Article

Bucket of Deep Transfer Learning Features and Classification Models for Melanoma Detection

Mario Manzo [1,*] **and Simone Pellino** [2]

[1] Information Technology Services, University of Naples "L'Orientale", 80121 Naples, Italy
[2] Computer Science Teacher, I.S. Mattei Aversa M.I.U.R., 81031 Rome, Italy; simonepellino@gmail.com
* Correspondence: mmanzo@unior.it

check for updates

Abstract: Malignant melanoma is the deadliest form of skin cancer and, in recent years, is rapidly growing in terms of the incidence worldwide rate. The most effective approach to targeted treatment is early diagnosis. Deep learning algorithms, specifically convolutional neural networks, represent a methodology for the image analysis and representation. They optimize the features design task, essential for an automatic approach on different types of images, including medical. In this paper, we adopted pretrained deep convolutional neural networks architectures for the image representation with purpose to predict skin lesion melanoma. Firstly, we applied a transfer learning approach to extract image features. Secondly, we adopted the transferred learning features inside an ensemble classification context. Specifically, the framework trains individual classifiers on balanced subspaces and combines the provided predictions through statistical measures. Experimental phase on datasets of skin lesion images is performed and results obtained show the effectiveness of the proposed approach with respect to state-of-the-art competitors.

Keywords: melanoma detection; deep learning; transfer learning; ensemble classification

1. Introduction

Among the types of malignant cancer, melanoma is the deadliest form of skin cancer and its incidence rate is growing rapidly around the world. Early diagnosis is particularly important since melanoma can be cured with a simple excision. In the majority, due to the similarity of the various skin lesions (melanoma and not-melanoma) [1], the visual analysis could be unsuitable and would lead to a wrong diagnosis. In this regard, image processing and artificial intelligence tools can provide a fundamental aid to a step of automatic classification [2]. Further improvement in diagnosis is provided by dermoscopy technique [3]. Dermoscopy technique can be applied to the skin, in order to capture illuminated and magnified images of the skin lesion in a non invasive way to highlight areas containing spots. Furthermore, the visual effect of the deeper skin layer can be improved if the skin surface reflection is removed. Anyhow, classification of melanoma dermoscopy images is a difficult task for different issues. First, the degree of similarity between melanoma and not-melanoma lesions. Second, the segmentation, and, therefore, the identification of the affected area is very complicated because of the variations in terms of texture, size, color, shape and location. The last issue and not the least, is the additional skin conditions such as hair, veins or variations due to image capturing. To this end, many solutions have been provided to improve the task. For example, low-level hand-crafted features [4] are adopted to discriminate non-melanoma and melanoma lesions. In some cases, these types of features are unable to discriminate clearly, leading to results that are

sometimes not very relevant [5]. Differently, segmentation is adopted to isolate the foreground elements from the background ones [6]. Consequently, the segmentation includes low-level features with a low representational power that provides unsatisfactory results [7]. In recent years, deep learning has become an effective solution for the extraction of significant features on large data. In particular, the diffusion of deep neural networks, applied to the image classification task, is connected to various factors such as the availability of software in terms of open source license, the constant growth of hardware power and the availability of large datasets [8]. Deep learning has proven effective for the management, analysis, representation and classification of medical images [9]. Specifically, for the treatment of melanoma, deep neural networks were adopted both in segmentation and classification phases [10]. However, the high variation of the types of melanoma and the imbalance of the data have a decisive impact on performance [11], hindering the generalization of the model and leading to over-fitting [12]. In order to overcome the aforementioned issues, in this paper, we introduce a novel framework based on transfer deep learning and ensemble classification for melanoma detection. It works based on three integrated stages. A first, which performs image preprocessing operations. A second, which extracts features using transfer deep learning. A third, including a layer of ensemble learning, in which different classification algorithms and features extracted are combined with the aim of making the best decision (melanoma/not-melanoma). Our approach provides the following main contributions:

- A deep and ensemble learning-based framework, to simultaneously address inter-class variation and class imbalance for the task of melanoma classification.
- A framework that, in the classification phase, at the same time, creates multiple image representation models, based on features extracted with deep transfer learning.
- The demonstration of how the choice of multiple features can enrich image representation by leading a lesion assessment like a skilled dermatologist.
- Some experimental greater improvements over existing methods on different state of art datasets about melanoma detection task.

The paper is structured as follows. Section 2 provides an overview of state-of-the-art about melanoma classification approaches. Section 3 describes in detail proposed framework. Section 4 provides a wide experimental phase, while Section 5 concludes the paper.

2. Related Work

In this section, we briefly analyze the most important approaches of skin lesions recognition literature. In this field are included numerous works that address the issue according to different aspects. Some works offer an important contribution about image representation, by implementing segmentation algorithms or new descriptors. Instead, others implement complex mechanisms of learning and classification.

