003
Precision	0.677 ± 0.029	0.667 ± 0.033	0.703 ± 0.029	0.674 ± 0.035	0.756 ± 0.032
F_1	0.634 ± 0.020	0.606 ± 0.020	0.649 ± 0.029	0.607 ± 0.027	0.677 ± 0.016
MCC	0.593 ± 0.021	0.566 ± 0.020	0.613 ± 0.028	0.568 ± 0.028	0.647 ± 0.017

TABLE 2. Evaluation metrics for the five compared models trained with data augmentation. Mean and standard deviation values of the cross-validation procedure are reported.

Function g	mean				
Measure	Raw MobileNetV2	Raw GoogLeNet	Shifted MobileNetV2	Shifted GoogLeNet	**Shifted 2-Nets**
Accuracy	0.810 ± 0.013	0.805 ± 0.010	0.819 ± 0.013	0.805 ± 0.008	0.835 ± 0.009
Sensitivity	0.649 ± 0.037	0.588 ± 0.042	0.659 ± 0.031	0.581 ± 0.046	0.656 ± 0.032
Specificity	0.951 ± 0.004	0.947 ± 0.005	0.952 ± 0.004	0.947 ± 0.005	0.954 ± 0.004
Precision	0.687 ± 0.040	0.689 ± 0.026	0.714 ± 0.037	0.685 ± 0.038	0.766 ± 0.024
F_1	0.653 ± 0.030	0.613 ± 0.040	0.670 ± 0.031	0.608 ± 0.046	0.688 ± 0.026
MCC	0.614 ± 0.030	0.576 ± 0.035	0.634 ± 0.029	0.571 ± 0.044	0.659 ± 0.023
Function g	median				
Accuracy	0.810 ± 0.013	0.805 ± 0.010	0.819 ± 0.013	0.805 ± 0.008	0.832 ± 0.009
Sensitivity	0.649 ± 0.037	0.588 ± 0.042	0.659 ± 0.031	0.581 ± 0.045	0.644 ± 0.034
Specificity	0.951 ± 0.004	0.947 ± 0.005	0.952 ± 0.004	0.947 ± 0.005	0.953 ± 0.004
Precision	0.687 ± 0.040	0.689 ± 0.026	0.714 ± 0.037	0.685 ± 0.035	0.761 ± 0.021
F_1	0.653 ± 0.030	0.613 ± 0.040	0.670 ± 0.031	0.608 ± 0.044	0.678 ± 0.027
MCC	0.614 ± 0.030	0.576 ± 0.035	0.634 ± 0.030	0.572 ± 0.041	0.649 ± 0.023

of six). Besides, for most classes, the number of correct predictions has increased. This especially the case for the nv, bcc, bkl, and vasc classes. The use of the median as the combination function yielded similar outcomes. Only two or three predictions by class varied. The raw models misclassify most the melanoma, basal cell carcinoma, benign keratosis and the nevi classes. Some improved results can be observed by the shifting models. However, the results can be a bit contradictory. For instance, *Shifted MobileNetV2* is good at predicting bkl. This not the case for the *Shifted GoogLeNet* model. The second model is good at predicting melanomas but not the first one. The ensemble model achieves an equilibrium.
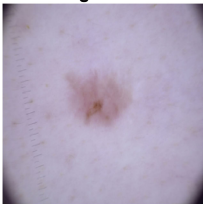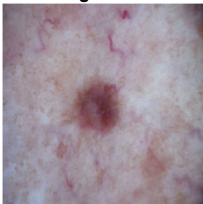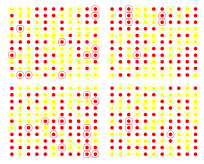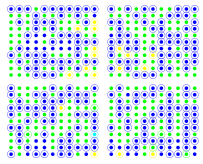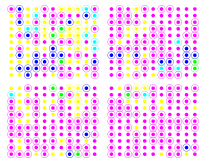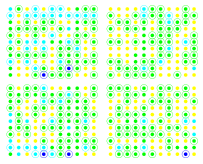
The outcomes of the augmented data are displayed in Fig. 11. The overall performance is better than the predictions of the previous figure. Analyzing the results class by class, we found that the number of true positives (the predominant class nv) is high, but the true negatives (the rest) are low. This fact depends on the training, so a direct comparison between both methodologies is not entirely fair. Nonetheless, the nv and bcc classes have improved their predictions (1279 and 75 correct classifications, resp.). This indicates that the data augmentation improved results but it is not enough to solve the imbalance between the seven classes. Focusing on the results of this specific comparison, the tendency is similar. Our proposal improves or adopts

an intermediate position between the two shifting models. The raw models provided a small improvement, but they are still far from the ensemble model. Finally, the values of the differences between the mean and median are not too high. The mean achieves 9 blank squares instead of the 8 of the median, so the latter achieved lower results.

