Neural Processing Letters
https://doi.org/10.1007/s11063-020-10364-y

A Convolutional Neural Network Framework for Accurate Skin Cancer Detection

Karl Thurnhofer-Hemsi[1,2] · **Enrique Domínguez**[1,2]

Accepted: 3 October 2020

Abstract
Skin diseases have become a challenge in medical diagnosis due to visual similarities. Although melanoma is the best-known type of skin cancer, there are other pathologies that are the cause of many death in recent years. The lack of large datasets is one of the main difficulties to develop a reliable automatic classification system. This paper presents a deep learning framework for skin cancer detection. Transfer learning was applied to five state-of-art convolutional neural networks to create both a plain and a hierarchical (with 2 levels) classifiers that are capable to distinguish between seven types of moles. The HAM10000 dataset, a large collection of dermatoscopic images, were used for experiments, with the help of data augmentation techniques to improve performance. Results demonstrate that the DenseNet201 network is suitable for this task, achieving high classification accuracies and F-measures with lower false negatives. The plain model performed better than the 2-levels model, although the first level, i.e. a binary classification, between nevi and non-nevi yielded the best outcomes.

Keywords Image processing · Deep learning · Classification · Skin cancer · Melanoma

1 Introduction

Skin alterations are caused due to multiple factors, like allergies, infections, exposition to the sun, etc. The last one is a common practice of most people, who looks for a tan of their skin. However, this search for beauty can have a negative effect on the appearance of skin lesions. This is a typical example of one of the reasons for skin cancer.

✉ Karl Thurnhofer-Hemsi
karlkhader@lcc.uma.es

Enrique Domínguez
enriqued@lcc.uma.es

[1] Department of Computer Languages and Computer Sciences, University of Málaga, Boulevar Louis Pasteur, 35, 29071 Málaga, Spain

[2] Biomedical Research Institute of Málaga (IBIMA), C/ Doctor Miguel Díaz Recio, 28, 29010 Málaga, Spain

Published online: 13 October 2020

Melanoma and non-melanoma skin cancer are highly present in Caucasians. The most common non-melanoma affections are basal cell carcinoma and squamous cell carcinoma. There were more than one million cases in 2018, being the 5th most common cancer. On the other side, melanoma is a less occurring cancer (in the 19th position), with around three hundred thousand new cases last year. Despite the lower number of detections, melanoma causes most of the mortality cases within the skin cancer area.

Melanoma is caused by an abnormal multiplication of melanocytes, the cells that produce pigment and give color to the skin. The sooner the melanoma is detected, the greater the chances of cure. Nevertheless, it could spread to other parts of the body if it is not detected early [1], causing an irremediable effect. The problem resides in the capacity of the detection of melanomas, as they are similar in characteristics to other benign nevi [12]. Dermatologists find hard to distinguish between a benign and a malign mole, being a challenge to find an appropriate rule to classify them.

A common methodology to detect melanomas is the use of the ABCDE rule [19]: asymmetry, borders, color, diameter, and evolving (Fig. 1). These are the warning signs that are monitored in order to diagnose melanoma. High levels of asymmetry or border irregularities are the first alert sign, as well as a strange color of the mole and more than 6 mm diameter. All these signs are monitored to analyze their evolution along the time. The more changes, the more probability to be a malignant mole.

Nowadays, physicians use their experience to analyze and diagnose the presence of skin melanoma by using the rule above mentioned. This methodology might be imprecise, and subject to measurement errors. This motivates to provide a lesion classification system that can support specialists in their clinical procedure, providing an additional accurate diagnosis of the lesion. Moreover, the correct establishment of the type of lesion is important to dispense the adequate treatment to the patient. On the other hand, the creation of an automatic tool that could be installed on any computer or mobile device is of interest for hospitals, physicians in underdeveloped environments and researchers, so they can evaluate patients in a cheap and fast way.

Most of the automatic classification systems in medical imaging have suffered the problem of data availability, provoking an insufficient capability of 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

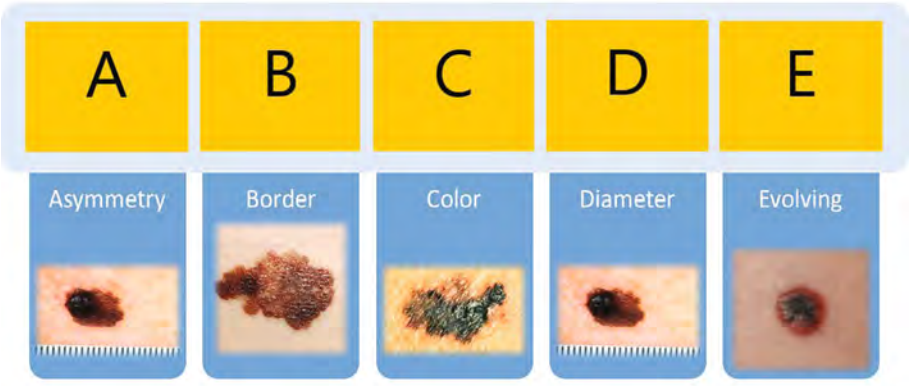


Fig. 1 Traditional clinical analysis followed by dermatologists (ABCDE rule)

it necessary to carry out several preprocessing steps [2] and segment the region of interest [10,11]. Moreover, another commonly used technique is the extraction of features that are used then to improve the classification rate [4,35]. The use of specific features extracted from the melanoma images was widely used to develop classification models [3,25,34], 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 segmentation of the skin lesion is carried out. Until a few years, the classical workflow was the use of these traditional techniques [22], 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 distinguish among classes of images.

Deep learning has been applied to resolve very complex classification and segmentation tasks [24,28] without the use of any image preprocessing method. The architecture of these networks is mostly based on convolutional layers, which filter and extract essential features of the images in order to learn the different lesions. For example, Zhou et al. [37] use 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 [14,16]. Nowadays, CNNs use the power of GPUs to compute a big amount of operations in a few seconds, allowing them to process large datasets to create a reliable model to be applied in image classification, recommender systems and object recognition and segmentation. Focusing on medical imaging, with the increase of publicly available datasets the deep networks has shown a great performance in medical image analysis [15]. Gao et al. [7] used neural networks fed with extra privileged information to carry out strain reconstruction in ultrasound elastography. Deep learning models have also used to detect vessel borders [5] and perceive blood flow from angiographies [6]. Specifically, recent works for skin lesion classification [20,36] have been published, although still there is a margin of improvement. These works are based on a two-stage process, so they can segment and extract features with deep networks and then make the prediction. Moreover, most of them focus on the two-classes problematic, and different types of skin pathologies are usually grouped into the same class and not classified.

The aim of our work is to implement an automatic classifier of several classes of moles without needing human intervention in the prediction. An end-to-end system is proposed to classify skin diseases with the use of deep neural networks, which will be already trained for its use by the user without requiring any parameter tuning. This paper evaluates the performance of the state-of-art pre-trained deep networks for melanoma detection by applying transfer learning. In this sense, the well-known HAM10000 dataset [33], which has been widely used for the benchmark and training of dermatologists, was selected for the experiments. This dataset contains more than 10000 images spread between 7 different classes (see Fig. 2) and more than 50% of the images belong to the nevi class, which makes harder the classification task. Our proposal aims to deal with the data imbalance to clearly differentiate between the 7 classes.