In Reference [13], a novel boundary descriptor based on the color variation of the skin lesion input images, achieved with standard cameras, is introduced. Furthermore, in order to reach higher performance, a set of textural and morphological features is added. Multilayer perceptron neural network as classifier is adopted.

In Reference [14], authors propose a complex framework that implements an illumination correction and features extraction on skin image lesions acquired using normal consumer-grade cameras. Applying a multi-stage illumination improvement algorithm and defining a set of high-level intuitive features (HLIF), that quantifies the level of asymmetry and border irregularity about a lesion, the proposed model can be used to classify accurate skin lesion diagnoses.

While in Reference [15], authors, to properly evaluate contents of the concave contours, introduce a novel border descriptor named boundary intersection-based signature (BIBS). Shape signature is a

one-dimensional illustration of shape border and cannot contribute to a proper description for concave borders that have more than one intersection points. For this reason, BIBS analyzes boundary contents of shape especially shapes with concave contours. Support vector machine (SVM) for classification process is adopted.

Another descriptor for the individualization of skin lesions is named high-level intuitive features (HLIFs) [16]. HLIFs are created to simulate a model of human-observable characteristics. It captures specific characteristics that are significant to the given application: color asymmetry—analyzing and clustering pixels colors, structural asymmetry—applying the Fourier descriptors of the shape, border irregularity—using morphological opening and closing, color characteristics—transforming the image to a perceptually uniform color space, building color-spatial representations that model the color information for a patch of pixels, clustering the patch representations into k color clusters, quantifying the variance found using the original lesion and the k representative colors.

A texture analysis method of Local Binary Patterns (LBP) and Block Difference of Inverse Probabilities is proposed in Reference [17]. A comparison is provided with classification results obtained by taking the raw pixel intensity values as input. Classification stage is achieved generating an automated model obtained by both convolutional neural networks (CNN) and SVM.

In Reference [18], authors propose a system that automatically extracts the lesion regions, using non-dermoscopic digital images, and then computes color and texture descriptors. Extracted features are adopted for automatic prediction step. The classification is managed using a majority vote of all predictions.

In Reference [19], non-dermoscopic clinical images to assist a dermatologist in early diagnosis of melanoma skin cancer are adopted. Images are preprocessed in order to reduce artifacts like noise effects. Subsequently, images are analyzed through a pretrained CNN which is a member of deep learning models. CNNs are trained by a large number of training samples in order to distinguish between melanoma and benign cases.

In Reference [20], Predict-Evaluate-Correct K-fold (PECK) algorithm is presented. The algorithm works by merging deep CNNs with SVM and random forest classifiers to achieve an introspective learning method. In addition, authors provides a novel segmentation algorithm, named Synthesis and Convergence of Intermediate Decaying Omnigradients (SCIDOG), to accurately detect lesion contours in non-dermoscopic images, even in the presence of significant noise, hair and fuzzy lesion boundaries.

In Reference [21], authors propose a novel solution to improve melanoma classification by defining a new feature that exploits the border-line characteristics of the lesion segmentation mask combining gradients with LBP. These border-line features are used together with the conventional ones and lead to higher accuracy in classification stage.

In Reference [22], an objective features extraction function for CNN is proposed. The goal is to acquire the variation separability as opposed to the categorical cross entropy which maximizes according to the target labels. The deep representative features increase the variance between the images making it more discriminative. In addition, the idea is to build a CNN and perform principal component analysis (PCA) during the training phase.

In Reference [23], a deep learning computer aided diagnosis system for automatic segmentation and classification of melanoma lesions is proposed. The system extracts CNN and statistical and contrast location features on the results of raw image segmentation. The combined features are utilized to obtain the final classification of melanoma, malignant or benign.

In Reference [24], authors propose an efficient algorithm for prescreening of pigmented skin lesions for malignancy using general-purpose digital cameras. The proposed method enhances borders and extracts a broad set of dermatologically important features. These discriminative features allow classification of lesions into two groups of melanoma and benign.

In Reference [25], a skin lesion detection system optimized to run entirely on the resource constrained smartphone is described. The system combines a lightweight method for skin detection with a hierarchical segmentation approach including two fast segmentation algorithms and proposes novel features to characterize a skin lesion. Furthermore, the system implements an improved features selection algorithm to determine a small set of discriminative features adopted by the final lightweight system.

Multiple-instance learning (MIL)-based approaches are of great interest in recent years. MIL is a type of supervised learning and works by receiving a set of instances, named bags, individually labeled. In Reference [26], authors present an MIL approach with application to melanoma detection. The goal was to discriminate between positive and negative sets of items. The main rule concerns a bag that is positive if at least one of its instances is positive and it is negative if all its instances are negative. Differently in Reference [27], MIL approaches are described with purpose to discriminate melanoma from dysplastic nevi. Specifically, authors introduce an MIL approach that adopts spherical separation surfaces. Finally, in Reference [28], a preliminary comparison between two different approaches, SVM and MIL, is proposed, focusing on the key role played by the feature selection (color and texture). In particular, the authors are inspired by the good results obtained applying MIL techniques for classifying some medical dermoscopic images.