The previous plots provided a detailed analysis of the accuracy for each method. In order to have a better comparison of the performance, the mean and standard deviation of several evaluation metrics are presented in Tables 1 and 2. In addition to the accuracy, which has been commented on before, the ensemble model (*Shifted 2-Nets*) generated the best statistics for most cases. That is, the specificity, precision, F_1-score, and *MCC* are quite better for both $g = $ mean and $g = $ median.

Analyzing the case of the models trained without data augmentation (Table 1), the mean accuracy reached 83.6% and the average specificity of the seven classes is 95.5%. This indicates a high level of negative predictions. The median function yielded similar results but slightly worse than the mean of scores. Across the tested models, a great difference is appreciable, having an increment of 3% of accuracy and almost 7% of *MCC* and F_1-score in some cases. *Shifted GoogLeNet* has not shown any improvement with respect to the raw model. However, the MobileNetV2 network is shows

TABLE 3. Examples of the outputs generated by the compared models (trained with data augmentation). The color represents the class, and the circle/no-circle is the division of the grid into two sets: circle are for MobileNetV2 and no-circle for GoogLeNet.

	ISIC image nº 0027708	ISIC image nº 0031513	ISIC image nº 0031309	ISIC image nº 0032910
Image				
Lattice				
Legend	* akiec　* bcc　* bkl　* df　* mel　* nv　* vasc			
Raw MobileNetV2	bkl	bkl	nv	bkl
Raw GoogLeNet	akiec	nv	df	bkl
Shifted MobileNetV2	bkl	bcc	df	nv
Shifted GoogLeNet	akiec	nv	df	bkl
Shifted 2-Nets	akiec	bcc	df	nv
GT	akiec	bcc	df	nv

higher accuracy with an improvement of around 2% in all measures.

Regarding the augmented models (reported in Table 2), the highest accuracy obtained is of 83.5%, while the highest F_1-score is 68.8%. Since the classes are unbalanced, the second value is more significant to compare the methods because it considers both the precision and recall of all classes. Thus, *Shifted MobileNetV2* model reached 67% in the F_1, and the raw models are far from this percentage. Similar error differences are found with the other measures. Comparing the mean and the median, the first one seemed to be more effective.

Comparing both types of training, the main conclusion that can be extracted is that data augmentation does not contribute significantly to the shifting and ensemble models, obtaining very similar results. However, for the raw models, it is necessary to include this preprocessing to have a better-generalized model. That is, our model eliminates the need to use data augmentation.

The comparison with the previous works reported in Section II should be made carefully since the datasets and evaluation employed are not always the same. The works that used the same dataset as us [38]–[40] also used transfer learning, but only one was evaluated using HAM10000. They reported 81.62% and 81.43% of overall accuracy, while our method yielded 83.6% accuracy. On the other hand, it should be noticed that the clinical appliance has to be decided by medicals since sometimes the obtained accuracy could be insufficient for a diagnostic in the medical field.

We carried out a visual inspection of the predictions to check and understand the behavior of the ensemble model. For that purpose, Table 3 depict four examples, with their

respective grid division and prediction and the outputs of each model. Please note that the number of circled stars (assigned to MobineNetV2) is the same that the number of non-circled stars (assigned to GoogLeNet) since the random division has the same size but different positions.

The first image corresponds to actinic keratosis. The MobileNetV2 model predicted more incorrect shifts (yellow) than correct ones (red). The final prediction was wrong. Nevertheless, all the predictions of GoogLeNet (non-circled red stars) were identified as the akiec class. That is because, in this case, the GoogLeNet model worked better.