The main challenge of this study is to generate an efficient classification model by using few images and dealing with unbalanced classes (nevi is highly the most prominent class). For that purpose, in addition to the usual plain classifier composed by only one deep network, an hierarchical classifier based on two levels (and two neural models) is proposed, where the first neural network at the first level separates the nevi class from the rest, and the second one classifies the other six classes. Additionally, finding adequate data augmentation techniques

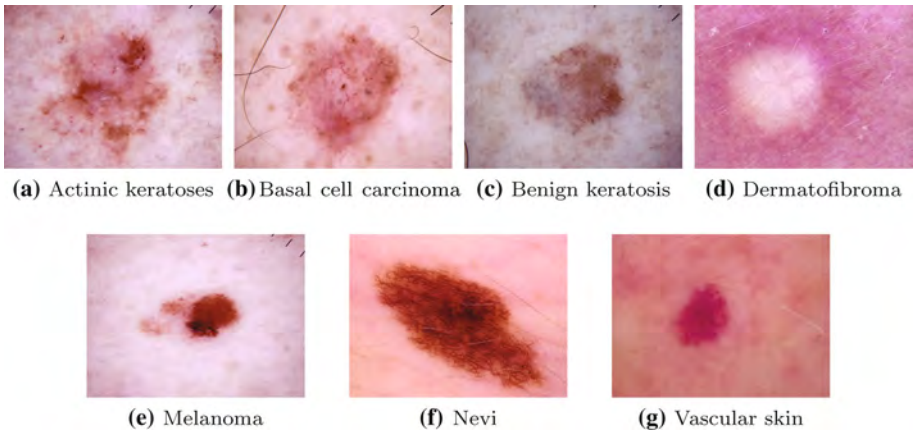


Fig. 2 Samples of the seven different classes of images

is also required to achieve the best results. Finally, overfitting also becomes a difficulty to overcome due to the different distribution of the data. Consequently, this works includes the following contributions: (i) the study of the applicability of transfer learning for skin lesions image classification; (ii) the implementation of a two-stage classifier to deal with class imbalance; (iii) a reliable study of the performance for the distinction among 7 types of lesions.

The rest of the paper is organized as follows: Sect. 2 summarizes the state-of-art convolutional neural networks oriented to object classification. Section 3 explains the proposed methodologies to carry out the classification of melanomas. Experiments and results are presented in Sect. 4, while discussions are made in Sect. 5. In the end, Sect. 6 extract the main conclusions of the work.

2 Deep Classification Networks

Convolutional neural networks have become an essential tool for object recognition, classification, and tracking, and benefited by the increase of data the performance, it has improved enough to create automatic non-assisted systems that are used in many fields, as video-surveillance or autonomous driving. In the medical imaging area, there are many tasks carried out by radiologists and doctors that need to be helped to improve their diagnostics due to the technical limitations of the images. The aim of this work is to use the power of deep learning with images to assist dermatologists in the detection and classification of melanoma.

The main assumption of machine learning models is that data has to have common features and similar distribution. Thus, in the case of the appearance of heterogeneity, deep learning models suffer and need to be adapted and retrained from scratch with new extra datasets. However, in most applications, this procedure in non-viable due to the lack of resources such us image availability or enough budget to carry out the expenses. In those cases, a well-known technique called transfer learning is appropriate, allowing to re-train an already good model to adapt it to a specific problem.

If the data we want to classify is heterogeneous, i.e., the classes are clearly differentiable, then the use of transfer learning may be suitable. There are a large number of classification

models trained on huge datasets in order to classify objects. They have been trained on a similar problem for hundreds of iterations in order to achieve good accuracy. Thus, some of these networks are easily adaptable to similar situations such as medical image classification. In this work, we analyze the performance of six widely known deep networks, which have proved a good performance using transfer learning techniques.

GoogLeNet [29] was presented at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 classification and detection challenges, also known as Inception. The main difference with respect to AlexNet, one of the first classification networks, is the deepness of the network, with a total of 22 layers. They increased the complexity of the network using more neurons at each level as well. It is composed of modules named as Inception, including a set of convolution layers and a max pooling, using a ReLU activation. The input is an RGB image of size 224×224 and the output is a probability vector of 1000 classes.

InceptionV3 [31] is an evolution of the original Inception, where the 7×7 convolution kernels are transformed into 3×3 including new extra convolutions. Moreover, the size of the filters in the inception modules was halved while the number of them was increased considerably. There are more inception modules than the previous version. In summary, this network has a total of 42 layers and the computation cost is about 2.5 times than GoogLeNet. The experiments on the ILSVRC 2012 challenge validation set showed a great improvement with respect to state-of-art methods.

DenseNet201 [8] is a deep network based on a modification of the connections between layers. Whereas in the traditional nets one layer is connected only to the following one, DenseNet connects one layer to all the subsequent layers, i.e. all the preceding feature maps are used as input for the next layers. The network architecture is based on multiple densely connected "dense blocks", including convolutional, max pooling and activation layers as a transition between one block and the following one, for a total of 201 layers. Experiments on ImageNet,[1] CIFAR-10 and 100,[2] and SVHN[3] datasets demonstrated that high performance can be achieved with less computation, due to the reduced number of parameters of the network. Another advantage of this approach is that the network performs feature reuse and propagation, alleviating the vanishing-gradient problem.

Inception-ResNetV2 [30] is a combination of two well known deep network, that tries to take benefit from the residual connections of ResNet accelerating the training of the Inception network. Concretely, more simple Inception blocks than the original one are used, followed by filter-expansion layers in order to scale up the dimensionality produced by the Inception block, and before applying the summations. This convolutional neural network has a total of 164 layers and it has been presented at the 2015 ILSVRC challenge, improving the performance of the ILSVRC 2012 classification task.

MobileNetV2 [27], unlike the previous networks, is a mobile neural network, optimized for constrained environments with small resources, but prepared for multiple tasks and benchmarks. The main novelty of this network is the inverted residual with linear bottleneck, a procedure that eliminates non-linearities and maintains the representational power. The architecture of MobileNetV2 has a total of 53 layers, where the initial layer is a full convolution followed by 19 residual bottleneck layers. A constant expansion rate is used throughout the network. The networks were tested with ImageNet, COCO,[4] and VOC datasets.[5]

[1] http://www.image-net.org.

[2] https://www.cs.toronto.edu/~kriz/cifar.html.

[3] http://ufldl.stanford.edu/housenumbers/.

[4] http://cocodataset.org.

[5] http://host.robots.ox.ac.uk/pascal/VOC/.

The network structures are summarized in Table 1, and the details can be found in the literature. The performance of the pre-trained models is depicted in Fig. 3, where the nets are ranked based on the accuracy versus the number of layers.

3 Proposed Methods

Given an image $\mathbf{X} = \{x_i\}_{i=1}^{N}$, $x_i \in [0, 255]^3$ belonging to one of the C classes, the deep network carries out a set of operations in order to determine on which class $k \in \{1, \ldots, C\}$, the image falls. These operations can be represented as a function

$$\mathcal{F}(\mathbf{X}, W) = \underset{k}{\operatorname{argmax}}\, \mathbf{y}_k \tag{1}$$

where W represents the parameters of the trained neural network and $\mathbf{y}_k$ are the class probabilities produced by the net.

One of the problems in the training stage of any classification model is the presence of unbalanced classes in the dataset. In order to analyze and deal with these situations, two different prediction methodologies are proposed for the above mentioned deep learning models:

1. *Plain classifier*: a neural model is used, whose weights are tuned by applying transfer learning. This method directly classifies the input image into one of the 7 classes. Here the impact of unbalanced classes may be high, being necessary to apply a preprocessing stage to improve its performance. The operation of the plain classifier can be expressed as:

$$\mathcal{F}^P(\mathbf{X}, W^P) = \underset{k}{\operatorname{argmax}}\, \mathbf{y}_k,\ k \in \{1, \ldots, 7\} \tag{2}$$

   where W^P represents the parameters of the fine-tuned network for the seven types of skin lesions.
2. *Hierarchical classifier*: this approach is composed of two neural models, assembling a 2-level classifier. The first level is trained to distinguish the nevi class from the rest, and the second one classifies the other 6 classes.

$$\mathcal{F}_1^H(\mathbf{X}, W_1^H) = \operatorname{argmax}\{\mathbf{y}_1, \bar{\mathbf{y}}_1\} \tag{3}$$

$$\mathcal{F}_2^H(\mathbf{X}, W_2^H) = \underset{k}{\operatorname{argmax}}\, \mathbf{y}_k,\ k \in \{2, \ldots, 7\} \tag{4}$$

   where W_i^H represents the fine-tuned model for two and six classes respectively. The first neural model becomes a specialist model acting as a discriminator for the nevi class, which contains most of the images of the dataset. The second classifier is only applied if $\mathcal{F}_1^H(\mathbf{X}, W_1^H)$ does not output the nevi class.