3. Materials and Methods

In this section, we describe the proposed framework which includes two well known methodologies: deep neural network and ensemble learning. The main idea is to combine algorithms of features extraction and classification. The result is a set of competitive models providing a range of confidential decisions useful for making choices during classification. The framework is composed of three levels. A first, which performs preprocessing operations such as image resize and data balancing. A second, of transfer learning, which extracts features using deep neural networks. A third level, of ensemble learning, in which different classification algorithms (SVM [29], Logistic Label Propagation (LLP) [30], KNN [31]) and features extracted are combined with the aim of making the best decision. Adopted classifiers are trained and tested through a bootstrapping policy. Finally, the framework iterates through a predetermined number of times in a supervised learning context. Figure 1 shows a graphic overview of the proposed framework.
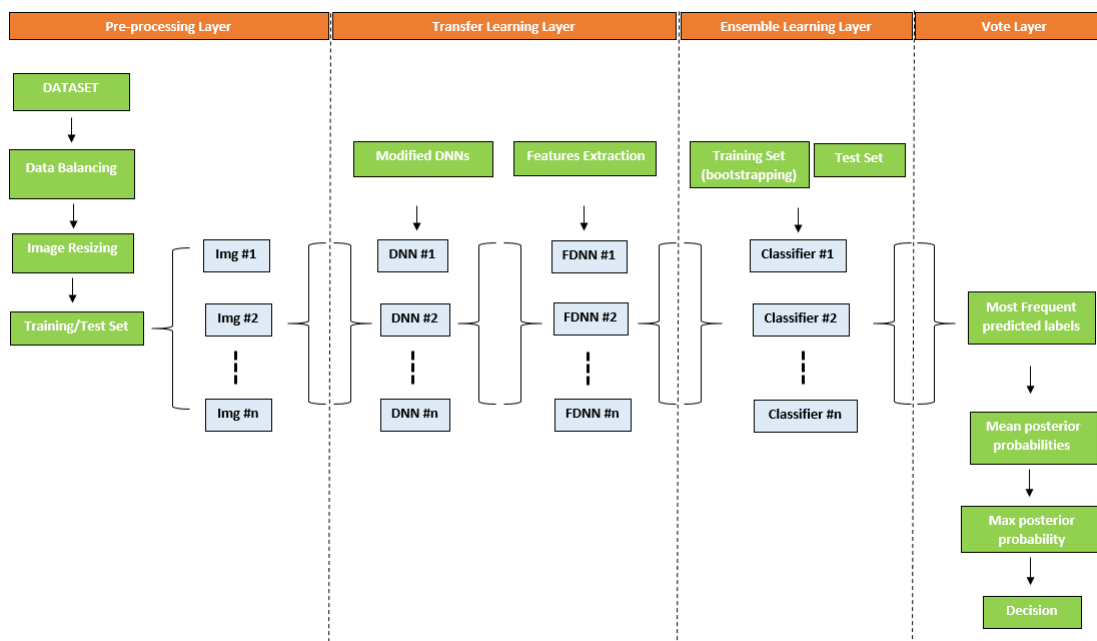
Figure 1. Overview of the proposed framework.

3.1. Data Balancing

Melanoma lesion analysis and classification is connected with accurate segmentation with purpose to isolate areas of the image containing information of interest. Moreover, the wide variety of skin lesions and the unpredictable obstructions on the skin make traditional segmentation an ineffective tool, especially for non-dermoscopic images. Furthermore, the problem of imbalance, present in many datasets, makes the classification difficult to address, especially when the samples of the minority class are very underrepresented. In the case under consideration, to compensate the strong imbalance between the two classes, a balancing phase was performed. The goal was to isolate segments of the image that could contain melanoma. In particular, the resampling of the minority class is performed by adding images altered through the application of K-Means color segmentation algorithm [32]. The application of segmentation algorithms for image augmentation [33], and consequently to provide a balancing between classes, represented a good compromise for this stage of the pipeline.

3.2. Image Resize

Images to be processed have been resized based on the dimension, related to the input layer, claimed by the deep neural networks (details can be found in Table 1 column 5). Many of the networks require this type of step but it does not alter the image information content in any way. This normalization step is essential because images of different or large dimensions cannot be processed for the features extraction stage.

Table 1. Description of the adopted pretrained network.