The second example is that of basal cell carcinoma. This case is very interesting since the *Raw MobileNetV2* outputted a bkl, while most of the shifted images yielded the nv or the bcc classes. The reason is unknown, but we can observe a yellow star near the center of the lattice ($\mathbf{s} = 0$), which would explain this behavior. The MobileNetV2 network predicted more bcc's than GoogLeNet, and that caused the different outputs. Thus, these two first examples showed that there are cases where one network works better than the other. Compensation is achieved when the predictions of both are merged.

The third example concerns the class dermatofibroma. This is am underrepresented class within the HAM10000 dataset. If the image is shifted to the upper-left, the networks were not able to have a firm decision on the predicted class. However, bottom-right shifts clearly yielded the df class. In this case, both networks have similar behavior. The last image is a nevus. This image is challenging because the skin contains many irregularities. MobileNetV2 predicted more nv and mel, while GoogLeNet classified it as bkl and nv. Here the combination function was essential to making an

adequate prediction. Our ensemble model dealt with many bkl predictions because the nevi class scores were higher in most of the predictions.

VI. CONCLUSION

A new methodology to perform skin lesion classification with deep convolutional neural networks was proposed. It consists of constructing an ensemble of convolutional neural networks that cooperate to yield a more accurate assessment of the lesion. This is attained by considering multiple shifted versions of the test input image so that the shift vectors form a regular lattice. Each shifted version is allocated to one of the networks of the ensemble. After that, the shifted versions of the test image are processed. The resulting class score vectors are combined by a suitable aggregation function in order to produce the final classification result. This strategy exploits the strengths of the networks that comprise the ensemble. The aggregation scheme alleviates the deleterious effect of individual classification failures. Therefore, our proposal is more robust than the standard convolutional neural network classification procedure. Also, it must be highlighted that our approach is not related to standard train time data augmentation by training image shifting.

Experimental results demonstrate how the proposed shifting technique outperforms traditional deep learning techniques for skin lesions classification. The proposed ensemble + shifting model is around 3% better than the deep networks with shifting and almost 6% better than the simple network in all classification performance measures. This is particularly the case in F_1-score that is the harmonic mean of the precision and recall. Specially, the major improvements of true predictions has been obtained for the bcc and bkl classes. The plain models behaved better when the ones trained with data augmentation. However this technique was unnecessary for the ensemble model to achieve the same results, with almost an 84% accuracy on the HAM10000 dataset. The lack of enough training images affected the generalization of all networks. The effect appears to be more severe for the raw models. The *Shifted MobileNetV2 + GoogLeNet* compensated this effect by defining an extensive set of displacements that covered many transformations of the original input image. Note that each deep network was trained and tested with the same configuration to fairly compare all models' performance with the same parameter values.

Further works will be focused on the testing of more deep networks and other topologies of the lattice. The image features are crucial to understand each class and to generate an adequate classifier for dermoscopic images. Other image transformations, such as rotations combined with the proposed shifts, may improve the generalization level of the model. The inclusion of more complex combination functions, such as probabilistic models, is another path to be explored to enhance predictive accuracy. Moreover, the proposed technique is planned to be applied to other medical image classification problems.

REFERENCES

[1] 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, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, Feb. 2021.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA, Cancer J. Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[3] *Cancer Facts & Figures*, American Cancer Society, Atlanta, GA, USA, 2016.

[4] J. Trufant and E. Jones, *Skin Cancer for Primary Care*. Cham, Switzerland: Springer, 2019, pp. 171–208.

[5] A. G. Priya H., J. Anitha, and J. Poonima J., "Identification of melanoma in dermoscopy images using image processing algorithms," in *Proc. Int. Conf. Control, Power, Commun. Comput. Technol. (ICCPCCT)*, Mar. 2018, pp. 553–557.

[6] M. H. Jafari, N. Karimi, E. Nasr-Esfahani, S. Samavi, S. M. R. Soroushmehr, K. Ward, and K. Najarian, "Skin lesion segmentation in clinical images using deep learning," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 337–342.

[7] M. H. Jafari, E. Nasr-Esfahani, N. Karimi, S. M. R. Soroushmehr, S. Samavi, and K. Najarian, "Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 6, pp. 1021–1030, Jun. 2017.