Figure 4 schematize the two proposals. A single image is provided to both classifiers and the output has to be one of the types of moles. In the case of the hierarchical methods, if the first level network determines that the predicted label is a nevus then it stops. Otherwise, the image is inputted into the second level classifier.

Another typical requirement of deep neural networks is the need for large datasets in order to properly train a model. Otherwise, the trained network may suffer from a lack of generalization, i.e. an over-fitting might appear in the final model. A common technique to reduce this effect is data augmentation, which has been applied with success in many

Table 1 Structures of the pre-trained deep networks

GoogLeNet		InceptionV3		DenseNet201		Inception-ResNetV2		MobileNetV2	
Layer	Output	Layer	Output	Layer	Output	Layer	Output	Layer	Output
Input	224 × 224	Input	229 × 229	Input	224 × 224	Input	299 × 299	Input	224 × 224
Conv	112 × 112	Conv	149 × 149	Conv	112 × 112	Stem	35 × 35	Conv	112 × 112
Max pool	56 × 56	Conv	147 × 147	Pool	56 × 56	5× IncResA	35 × 35	Bottleneck	112 × 112
Conv	56 × 56	Conv pad	147 × 147	Denseblock 6[conv, conv]	56 × 56	Reduction-A	17 × 17	Bottleneck	56 × 56
Max pool	28 × 28	Max pool	73 × 73	Transition	56 × 56	10× IncResB	17 × 17	Bottleneck	28 × 28
Inception	28 × 28	Conv	71 × 71	Conv, pool	28 × 28	Reduction-A	8 × 8	Bottleneck	14 × 14
Inception	28 × 28	Conv	35 × 35	Denseblock 12[conv, conv]	28 × 28	5×IncResC	8 × 8	Bottleneck	14 × 14
Max pool	14 × 14	Conv	35 × 35	Transition	28 × 28	Avg pool	1 × 1	Bottleneck	7 × 7
Inception	14 × 14	3× Inception	17 × 17	Conv, pool	14 × 14	Dropout	1 × 1	Bottleneck	7 × 7
Inception	14 × 14	5× Inception	8 × 8	Denseblock 48[conv, conv]	14 × 14	Softmax	1 × 1	Conv	7 × 7
Inception	14 × 14	5× Inception	8 × 8	Transition	14 × 14			Avg pool	1 × 1
Inception	14 × 14	Pool	1 × 1	Conv, pool	7 × 7			Conv	1 × 1
Inception	14 × 14	Linear	1 × 1	Denseblock 32[conv, conv]	7 × 7				
Max pool	7 × 7	Softmax	1 × 1	Pool	1 × 1				
Inception	7 × 7			Fully	1 × 1				
Inception	7 × 7								
Avg pool	1 × 1								
Dropout	1 × 1								
Linear	1 × 1								
Softmax	1 × 1								

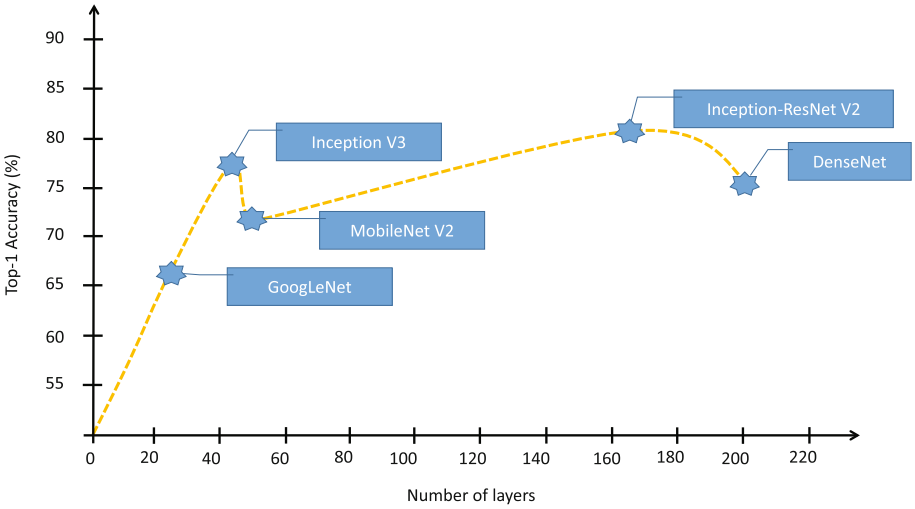


Fig. 3 Performance of the studied deep networks with respect to their number of layers

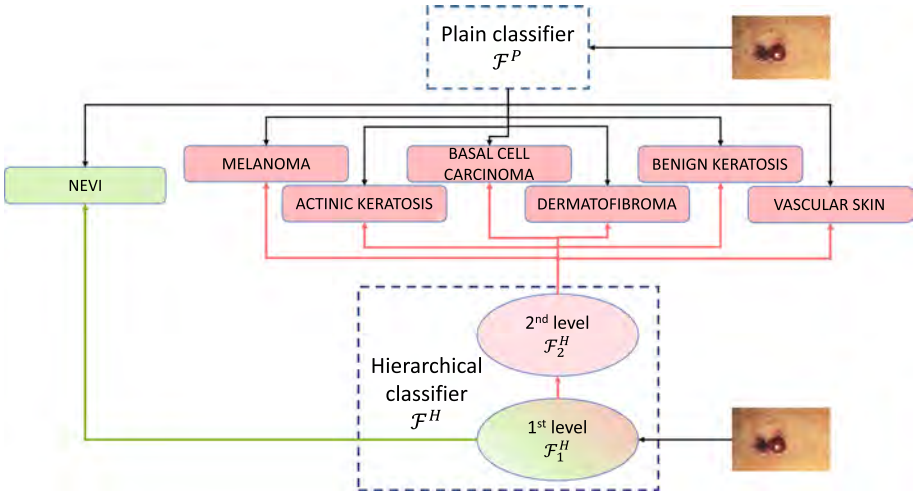
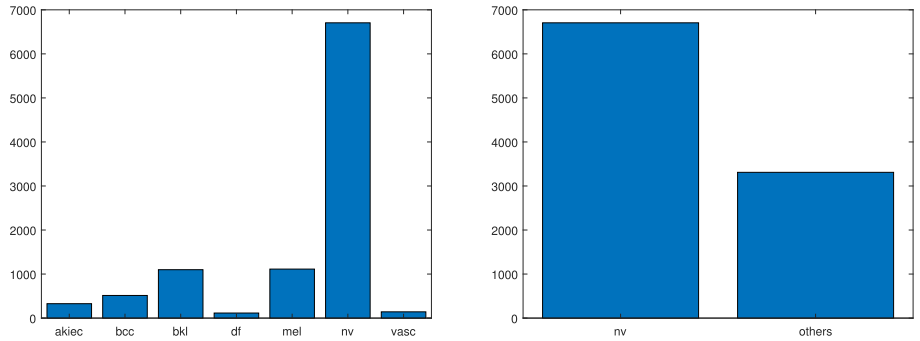


Fig. 4 Scheme of the operation of the proposed classifiers

classification problems, including in the medical field [9]. A robust convolutional neural network can be invariant to translation, viewpoint, size or illumination and this is the premise of data augmentation. There are different kinds of methods that can be applied to enlarge the input dataset, such as geometric deformations, data wrapping or color transformations. The most used are an affine transformation of the original image, like reflections, random crops, and translations of the image.

The idea of a hierarchical model comes up because the HAM10000 dataset is a completely unbalanced set in favor of the nevi class, as it is described in Sect. 4. Although the use of data augmentation might level the training of the model, our aim is to remove the differences between classes by the first CNN and then specializes in the second one on the non-nevi class to achieve better discrimination.