Network	Depth	Size (MB)	Parameters (Millions)	Input Size	Features Layer
Alexnet	8	227	61	227×227	fc7
Googlenet	8	27	7	224×224	pool5-7x7_s1
Resnet18	18	44	11.7	224×224	pool5
Resnet50	50	96	25.6	224×224	avg_pool

3.3. Transfer Learning and Features Extraction

The transfer learning approach has been chosen for features extraction purpose. Commonly, a pretrained network is adopted as starting point to learn a new task. It is the easiest and fastest way to exploit the representational power of pretrained deep networks. It is usually much faster and easier to tune a network with transfer learning than training a new network from scratch with randomly initialized weights. We have selected deep learning architectures for image classification based on their structure and performance skills. The goal was to extract features from images through neural networks by redesigning their structures in the final layer according to the needs of the addressed task (two outgoing classes: melanoma and not-melanoma). The features extraction is performed through a chosen layer (different for each network and specified in the Table 1), placed in the final part of the structure. The image will be encoded through a vector of real numbers produced by consecutive convolution steps, from the input layer to the layer chosen for the representation. Below, a description of the adopted networks is reported.

Alexnet [8] consists of 5 convolutional layers and 3 fully connected layers. It includes the non-saturating ReLU activation function, better then tanh and sigmoid during training phase. For features extraction, we have chosen a fully connected 7 (fc7) layer composed of 4096 neurons.

Googlenet [34] is composed of 22 layers deep. The network is inspired by LeNet [35] but implemented a novel element which is dubbed an inception module. This module is based on several very small convolutions in order to drastically reduce the number of parameters. Their architecture reduced the number of parameters from 60 million (AlexNet) to 4 million. Furthermore, it includes batch normalization, image distortions and Root Mean Square Propagation algorithm. For features extraction, we have chosen global average pooling (pool5-7x7_s1) layer composed of 1024 neurons.

Resnet18 and Resnet50 [36] are inspired by pyramidal cells contained in the cerebral cortex. They use particular skip connections or shortcuts to jump over some layers. They are composed of 18 and 50 layers deep, which with the help of a technique known as skip connection has paved the way for residual networks. For feature extraction, we have chosen two global average pooling (pool5 and avg-pool) layers composed of 512 and 2048 neurons, respectively.

3.4. Network Design

The adopted networks have been adapted to the melanoma classification problem. Originally, they have been trained on the Imagenet dataset [37], composed of a million images and classified into 1000 classes. The result is a rich features representation for a wide range of images. The network processes an image and provides a label along with probabilities for each of the classes. Commonly, the first layer of the network is the image input layer. This requires input images with 3 color channels. Just after, convolutional layers work to extract image features in which the last learnable layer and the final classification layer adopt to classify the input image. In order to make suitable the pretrained network to classify new images, the two last layers with new layers are replaced. In many cases, the last layer, including learnable weights, is a fully connected layer. This is replaced with a new fully connected layer related to the number of outputs equal to the number of classes of new data. Moreover, to speedup the learning in the new layer with respect to transferred layers, it is recommended to increase the learning rate factors. As an optional choice, the weights of earlier layers can be frozen by setting the related learning rate to zero. This setting produces a failure of update of the weights during the training, and a consequent lowering of the execution time as the gradients of the related layers must not be calculated. This aspect is very interesting to avoid overfitting in the case of small datasets.

3.5. Ensemble Learning

The contribution of different transfer learning features and classifiers can be mixed in an ensemble context. Considering the set of images, with cardinality k, belonging to x classes, to be classified

$$Imgs = \{i_1, i_2, \ldots, i_k\} \tag{1}$$

each element of the set will be treated with the procedure below. Let us consider the set C composed of n classifiers

$$C = \{\beta_1, \beta_2, \ldots, \beta_n\} \tag{2}$$

and set F composed of m vectors of transferred learning features

$$F = \{\Theta_1, \Theta_2, \ldots \Theta_m\} \tag{3}$$

the goal is the combination each element of the set C with the elements of the set F. The set of combinations can be defined as CF

$$CF = \begin{bmatrix} \beta_1\Theta_1 & \cdots & \beta_1\Theta_m \\ \vdots & \ddots & \\ \beta_n\Theta_1 & & \beta_n\Theta_m \end{bmatrix} \tag{4}$$

each combination provides a decision $i \in I\{-1, 1\}$, where 1 stands for melanoma and -1 for not-melanoma, related to image of the set $Imgs$. The set of decisions D can be defined as follows

$$D = \begin{bmatrix} d_{\beta_1\Theta_1} & \cdots & d_{\beta_1\Theta_m} \\ \vdots & \ddots & \\ d_{\beta_n\Theta_1} & & d_{\beta_n\Theta_m} \end{bmatrix} \tag{5}$$