[8] B. Devassy, S. Yildirim-Yayilgan, and J. Hardeberg, "The impact of replacing complex hand-crafted features with standard features for melanoma classification using both hand-crafted and deep features," *Adv. Intell. Syst. Comput.*, vol. 868, pp. 150–159, Sep. 2019.

[9] V. Yadav and V. Kaushik, "Detection of melanoma skin disease by extracting high level features for skin lesions," *Int. J. Adv. Intell. Paradigms*, vol. 11, nos. 3–4, pp. 397–408, 2018.

[10] M. Ruela, C. Barata, J. S. Marques, and J. Rozeira, "A system for the detection of melanomas in dermoscopy images using shape and symmetry features," *Comput. Methods Biomech. Biomed. Eng., Imag. Visualizat.*, vol. 5, no. 2, pp. 127–137, Mar. 2017.

[11] S. Bakheet, "An SVM framework for malignant melanoma detection based on optimized HOG features," *Computation*, vol. 5, no. 4, p. 4, Jan. 2017.

[12] A. Victor and M. Ghalib, "Automatic detection and classification of skin cancer," *Int. J. Intell. Eng. Syst.*, vol. 10, no. 3, pp. 444–451, Jun. 2017.

[13] R. B. Oliveira, J. P. Papa, A. S. Pereira, and J. M. R. S. Tavares, "Computational methods for pigmented skin lesion classification in images: Review and future trends," *Neural Comput. Appl.*, vol. 29, no. 3, pp. 613–636, Feb. 2018.

[14] F. P. dos Santos and M. A. Ponti, "Robust feature spaces from pre-trained deep network layers for skin lesion classification," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 189–196.

[15] A. H. Shahin, A. Kamal, and M. A. Elattar, "Deep ensemble learning for skin lesion classification from dermoscopic images," in *Proc. 9th Cairo Int. Biomed. Eng. Conf. (CIBEC)*, Dec. 2018, pp. 150–153.

[16] T. Zhou, K.-H. Thung, X. Zhu, and D. Shen, "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis," *Hum. Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019.

[17] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, 2018.

[18] J. Li, G. Zhou, Y. Qiu, Y. Wang, Y. Zhang, and S. Xie, "Deep graph regularized non-negative matrix factorization for multi-view clustering," *Neurocomputing*, vol. 390, pp. 108–116, May 2020.

[19] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[20] Z. Gao, S. Wu, Z. Liu, J. Luo, H. Zhang, M. Gong, and S. Li, "Learning the implicit strain reconstruction in ultrasound elastography using privileged information," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101534.

[21] Z. Gao, J. Chung, M. Abdelrazek, S. Leung, W. K. Hau, Z. Xian, H. Zhang, and S. Li, "Privileged modality distillation for vessel border detection in intracoronary imaging," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1524–1534, May 2020.

[22] Z. Gao, X. Wang, S. Sun, D. Wu, J. Bai, Y. Yin, X. Liu, H. Zhang, and V. H. C. de Albuquerque, "Learning physical properties in complex visual scenes: An intelligent machine for perceiving blood flow dynamics from static CT angiography imaging," *Neural Netw.*, vol. 123, pp. 82–93, Mar. 2020.

[23] S. F. Qadri, D. Ai, G. Hu, M. Ahmad, Y. Huang, Y. Wang, and J. Yang, "Automatic deep feature learning via patch-based deep belief network for vertebrae segmentation in CT images," *Appl. Sci.*, vol. 9, no. 1, p. 69, 2019.

[24] M. Ahmad, D. Ai, G. Xie, S. F. Qadri, H. Song, Y. Huang, Y. Wang, and J. Yang, "Deep belief network modeling for automatic liver segmentation," *IEEE Access*, vol. 7, pp. 20585–20595, 2019.

[25] N. Nida, A. Irtaza, A. Javed, M. H. Yousaf, and M. T. Mahmood, "Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy c-means clustering," *Int. J. Med. Informat.*, vol. 124, pp. 37–48, Apr. 2019.

[26] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[27] C. Cui, K. Thurnhofer-Hemsi, R. Soroushmehr, A. Mishra, J. Gryak, E. Dominguez, K. Najarian, and E. Lopez-Rubio, "Diabetic wound segmentation using convolutional neural networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 1002–1005.