(a) Original distribution **(b)** 2-class distribution

Fig. 5 Class distribution of the HAM10000 dataset

4 Experimental Results

This section summarizes the experiments we carried out in order to evaluate the performance of the proposed models.

4.1 Dataset and Data Augmentation

The publicly available HAM10000 dataset [33] has been utilized for the experimental results and analysis of the proposed approaches. The HAM10000 dataset consists of 10,050 dermoscopic images belonging to seven different classes including actinic keratosis (akiec), basal cell carcinoma (bcc), benign keratosis (bkl), dermatofibroma (df), nevi (nv), melanoma (mel), and vascular skin (vasc). This dataset has been widely used as a benchmark for comparisons of humans and machines, even in several classification challenges. The 10,015 dermoscopic images were collected over 20 years from the department of dermatology at the Medical University of Viena (Austria) and the skin cancer practice of Cliff Rosendahl in Queensland (Australia).

The first important drawback of this dataset is related to the unbalanced distribution of the data, which can be seen in Fig. 5a. There are almost 7000 images belonging to the nevi class, while the others contain no more than 1000. This may provoke the network to specialize in those images with similar characteristics to nevi, like benign keratosis. In addition, there are few dermatofibromas and vascular skin images. Thus, the performance on the test set should be analyzed carefully because the proportion of images is considerably lower than in other classes. Consequently, the use of augmentation techniques is necessary to balance the skewed distribution in the training stage. This data augmentation was done by using different reflections and rotations of the original images, and it was applied during the training process. The selected data augmentation techniques were:

- Horizontal flipping with a probability of 0.5.
- Vertical flipping with another probability of 0.5.
- Image rotations with a probability of 0.75 using a random angle in the range $[-90, 90]$.

However, data augmentation can smooth the effects, but not solve the 7:1 ratio between nevi and other skin lesions. Therefore, the 2-stages model was implemented to first equilibrate the dataset as shown in Fig. 5b, and then to improve the classification of the non-nevi images.

Regarding to the class imbalance effects, the low inter-class variability of some lesions makes difficult their correct predictions by the neural network. Nevi are benign neoplasms of melanocytes with a symmetric distribution of color and structure. In contrast, melanoma is usually chaotic, which makes them easily distinguishable from nevi. Something similar happens with dermatofibroma and vascular skin, presenting a strong color contrast between the center and the outer part. The most problematic lesions are the actinic keratoses, the basal cell carcinoma, and the benign keratosis, which are very similar in color and shape. The classification model might fail more often with these types of images.

4.2 Training and Performance

There are several parameters that can be tuned in a deep neural network. However, the main ones we are focusing on (with their selected values) are given below. This configuration was deduced from our previous work [32] with the same problem, where several different configurations and neural models were tested.

- Batch size (16): indicates the number of images processed in one iteration.
- Initial learning rate (0.0001): establishes with which rate is going to start the learning procedure.
- Validation frequencies (10): is the number of iterations between evaluations of the validation metrics.
- Optimizer (SGDM): the algorithm which updates the weights and biases of the network in order to minimize the loss function. In our experiments, the Standard Gradient Descent algorithm with Momentum of 0.9 was selected.
- Maximum number of epochs (10): the maximum number of times the full dataset is passed to the neural network.

Both the plain and the hierarchical models were previously trained on the ImageNet database and fine-tuning was performed on the HAM10000 dataset in order to transfer their knowledge to the skin diseases classification problem. The dataset was split into three sets in order to carry out the fine-tuning and posterior evaluation: training (70%), validation (20%) and testing (10%). Repeated holdout was employed in order to provide a more reliable evaluation of the classification performance. Therefore, $K = 10$ repetitions of the experiments were executed by randomly splitting the dataset multiple times according to the previous division.

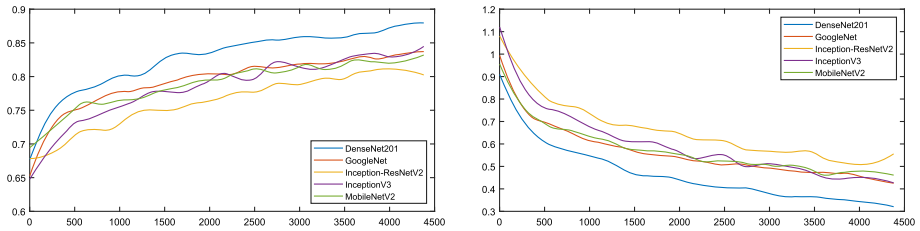
The performance of each model and each deep network was compared with the standard classification measures: Accuracy, Precision, Recall and F-measure. It is well-known that the accuracy measures the proportion of the true detections, while the precision is the proportion of relevant instances among the retrieved instances and recall is the proportion of relevant instances that have been retrieved over the number of relevant instances. The F-measure provides a good overall evaluation of the performance of a given method considering both the precision and the recall. All measures are ranged between 0 and 1, being the higher the better.

The formulation of each measure is written as follow:

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

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

$$Precision = \frac{TP}{TP + FP} \tag{7}$$



(a) Training Accuracy **(b)** Training Loss

Fig. 6 Performance of the addressed models in the training process

Table 2 Accuracy results of the plain classifier

Model	Accuracy		
	Training	Validation	Test
DenseNet201	**0.9618**	**0.8787**	**0.8773**
GoogLeNet	0.8465	0.8093	0.8013
Inception-ResNetV2	0.8669	0.8347	0.8322
InceptionV3	0.8955	0.8487	0.8412
MobileNetV2	0.8928	0.8457	0.8431

Each deep network was tested with the training, validation and test sets
The best performance is shown in bold

$$F\text{-}measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

where TP, TN, FP, and FN denote the true positives, true negatives, false positives, and false negatives, respectively. In addition, the average of the recall, precision, and F-measure among classes will be computed in order to obtain a precise measure that it is not corrupted by the class imbalance. Note that the average recall is the equivalent to the balanced accuracy for multi-class problems.

The training curves for the fine-tuning of the five deep networks studied in this work are presented in Fig. 6. It can be seen that DenseNet201 is clearly the most efficient model achieving the highest accuracy and the lowest loss. InceptionV3 seems to be rather unstable, while Inception-ResNetV2 has the worst training performance reaching a bit of overfitting.

4.3 Results

The results on the training, validation and test sets of the two methods, are shown in Tables 2 and 3. In the case of the plain classifier shown in the former Table, most of the neural networks carried out good training accuracy but DenseNet201 excels with respect to the others achieving 96% of correct classifications. The tendency in the validation and test sets is similar. It is remarkable that DenseNet201 has worsened its performance by 9% with respect to the training set, while Inception-ResNetV2 has only fallen 3% its accuracy.

On the other hand, Table 3 shows the performance of each level of the hierarchical classifier. In the first level (Table 3a) DenseNet201 again yields the best accuracy in the training set, with more than 96%. However, on the validation and test sets, it suffers again of a lack of generalization distinguishing between a nevi or a non-nevi image, and being the InceptionV3

Table 3 Accuracy results of the hierarchical classifiers for each level and network

Model	Accuracy		
	Training	Validation	Test
(a) Level 1			
DenseNet201	**0.9657**	0.8922	**0.8912**
GoogLeNet	0.9045	0.8717	0.8698
Inception-ResNetV2	0.8968	0.8637	0.8610
InceptionV3	0.9393	**0.8937**	0.8911
MobileNetV2	0.9256	0.8717	0.8691
(b) Level 2			
DenseNet201	**0.9252**	**0.7628**	**0.7620**
GoogLeNet	0.7481	0.6692	0.6671
Inception-ResNetV2	0.8082	0.7266	0.7259
InceptionV3	0.6488	0.6269	0.6258
MobileNetV2	0.8187	0.7190	0.7188

Each deep network was tested with the training, validation and test sets
The best performance is shown in bold

network very competitive in this matter. MobileNetV2 also achieves high rates in this first step. Compared with the plain model, the binary classification accuracy rate outperforms the seven-class classification model, which may indicate that the networks distinguish better between benign and malign moles that between classes individually. Thus, the inputs of the second level should be adequate enough to avoid misclassifications.