Each $d_{\beta_i\Theta_j}$ value represents a decision based on the combination of sets C and F. In addition, the set of scores S can be defined as follows

$$S = \begin{bmatrix} P(i|x)_{d_{\beta_1\Theta_1}} & \cdots & P(i|x)_{d_{\beta_1\Theta_m}} \\ \vdots & \ddots & \\ P(i|x)_{d_{\beta_n\Theta_1}} & & P(i|x)_{d_{\beta_n\Theta_m}} \end{bmatrix} \tag{6}$$

a score value, $s \in S\{0, \ldots, 1\}$, is associated with each decision d and represents the posterior probability $P(i|x)$ that an image i belongs to class x. At this point, let us introduce the concept of mode, defined as the value which is repeatedly occurred in a given set

$$mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h \tag{7}$$

where l is the lower limit of the modal class, h is the size of the class interval, f_1 is the frequency of the modal class, f_0 is the frequency of the class which precedes the modal class and f_2 is the frequency of the class which successes the modal class. The columns of matrix D are analyzed with the mode, in order to obtain the values of the most frequent decisions. This step is carried out in order to verify the best response of the different classifiers, contained in the set C, which adopt the same type of features. Moreover, the *mode* provides two indications. The most frequent value and its occurrences (indices). For each most

frequent occurrence, modal value, the corresponding score of the matrix S is extracted. In this regard, a new vector is generated

$$DS = \{ds_{P(i|x)_{d_{\beta_{1,\ldots,n}\Theta_1}}}, \ldots, ds_{P(i|x)_{d_{\beta_{1,\ldots,n}\Theta_m}}}\}, \tag{8}$$

where each element ds contains the average of the scores that have a higher frequency, extracted through the *mode*, in the related column of the matrix D. In addition, the modal value of each column of the matrix D is stored in the vector DM

$$DM = \{dm_{d_{\beta_{1,\ldots,n}\Theta_1}}, \ldots, dm_{d_{\beta_{1,\ldots,n}\Theta_m}}\}, \tag{9}$$

the final decision will consist in the selection of the element of the vector DM with the same position of the maximum score value of the vector DS. This last step verifies the best prediction based on the different features adopted, essentially the best features suitable for the classification of the image.

3.6. Train and Test Strategy: Bootstrapping

Bootstrapping is a statistical technique which consists of creating samples of size B, named bootstrap samples, from a dataset of size N. The bootstrap samples are randomly inserted with replacement on the dataset. This strategy has important statistical properties. First, subsets can be considered as directly extracted from the original distribution, independently of each others, containing representative and independent samples, almost independent and identically distributed (idd). Two considerations must be made in order to validate the hypotheses. First, the N dimension of the original dataset should be large enough to detect the underlying distribution. Sampling the original data is a good approximation of real distribution (representativeness). Second, the N dimension of the dataset should be better than the B dimension of the bootstrap samples so that the samples are not too correlated (independence). Commonly, considering the samples to be truly independent means requiring too much data compared to the amount actually available. This strategy can be adopted to generate several bootstrap samples that can be considered nearly representative and almost independent (almost iid samples). In the proposed framework, bootstrapping is applied to set F (Equation (3)) in order to perform the training and testing stages of classifiers. This strategy seemed suitable for the problem faced in order to create a competitive environment capable of providing the best performance.

4. Experimental Results

This section describes the experiments performed on public datasets. In order to produce compliant performance, the settings included in well-known melanoma classification methods, in which the main critical issue concerns the features extraction for image representation, are adopted.

4.1. Datasets

The first adopted dataset is MED-NODE (http://www.cs.rug.nl/~imaging/databases/melanoma_naevi/). It was created by the Department of Dermatology of the University Medical Center Groningen (UMCG). The dataset was initially used to train the MED-NODE computer assisted melanoma detection system [18]. It is composed of 170 non-dermoscopic images, where 70 are melanoma and 100 are nevi. The image dimensions vary greatly, ranging from 201×257 to 3177×1333 pixels.

The second adopted dataset, Skin-lesion (from now), is described in Reference [16]. It is composed of 206 images of skin lesion, which were obtained using standard consumer-grade cameras in varying and unconstrained environmental conditions. These images were extracted from the online public databases Dermatology Information System (http://www.dermis.net) and DermQuest (http://www.dermquest.com). Of these images, 119 are melanomas, and 87 are not-melanoma. Each image contains a single lesion of interest.

4.2. Settings

The framework consists of different modules written in Matlab language. Moreover, we applied pretrained networks available which are included in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [38]. Among all the computational stages, the features extraction process, described in Section 3.3, was certainly the most expensive. As is certainly known, the networks are composed of fully connected layers that make the structure extremely dense and complex. This aspect certainly increases the computational load. Alexnet, Googlenet, Resnet50 are adopted to extract features on MED-NODE dataset. Differently, Resnet50 and Resnet18 are adopted for Skin-lesion dataset. The choice is not random but was made based on two criteria. Primarily, a study about the network specifications and characteristics most suitable for the problem faced in literature and, secondly, the performance obtained. A different combination did not provide expected feedback. In the Table 1, some important details related to the layers chosen for feature extraction are shown. Networks were trained by setting the mini batch size to 5, the maximum epochs to 10, the initial learning rate to 3×10^{-4} and the optimizer is stochastic gradient descent with momentum (SGDM) algorithm. For both experimental procedures, in order to train the classifiers, 80% and 20% of images are included in train and test sets, respectively, for a number of iterations equal to 10. Table 2 enumerates classification algorithms included in the framework and related settings (some algorithms appear more times with different configurations). For completeness and clarity, and in order to demonstrate the best solution, both results of combinations adopted, even if they did not provide the best performance, are indicated in Tables 4 and 5.