[28] K. Thurnhofer-Hemsi and E. Domínguez, "Analyzing digital image by deep learning for melanoma diagnosis," in *Advances in Computational Intelligence*, I. Rojas, G. Joya, and A. Catala, Eds. Cham, Switzerland: Springer, 2019, pp. 270–279.

[29] K. Thurnhofer-Hemsi and E. Domínguez, "A convolutional neural network framework for accurate skin cancer detection," *Neural Process. Lett.*, pp. 1–21, Oct. 2020, doi: 10.1007/s11063-020-10364-y.

[30] 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.

[31] K. Thurnhofer-Hemsi, E. Lopez-Rubio, N. Roe-Vellve, E. Dominguez, and M. A. Molina-Cabello, "Super-resolution of 3D magnetic resonance images by random shifting and convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[32] Z. Yu, X. Jiang, T. Wang, and B. Lei, "Aggregating deep convolutional features for melanoma recognition in dermoscopy images," in *Machine Learning in Medical Imaging*, Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, Eds. Cham, Switzerland: Springer, 2017, pp. 238–246.

[33] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data augmentation for skin lesion analysis," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, D. Stoyanov, Z. Taylor, D. Sarikaya, Eds. Cham, Switzerland: Springer, 2018, pp. 303–311.

[34] A. Romero-Lopez, X. Giro-i-Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *Proc. Biomed. Eng.*, Feb. 2017, pp. 49–54.

[35] H. Xu, L. Jin, T. Shen, and F. Huang, "Skin cancer diagnosis based on improved multiattention convolutional neural network," in *Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2021, pp. 761–765.

[36] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, 2018.

[37] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinge, "Skin lesion classification using hybrid deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1229–1233.

[38] A. Singhal, R. Shukla, P. K. Kankar, S. Dubey, S. Singh, and R. B. Pachori, "Comparing the capabilities of transfer learning models to detect skin lesion in humans," *Proc. Inst. Mech. Eng., H, J. Eng. Med.*, vol. 234, no. 10, pp. 1083–1093, Oct. 2020.

[39] M. A. Kadampur and S. Al Riyaee, "Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images," *Informat. Med. Unlocked*, vol. 18, Jan. 2020, Art. no. 100282.

[40] A. Mahbod, G. Schaefer, C. Wang, R. 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.

[41] 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)*, Jun. 2015, pp. 1–9.

[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[43] 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, Dec. 2018, Art. no. 180161.

KARL THURNHOFER-HEMSI was born in 1990. He received the dual B.Sc. (computer engineering) and M.Sc. (mathematics) degree and the Ph.D. degree from the University of Málaga, Spain, in 2014 and 2021. In 2016, he joined the Department of Computer Languages and Computer Science, University of Málaga. He is currently a Postdoctoral Researcher with the University of Málaga. His research interests include medical image analysis, pattern recognition, and image processing.

EZEQUIEL LÓPEZ-RUBIO was born in 1976. He received the M.Sc. and Ph.D. (Hons.) degrees in computer engineering from the University of Málaga, in 1999 and 2002, respectively. In 2000, he joined the Department of Computer Languages and Computer Science, University of Málaga, where he is currently a Professor of computer science and artificial intelligence. His technical interests include deep learning, pattern recognition, image processing, and computer vision.
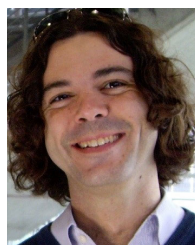
ENRIQUE DOMÍNGUEZ was born in 1975. He received the B.Sc., M.Sc., and Ph.D. degree in computer science from the University of Málaga, in 1999, 2000, and 2007, respectively. In 2000, he joined the Department of Computer Languages and Computer Science, University of Málaga, where he is currently an Associate Professor. He has participated as a member of the program committee of prestigious international conferences and journals. His research interests include neurocomputation, optimization, image/video processing, and computer vision.

DAVID A. ELIZONDO (Senior Member, IEEE) received the B.A. degree in computer science from the Knox College, Galesbourg, IL, USA, the M.S. degree in artificial intelligence from the Department of Artificial Intelligence and Cognitive Computing, University of Georgia, Athens, GA, USA, and the Ph.D. degree in computer science from the University of Strasbourg, France, in cooperation with Swiss Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP). He is currently a Professor of intelligent transport systems with the Department of Computer Technology, De Montfort University, U.K.

• • •