However, the results of the second level are a bit worse. GoogLeNet and InceptionV3 do not get to have a good convergence in training, yielding bad accuracies ($< 75\%$) not only in the validation and test sets, actually also in the training data. MobileNetV2 and Inception-ResNetV2 perform better with around 81% but far from the desired results. The case of DenseNet201 is very particular because the training accuracy is good enough but its capacity of generalization on the validation and test sets falls 15%. This indicates symptoms of over-fitting which might be caused by a lack of enough images to fine-tune properly these deep networks.

The next three tables sum up the detailed measures of each model and network using the whole HAM10000 dataset. Average values of recall, precision and F-measure among classes were computed. Table 4 presents the results of the first level of the hierarchical classifier, i.e the binary classifier. InceptionV3 and DenseNet201 yield measures greater than 90%, with the exception of recall and F-measure for InceptionV3. The 92% of F-measure obtained by DenseNet201 verifies that this network is very appropriate for the classification of nevi and non-nevi. A lower precision value for Inception-ResNetV2 means that this network provokes many false positives but its high recall indicates that the false negatives are not predominant. In medical applications, the absence of false negatives is crucial to avoid the non-diagnosis of malignant moles. However, MobileNetV2 has the opposite effect, which is not recommended for this task.

Compared with the classifier of seven types of moles showed in Table 5, the performance of DenseNet201 is better in the binary model, being also the best among the other networks in this modality. In this occasion, the second position is for MobileNetV2 since the true positive rate is around 80%, much better than the rest of the models, which have good accuracy values but poor F-measure. These results indicate that transfer learning is more effective when we

Table 4 First level results (hierarchical classifier) for each network

Model	Accuracy	Avg. recall	Avg. precision	Avg. F-measure
DenseNet201	**0.9509**	**0.9157**	0.9346	**0.9251**
GoogLeNet	0.8980	0.8520	0.8413	0.8466
Inception-ResNetV2	0.8902	0.9115	0.7890	0.8458
InceptionV3	0.9302	0.8438	**0.9388**	0.8888
MobileNetV2	0.9148	0.8088	0.9241	0.8626

Accuracy and average recall (i.e balanced accuracy), precision and F-measure among classes measured on the overall dataset are presented
The best performance is shown in bold

Table 5 Results of the plain classifiers for each network

Model	Accuracy	Avg. recall	Avg. precision	Avg. F-measure
DenseNet201	**0.9452**	**0.9050**	**0.9203**	**0.9126**
GoogLeNet	0.8390	0.6523	0.7667	0.7049
Inception-ResNetV2	0.8605	0.6661	0.8035	0.7284
InceptionV3	0.8862	0.7492	0.8657	0.8032
MobileNetV2	0.8834	0.8009	0.8175	0.8091

Accuracy and average Recall (i.e balanced Accuracy), Precision and F-measure measured on the overall dataset are presented
The best performance is shown in bold

Table 6 Results of the hierarchical classifiers for each network

Model	Accuracy	Avg. Recall	Avg. Precision	Avg. F-measure
DenseNet201	**0.9173**	**0.8480**	**0.8530**	**0.8505**
GoogLeNet	0.8234	0.6234	0.7323	0.6735
Inception-ResNetV2	0.8280	0.7086	0.6855	0.6969
InceptionV3	0.8277	0.4992	0.7522	0.6002
MobileNetV2	0.8604	0.6845	0.7818	0.7299

Accuracy and average recall (i.e balanced accuracy), precision and F-measure among classes measured on the overall dataset are presented
The best performance is shown in bold

tried to classify between two classes than with seven, although in both cases the best deep network achieved an accuracy of around 95% and a balanced accuracy of around 91%.

Focusing on the performance of the hierarchical model (Table 6), the best outcomes are produced again by DenseNet201, with 91% accuracy and 85% of F-measure and precision as reference values. The other networks behaved badly, being MobileNetV2 the most close to the winner. However, the results are worst compared to the plain classifier. The results of the first level showed in the previous table are quite good but the final classification of the malignant moles is not good enough. The main reason for this might be the bad accuracies of the second level yielded in Table 3. The same configuration was applied for all networks and more over-fitting is present in DenseNet201.

A detailed analysis can be extracted from the confusion matrices depicted in Figs. 7 and 8. The first figure shows the predictions for the plain classifier. The blue diagonal corresponds to the true positives, where the saturation indicates the percentage of images correctly identified.

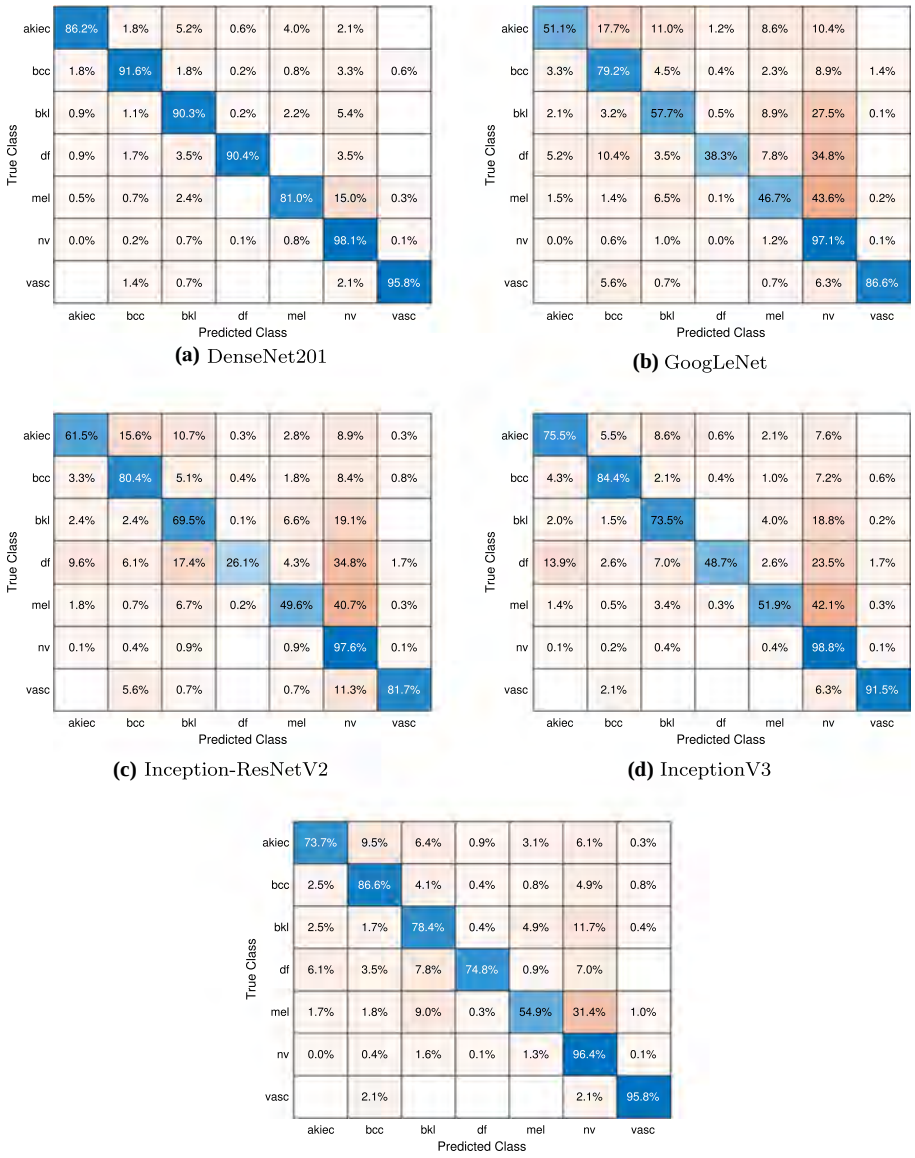


Fig. 7 Confusion matrices of the plain classifiers generated using the whole dataset

Also, note that the balanced accuracy (the average of the diagonal values) is equal to the average recall shown in the previous tables.