Table 2. Classification algorithms and related settings.

Algorithms	Setting
SVM [29]	KernelFunction:polynomial, KernelScale: auto
SVM [29]	KernelFunction: Gaussian, KernelScale: auto
LLP [30]	KernelFunction: rbf, Regularization parameter: 1, init: 0, maxiter: 1000
KNN [31]	NumNeighbors: 3, Distance: spearman
KNN [31]	NumNeighbors: 4, Distance: correlation

4.3. Discussion

Table 3 shows the metrics adopted for the performance evaluation, in order to provide a uniform comparison with algorithms working on the same task.

Table 3. Evaluation metrics adopted during the relevance feedback stage.

Metric	Equation
True Positive Rate	$TPR = \dfrac{TP}{TP + FN}$
True Negative Rate	$TNR = \dfrac{TN}{TN + FP}$
Positive Predictive Value	$PPV = \dfrac{TP}{TP + FP}$
Negative Predictive Value	$NPV = \dfrac{TN}{TN + FN}$
Accuracy	$ACC = \dfrac{TP + FN}{TP + FP + TN + FN}$
F_1-Score(Positive)	$F_1^P = \dfrac{2 \cdot PPV \cdot TPR}{PPV + TPR}$
F_1-Score(Negative)	$F_1^N = \dfrac{2 \cdot NPV \cdot TNR}{NPV + TNR}$
Matthew's Correlation Coefficient	$MCC = \dfrac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$

Looking carefully at the table, it is important to focus on the meaning of the individual measures with reference to melanoma detection. The True Positive rate, also known as Sensitivity, concerns the portion of positives melanoma images that are correctly identified. This provide important information because highlights the skill to identify images containing skin lesions and contributes to increase the degree of robustness of result. The same concept is true for the True Negative rate, also known as Specificity, which instead measures the portion of negatives, not containing skin lesions, that have been correctly identified. The Positive and Negative Predictive values, also known as Precision and Recall, respectively, are probabilistic measures that indicate whether an image with a positive or negative melanoma test may or may not have a skin lesion. In essence, Recall expresses the ability to find all relevant instances in the dataset, Precision expresses the proportion of instances that the framework claims to be relevant were actually relevant. Accuracy, a well-known performance measure, is the proportion of true results among the total number of cases examined. In our case, it provides an overall analysis, certainly a rough measurement compared to the previous ones, about the skill of a classifier to distinguish a skin lesion from an image without lesions. F_1-Score measure combines the Precision and Recall of the model, as the harmonic mean, in order to find an optimal blend. The choice of the harmonic mean instead of a simple mean concerns the possibility of eliminating extreme values. Finally, Matthew's correlation coefficient is another overall well-known quality measure. It takes into account True/False Positives/Negatives values and is generally regarded as a balanced measure which can be adopted even if the classes are of very different sizes.

The Tables 4 and 5 describe the comparison with existing skin cancer classification methods (we referred with the results which appear in the corresponding papers). The provided performance can be considered satisfactory compared to competitors. In terms of accuracy, although it provides a rough measurement, we have provided the best result for MED-NODE and the second for Skin-lesion (only surpassed by BIBS). Differently, predictive positive value and negative positive value give good indications on the classification ability. True positive rate, a measure that provides greater confidence about addressed problem, is very high for both datasets. Otherwise, true negative rate, which also provides a high degree of sensitivity related to the absence of tumors within the image, is the best value for both datasets. Regarding the remaining measures, F_1^p, F_1^n and Matthew's Correlation Coefficient, considerable values were obtained

but, unfortunately, not available for all competitors. We can certainly attribute the satisfactory performance to two main aspects. First, the deep learning features, which even if abstract, are able to best represent the images. Furthermore, the framework provides multiple representation models that certainly constitute a different starting point than a standard approach, in which a single representation is provided. This aspect is relevant for improving performance. A non negligible issue, the normalization of the image size, with respect to the request of the first layer of the neural network, before the features extraction phase, does not produce a performance degradation. In other cases, normalization causes loss of quality of the image content and a consequent degradation of details. Otherwise, the weak point is the computational load even if pretrained networks include layers with already tuned weights. Surely, the time required for training is long but less than a network created from scratch. Second, the classification scheme, which provides multiple choices in decision making. In fact, at each iteration, the framework chooses which classifier is suitable for recognizing melanoma in the images included in the proposed set. Certainly, this approach is more computationally expensive but produces better results than a single classifier.

Table 4. Experimental results on the MED-NODE dataset.

Method	TPR	TNR	PPV	NPV	ACC	F_1^p	F_1^n	MCC
MED-NODE annoted [18]	0.78	0.59	0.56	0.80	0.66	0.65	0.68	0.36
Spotmole [39]	0.82	0.57	0.56	0.83	0.67	0.67	0.68	0.39
Barhoumi and Zagrouba [40]	0.46	0.87	0.70	0.71	0.70	0.56	0.78	0.37
MED-NODE color [18]	0.74	0.72	0.64	0.81	0.73	0.69	0.76	0.45
MED-NODE texture [18]	0.62	0.85	0.74	0.77	0.76	0.67	0.81	0.49
Jafari et al. [24]	0.90	0.72	0.70	0.91	0.79	0.79	0.80	0.61
MED-NODE combined [18]	0.80	0.81	0.74	0.86	0.81	0.77	0.83	0.61
Nasr Esfahani et al. [19]	0.81	0.80	0.75	0.86	0.81	0.78	0.83	0.61
Benjamin Albert [20]	0.89	0.93	0.92	0.93	0.91	0.89	0.92	0.83
Pereira et [21] ght/svm-smo/f23-32	0.45	0.92	-	-	0.73	-	-	-
Pereira et [21] ght/svm-smo/f1-32	0.56	0.86	-	-	0.74	-	-	-
Pereira et al. [21] lbpc/svm-smo/f23-32	0.49	0.93	-	-	0.75	-	-	-
Pereira et al. [21] lbpc/svm-smo/f1-32	0.58	0.91	-	-	0.78	-	-	-
Pereira et al. [21] ght/svm-sda/f23-32	0.66	0.83	-	-	0.76	-	-	-
Pereira et al. [21] ght/svm-sda/f1-32	0.66	0.86	-	-	0.78	-	-	-
Pereira et al. [21] lbpc/svm-isda/f23-32	0.69	0.83	-	-	0.77	-	-	-
Pereira et al. [21] lbpc/svm-isda/f1-32	0.65	0.88	-	-	0.79	-	-	-
Pereira et al. [21] ght/ffn/f23-32	0.63	0.84	-	-	0.76	-	-	-
Pereira et al. [21] ght/ffn/f1-32	0.63	0.84	-	-	0.76	-	-	-
Pereira et al. [21] lbpc/ffn/f23-32	0.64	0.83	-	-	0.75	-	-	-
Pereira et al. [21] lbpc/ffn/f1-32	0.66	0.86	-	-	0.77	-	-	-
Sultana et al. [22]	0.73	0.86	0.77	0.83	0.81	-	-	-
Ge, Yunhao and Liet al. [23]	0.94	0.93	-	-	0.92	-	-	-
Mandal et al.[41] Case 1	0.61	0.65	0.74	0.87	0.65	-	-	-
Mandal et al.[41] Case 2	0.80	0.73	0.74	0.87	0.71	-	-	-
Mandal et al.[41] Case 3	0.84	0.66	0.68	0.86	0.71	-	-	-
Jafari et al. [42]	0.82	0.71	0.67	0.85	0.76	-	-	-
T. Do et al. [25] Color	0.81	0.73	0.66	0.85	0.75	-	-	-
T. Do et al. [25] Texture	0.66	0.85	0.75	0.79	0.78	-	-	-
T. Do et al. [25] Color and Texture	0.84	0.72	0.70	0.87	0.77	-	-	-
E. Nasr-Esfahani et al. [19]	0.81	0.80	0.75	0.86	0.81	-	-	-
Resnet50+Resnet18	0.80	1.00	1.00	0.83	0.90	0.88	0.90	0.81
Resnet50+Googlenet+Alexnet	0.90	0.97	0.97	0.90	0.93	0.93	0.94	0.87

Table 5. Experimental results on the Skin-lesion dataset.

Method	TPR	TNR	PPV	NPV	ACC	F_l^p	F_l^n	MCC
Texture analysis [17]	0.87	0.71	0.76	-	0.75	-	-	-
HLIFs [16]	0.96	0.73	-	-	0.83	-	-	-
BIBS [15]	0.92	0.88	0.91	-	0.90	-	-	-
Decision Support [14]	0.84	0.79	-	-	0.81	-	-	-
Color pigment boundary [13]	0.95	0.88	0.92	-	0.82	-	-	-
R. Amelard et al. [43] Asymmetry F_C	0.73	0.64	-	-	0.69	-	-	-
R. Amelard et al. [43] Proposed HLIFs	0.79	0.68	-	-	0.75	-	-	-
R. Amelard et al. [43] Cavalcanti feature set	0.84	0.78	-	-	0.82	-	-	-
R. Amelard et al. [43] Modified F_C	0.86	0.75	-	-	0.72	-	-	-
R. Amelard et al. [43] Combined F_{MC} F_A^{HLIFS}	0.91	0.80	-	-	0.86	-	-	-
Resnet50+Resnet18	0.84	0.92	0.91	0.85	0.88	0.87	0.88	0.76
Resnet50+Googlenet+Alexnet	0.87	0.65	0.71	0.84	0.76	0.78	0.73	0.54