The differences among networks are very small for all classes with the exception of the melanoma class (mel). DenseNet201 classifies 81% of melanoma images correctly where the rest of the networks almost do not overpass the 55%. This set of images is usually misclassified as benign keratosis (bkl) and could be the main reason for the accuracy drop.

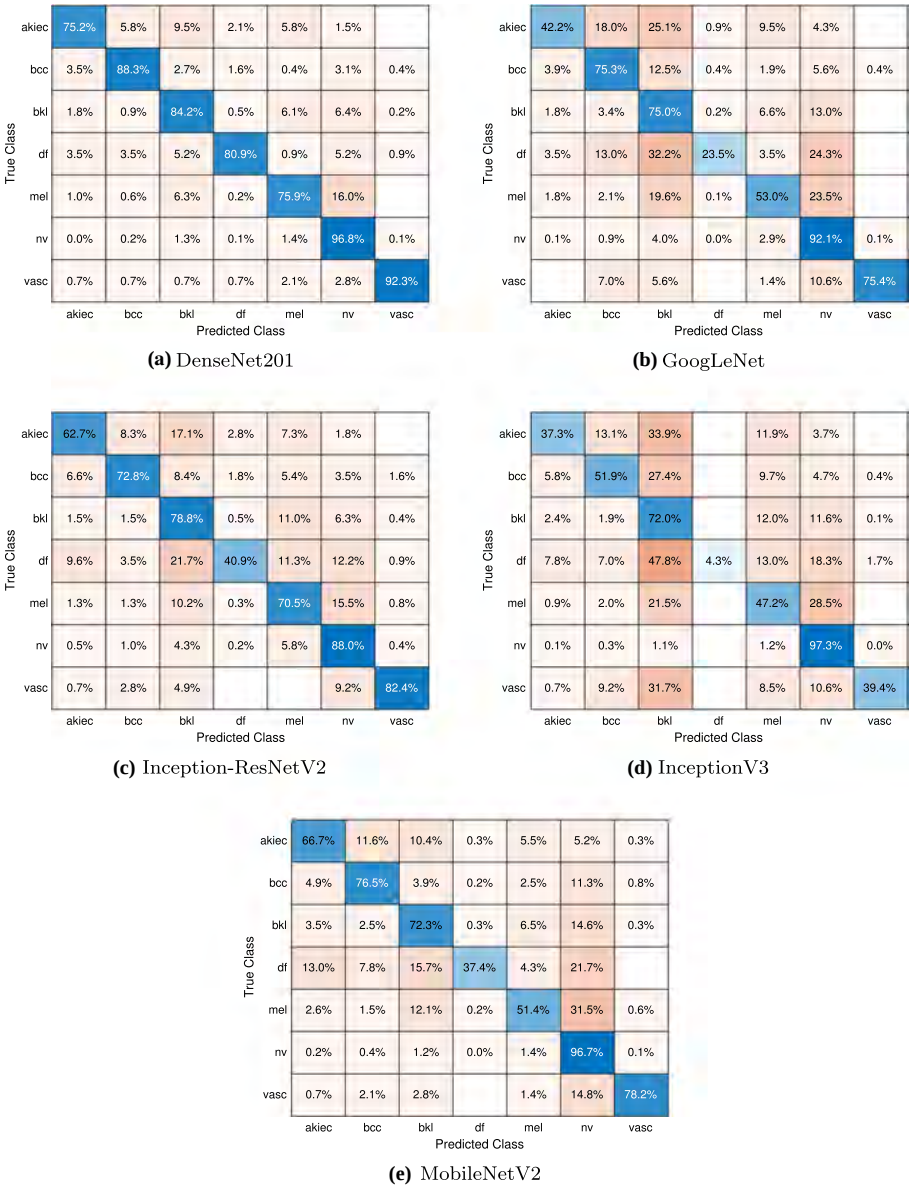
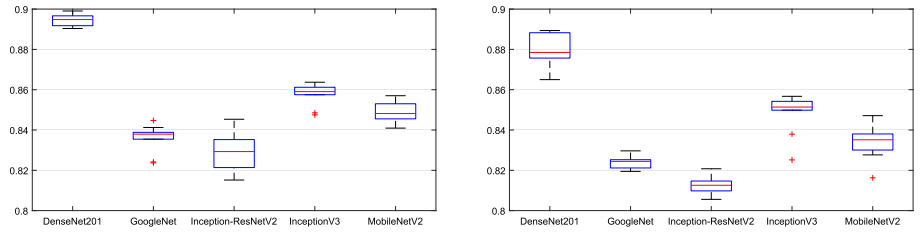


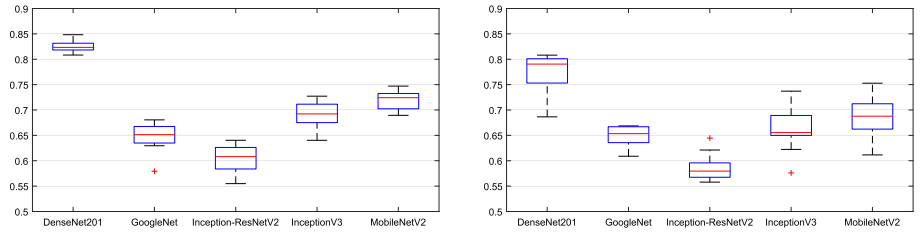
Fig. 8 Confusion matrices of the hierarchical classifiers generated using the whole dataset

On the other hand, the confusion matrices of the hierarchical model presented in Fig. 8 reflect a general spread of the predictions. For example, InceptionV3 is not able to identify dermatofibromas (df) and vascular skin (vas) as good as before, and either GoogLeNet Inception-ResNetV2 and MobileNetV2 confuses the basal cell carcinoma (bcc) with benign keratosis. If we taking into account the image examples presented before in Fig. 2, these inconsistencies are reasonable due to the similarity between classes and the lack of generalization of the second classifier.



(a) Plain classifiers **(b)** Hierarchical classifiers

Fig. 9 Accuracy for 10 repeated holdout cross-validation



(a) Plain classifiers **(b)** Hierarchical classifiers

Fig. 10 Balanced accuracy for 10 repeated holdout cross-validation

In general, the balanced accuracy yielded by the DenseNet201 model is the highest one for both the plain and hierarchical classifiers. In order to have a better quality assessment of the models, repeated holdout cross-validation was carried out and the results are shown in Figs. 9 and 10. Boxplots summarize the 10 repetitions of the classification measures for the whole dataset.

The accuracy outcomes depicted in Fig. 9 showed that the previous results were coherent. There is low dispersion among trainings except for the Inception-ResNetV2 in the plain classifier and DenseNet201 in the hierarchical, although the tendency for both models is very similar. DenseNet201 presents the best outcomes arriving at 90% accuracy, while GoogleNet and Inception-ResNetV2 are always below 85%.

If we focus on the balanced accuracy (see Fig. 10), the results are slightly worse although the behavior of the methods is very similar to global accuracy. InceptionV3 differs a bit yielding worse balanced accuracy than MobileNetV2. 82.5% of balanced accuracy is the best measure obtained, being DenseNet201 the model who is near this performance.

A visual representation of the features learned by the DenseNet201 network is depicted in Fig. 11. Here the main differences are not manifested as edges or shapes, but with color variations. For example, melanoma class (mel) is clearly distinguishable due to the brown and orange specks. However, there few similarities with benign keratosis (bkl) and nevi (nv) classes. This fact can be also seen in the confusion matrices, where the predicted class of a melanoma image is sometimes confused by one of these ones. Something similar occurs with the actinic keratosis (akiec), whose features have not a clear pattern and are confused mostly with the bkl class. On the other hand, vascular skin (vasc) is the most differentiable class since its purple and circular specks are unique. It presents the lowest percentage of confusion.

In Table 7 the measures of each class of the best method (DenseNet201) are detailed. It can be seen that the nevi class (nv) is well identified in both plain and hierarchical models,

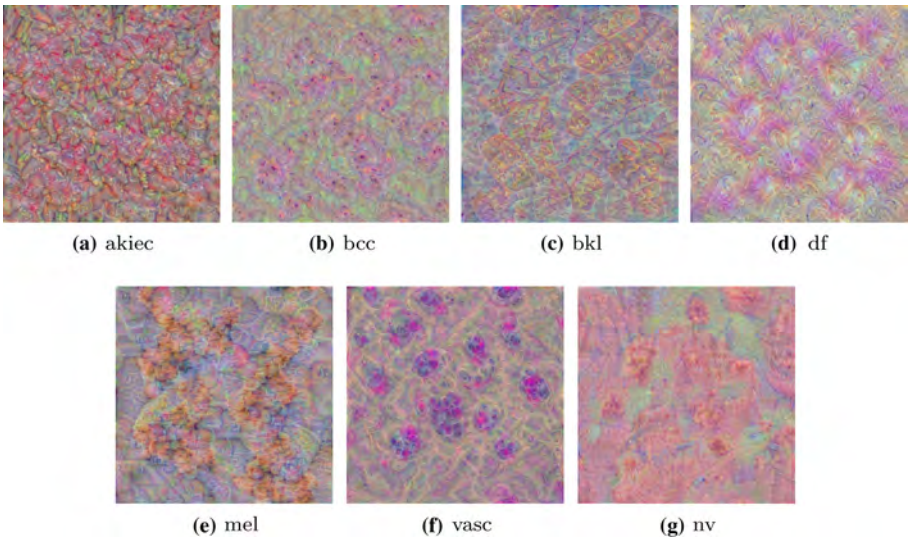


Fig. 11 Feature maps for each class of the final fully convolutional layer of the DenseNet201

Table 7 Results of the Densenet201 model for each class

Class	Plain			Hierarchical		
	Recall	Precision	F-measure	Recall	Precision	F-measure
akiec	0.8624	0.9126	0.8868	0.7523	0.8119	0.7810
bcc	0.9163	0.9128	0.9146	0.8833	0.8919	0.8876
bkl	0.9026	0.9059	0.9043	0.8417	0.8171	0.8292
df	0.9043	0.9204	0.9123	0.8087	0.7440	0.7750
mel	0.8104	0.9029	0.8542	0.7592	0.8180	0.7875
nv	0.9812	0.9624	0.9717	0.9684	0.9588	0.9636
vasc	0.9577	0.9252	0.9412	0.9225	0.9291	0.9258

due to the presence of a large number of images within this class. The vascular skin (vas) is also recognized with more than 92% of F-measure because the net detected strong features in this type of image. The main problem of the 2-level classifier resides in the misclassification of actinic keratosis, dermatofibroma and melanoma.

4.4 Times Comparison

In addition to the classification performance of the tested models, a comparison in terms of computational time was carried out. Training and testing times were measured and the results are summarized in Table 8, where the mean and standard deviation of the measurements among the 10 repetitions is presented.

The training stage takes around 8 h for DenseNet201 and Inception-ResNetV2 for both the plain and hierarchical classifiers, being the slowest models. It is important to remark that this training is a fine-tuning of a pre-trained network, which needed several days to be trained.

Table 8 CPU times comparison between the tested models

Model	Plain		Hierarchical	
	Fine-tuning (h)	Testing (s)	Fine-tuning (h)	Testing (s)
DenseNet201	7.311 ± 0.076	$0.017 \pm 1.279e{-5}$	9.084 ± 0.100	$0.025 \pm 5.733e{-4}$
GoogLeNet	1.113 ± 0.008	$0.003 \pm 1.277e{-5}$	1.298 ± 0.006	$0.017 \pm 6.531e{-5}$
Inception-ResNetV2	7.519 ± 0.094	$0.016 \pm 5.092e{-5}$	9.256 ± 0.141	$0.017 \pm 5.417e{-4}$
InceptionV3	2.590 ± 0.004	$0.007 \pm 1.013e{-5}$	3.094 ± 0.007	$0.017 \pm 3.182e{-4}$
MobileNetV2	1.635 ± 0.010	$0.006 \pm 1.755e{-5}$	1.950 ± 0.015	$0.017 \pm 2.760e{-4}$

Mean and standard deviations were computed. The fine-tuning columns present to the overall training stage in hours, while the testing columns present the time needed to test one image in seconds

Table 9 Comparison with other state-of-art methods using HAM10000 dataset for classification of seven classes of skin lesions

Approach	Best results
Ours	Accuracy in training: **0.961**, in validation: 0.878, in testing: **0.877**
Nugroho et al. [21]	Accuracy in training: 0.800, in testing: 0.780
Shahin et al. [28]	Accuracy in validation **0.899**
Khan et al. [13]	Accuracy in testing: 0.898
Moldovan [18]	Accuracy 1st level: 0.850, 2nd level: 0.750
Mobiny et al. [17]	Accuracy in testing: 0.8359 ± 0.170
Sae-Lim et al. [26]	Accuracy: 0.832, Recall: 0.850, F-measure: 0.820
Pai and Giridharan [23]	Accuracy in testing: 0.780

The best performance is shown in bold

Thus, we are reducing the time considerably. Nevertheless, the testing of an image is quite fast for all methods, and much faster that traditional methods than usually need to extract features before making a prediction.

On the other hand, GoogLeNet is the fastest model for training and testing, followed by MobileNetV2. While GoogLeNet does not achieve a good classification performance, MobileNetV2 is suitable for its use in skin lesion diagnostic in a fast way and integrated into low-cost hardware.

5 Discussion

From the previous section, it can be concluded that it is possible to achieve a good classification performance with an adequate combination of fine-tuning and data augmentation. Nevertheless, this work would be useless if we do not make a comparison of the performance of our proposal with the state-of-art. For that purpose, we have searched the recently published works that employ the same HAM10000 dataset for their work, and their best results were summarized in Table 9.

Nugroho et al. [21] presented a custom convolutional neural network with only 9 layers, which is very efficient in terms of resource usage, although the classification accuracy is only 80%. Shahin et al. [28] combined the outputs of the well-known ResNet50 and the InceptionV3 with an ensemble technique to obtain 89.9% accuracy on the validation set. A

similar result was obtained by Khan et al. [13] on the test set, where first the features are extracted by a multi-modal CNN and then pass through a support vector machine with an RBF kernel. Moldovan [18] developed a 2-step classification method using DenseNet121, dividing into three and four classes. The first one yielded an accuracy of 85% and the second 75%. Bayesian DenseNet169 was used by Mobiny et al. [17] to obtain an accuracy of 84% on the test set. while a modified version of MobileNet implemented by Sae-Lim et al. [26] performed a few worse. Finally, Pai and Giridharan [23] created a website application based on VGGNet to provide 78% accuracy for testing. Therefore, we can say that our work overcomes some of the recent methods based on deep learning for skin lesion classification.

Comparing the plain and the hierarchical classifiers, we found the first one more appropriate for this task. The data augmentation techniques worked better when the class imbalance is pronounced. The hierarchical classifier provides an effective way to discriminate between nevi and non-nevi moles with 96% accuracy. More work should be done at the second level to distinguish the rest of the lesions.

The main limitation of the proposed method is its dependency on the type of images contained within each class. The intra-class variance could be difficult to overcome because each image would need a specific preprocessing. The basis of deep learning is that it works well with big datasets but with low variance in each class. Another drawback we encountered for the hierarchical classifier is that the second level was not able to distinguish adequately the non-nevi images. This may indicate similarities between classes that should be considered to avoid over-fitting.

6 Conclusions

We have presented two frameworks based on transfer learning for computer-aided diagnosis of malignant moles in the skin. In addition to the nevi and melanoma classes, the usual classification, we tried to classify the other five types of skin diseases. Five well-known convolutional neural networks were fine-tuned using the HAM10000 dataset, and used in two different proposed frameworks: a plain model and a hierarchical model with 2 levels, the first level deals with the distinguishing between nevi and non-nevi images and the second one to classify the malign moles (non-nevi).

Experiments showed that the best deep network is DenseNet201, being around 10% better than the rest of the networks in all measures, and specifically in recall that indicates a low level of false negative detections, which is essential in medical diagnosis. The plain model behaved better for both binary (first level of the hierarchical model) and seven classes classification, with almost a 95% accuracy and 92% of F-measure on the whole HAM10000 dataset. The imbalance of the dataset and the absence of enough images, despite the use of data augmentation, are the reason for the lower generalization of DenseNet201 in the second level classifier, dragging its results to the complete model. Note that the same configuration was used for each deep network in order to compare the performance of all neural networks with the same parameter values.

Further works will be focused on the testing of more deep networks and other hierarchies, including preprocessing steps in order to distinguish better between the six non-nevi classes. A specific analysis of the features of each class is a necessary task to generate an adequate classifier. The inclusion of probabilistic techniques to make an accurate prediction of different classifiers is another research line.

Acknowledgements This work is partially supported by the Ministry of Economy and Competitiveness of Spain under Grants TIN2016-75097-P and PPIT.UMA.B1.2017. It is also partially supported by the Ministry of Science, Innovation and Universities of Spain under Grant RTI2018-094645-B-I00, project name Automated detection with low-cost hardware of unusual activities in video sequences. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project UMA18-FEDERJA-084, project name Detection of anomalous behavior agents by deep learning in low-cost video surveillance intelligent systems. All of them include funds from the European Regional Development Fund (ERDF). The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X GPUs used for this research. The authors acknowledge the funding from the Universidad de Málaga. Karl Thurnhofer-Hemsi (FPU15/06512) is funded by a PhD scholarship from the Spanish Ministry of Education, Culture and Sport under the FPU program.

References

1. American Cancer Society I (ed) (2016) Cancer facts & figures. American Cancer Society, Atlanta
2. Asha Gnana Priya H, Anitha J, Poonima Jacinth J (2018) Identification of melanoma in dermoscopy images using image processing algorithms. In: 2018 international conference on control, power, communication and computing technologies, ICCPCCT 2018, pp 553–557
3. Bakheet S (2017) An SVM framework for malignant melanoma detection based on optimized HOG features. Computation 5(1):1–13
4. Devassy B, Yildirim-Yayilgan S, Hardeberg J (2019) 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 868:150–159
5. Gao Z et al (2019) Privileged modality distillation for vessel border detection in intracoronary imaging. IEEE Trans Med Imaging 39(5):1524–1534
6. Gao Z, Wang X, Sun S, Wu D, Bai J, Yin Y, Liu X, Zhang H, de Albuquerque VHC (2020) Learning physical properties in complex visual scenes: an intelligent machine for perceiving blood flow dynamics from static CT angiography imaging. Neural Netw 123:82–93
7. Gao Z, Wu S, Liu Z, Luo J, Zhang H, Gong M, Li S (2019) Learning the implicit strain reconstruction in ultrasound elastography using privileged information. Med Image Anal 58:101534
8. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
9. Hussain Z, Gimenez F, Yi D, Rubin D (2017) Differential data augmentation techniques for medical imaging classification tasks. In: AMIA annual symposium proceedings, vol 2017. American Medical Informatics Association, p 979
10. Jafari MH, Karimi N, Nasr-Esfahani E, Samavi S, Soroushmehr SMR, Ward K, Najarian K (2016) Skin lesion segmentation in clinical images using deep learning. In: 2016 23rd international conference on pattern recognition (ICPR), pp 337–342
11. Jafari MH, Nasr-Esfahani E, Karimi N, Soroushmehr SMR, Samavi S, Najarian K (2017) Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma. Int J Comput Assist Radiol Surg 12(6):1021–1030
12. Jerant AF, Johnson JT, Sheridan C, Caffrey TJ (2000) Early detection and treatment of skin cancer. Am Fam Phys 62(2):357–368, 375–376, 381–382
13. Khan MA, Javed MY, Sharif M, Saba T, Rehman A (2019) Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification. In: 2019 international conference on computer and information sciences (ICCIS). IEEE, pp 1–7
14. Li J, Zhou G, Qiu Y, Wang Y, Zhang Y, Xie S (2019) Deep graph regularized non-negative matrix factorization for multi-view clustering. Neurocomputing 390:108–116
15. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
16. Liu N, Wan L, Zhang Y, Zhou T, Huo H, Fang T (2018) Exploiting convolutional neural networks with deeply local description for remote sensing image classification. IEEE Access 6:11215–11228
17. Mobiny A, Singh A, Van Nguyen H (2019) Risk-aware machine learning classifier for skin lesion diagnosis. J Clin Med 8(8):1241
18. Moldovan D (2019) Transfer learning based method for two-step skin cancer images classification. In: 2019 E-health and bioengineering conference (EHB), pp 1–4

19. Nachbar F, Stolz W, Merkle T, Cognetta AB, Vogt T, Landthaler M, Bilek P, B-Falco O, Plewig G (1994) The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. J Am Acad Dermatol 30(4):551–559
20. Nida N, Irtaza A, Javed A, Yousaf M, Mahmood M (2019) Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering. Int J Med Inf 124:37–48
21. Nugroho AA, Slamet I, Sugiyanto (2019) Skins cancer identification system of HAMl0000 skin cancer dataset using convolutional neural network. AIP Conf Proc 2202(1):020039
22. Oliveira RB, Papa JP, Pereira AS, Tavares JMR (2018) Computational methods for pigmented skin lesion classification in images: review and future trends. Neural Comput Appl 29(3):613–636
23. Pai K, Giridharan A (2019) Convolutional neural networks for classifying skin lesions. In: TENCON 2019—2019 IEEE region 10 conference (TENCON). IEEE, pp 1794–1796
24. Pereira dos Santos F, Antonelli Ponti M (2018) Robust feature spaces from pre-trained deep network layers for skin lesion classification. In: 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, pp 189–196
25. Ruela M, Barata C, Marques J, Rozeira J (2017) A system for the detection of melanomas in dermoscopy images using shape and symmetry features. Comput Methods Biomech Biomed Eng: Imaging Vis 5(2):127–137
26. Sae-Lim W, Wettayaprasit W, Aiyarak P (2019) Convolutional neural networks using mobileNet for skin lesion classification. In: 2019 16th international joint conference on computer science and software engineering (JCSSE), pp 242–247
27. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
28. Shahin AH, Kamal A, Elattar MA (2018) Deep ensemble learning for skin lesion classification from dermoscopic images. In: 2018 9th Cairo international biomedical engineering conference (CIBEC). IEEE, pp 150–153
29. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 1–9
30. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: 31st AAAI conference on artificial intelligence
31. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
32. Thurnhofer-Hemsi K, Domínguez E (2019) Analyzing digital image by deep learning for melanoma diagnosis. In: Proceedings of the 15th international work-conference on artificial neural networks (IWANN), pp 270–279
33. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5:180161
34. Victor A, Ghalib M (2017) Automatic detection and classification of skin cancer. Int J Intell Eng Syst 10(3):444–451
35. Yadav V, Kaushik V (2018) Detection of melanoma skin disease by extracting high level features for skin lesions. Int J Adv Intell Paradig 11(3–4):397–408
36. Yu L, Chen H, Dou Q, Qin J, Heng PA (2017) Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 36(4):994–1004
37. Zhou T, Thung K, Zhu X, Shen D (2019) Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum Brain Mapp 40(3):1001–1016