5. Conclusions and Future Works

The challenge in the discrimination of melanoma and nevi has resulted to be very interesting in recent years. The complexity of the task is linked to different factors such as the large amount of types of melanomas or the difficulties for digital phase acquisition (noise, lighting, angle, distance and much more). Machine learning classifiers suffer greatly these factors and inevitably reflect on the quality of the results. In support, the convolutional neural networks give a big hand for both classification and features extraction phases. In this context, we have proposed a framework that combines standard classifiers and features extracted with convolutional neural networks using a transfer learning approach. The results produced certainly support the theoretical thesis. A multiple representation of the image compared to a single one is a high discrimination factor even if the features adopted are completely abstract. The extensive experimental phase has shown how the proposed approach is competitive, and in some cases surpassing, with respect to state-of-the-art methods. Certainly, the main weak point concerns the computational complexity relating to features extraction phase, as it is known, takes a long time especially when the data to be processed grows. Future work will certainly concern the study and analysis of convolutional neural networks still unexplored for this type of problem, the application of the proposed framework to additional datasets (such as PH^2 [44]) and alternative tasks from the melanoma detection. Finally, also interesting are different dataset balancing approaches, such as proposed in [45] where all the melanom images are duplicated by including zero-mean Gaussian noise with variance equal to 0.0001.

References

1. Codella, N.; Cai, J.; Abedini, M.; Garnavi, R.; Halpern, A.; Smith, J.R. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 118–126.
2. Mishra, N.K.; Celebi, M.E. An overview of melanoma detection in dermoscopy images using image processing and machine learning. *arXiv* **2016**, arXiv:1601.07843.

3. Binder, M.; Schwarz, M.; Winkler, A.; Steiner, A.; Kaider, A.; Wolff, K.; Pehamberger, H. Epiluminescence microscopy: A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Arch. Dermatol.* **1995**, *131*, 286–291. [CrossRef] [PubMed]

4. Barata, C.; Celebi, M.E.; Marques, J.S. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 1096–1109. [CrossRef] [PubMed]

5. Celebi, M.E.; Kingravi, H.A.; Uddin, B.; Iyatomi, H.; Aslandogan, Y.A.; Stoecker, W.V.; Moss, R.H. A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph.* **2007**, *31*, 362–373. [CrossRef]

6. Tommasi, T.; La Torre, E.; Caputo, B. Melanoma recognition using representative and discriminative kernel classifiers. In *International Workshop on Computer Vision Approaches to Medical Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.

7. Pathan, S.; Prabhu, K.G.; Siddalingaswamy, P. A methodological approach to classify typical and atypical pigment network patterns for melanoma diagnosis. *Biomed. Signal Process. Control.* **2018**, *44*, 25–37. [CrossRef]

8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*; ACM Digital Library: New York, NY, USA, 2012 , pp. 1097–1105.

9. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015, pp. 234–241.

10. Yu, L.; Chen, H.; Dou, Q.; Qin, J.; Heng, P.A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans. Med Imaging* **2016**, *36*, 994–1004. [CrossRef]

11. Shie, C.K.; Chuang, C.H.; Chou, C.N.; Wu, M.H.; Chang, E.Y. Transfer representation learning for medical image analysis. In Proceedings of the 2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 711–714.

12. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef]

13. Jayant Sachdev, Shashank Shekhar, D.S.I. Skin Lesion Images Classification Using New Color Pigmented Boundary Descriptors. In Proceedings of the 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA 2017), Shahrekord, Iran, 19–20 April 2017.

14. Amelard, R.; Glaister, J.; Wong, A.; Clausi, D.A. Melanoma Decision Support Using Lighting-Corrected Intuitive Feature Models. In *Computer Vision Techniques for the Diagnosis of Skin Cancer*; Springer: Berlin/Heidelberg, Germany, 2013, pp. 193–219.

15. Mahdiraji, S.A.; Baleghi, Y.; Sakhaei, S.M. BIBS, a New Descriptor for Melanoma/Non-Melanoma Discrimination. In Proceedings of the Iranian Conference on Electrical Engineering (ICEE), Mashhad, Iran, 8–10 May 2018, pp. 1397–1402.

16. Amelard, R.; Glaister, J.; Wong, A.; Clausi, D.A. High-Level Intuitive Features (HLIFs) for Intuitive Skin Lesion Description. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 820–831. [CrossRef]

17. Karabulut, E.; Ibrikci, T. Texture analysis of melanoma images for computer-aided diagnosis. In Proceedings of the International Conference on Intelligent Computing, Computer Science & Information Systems (ICCSIS 16), Pattaya, Thailand, 28–29 April 2016; Volume 2, pp. 26–29.

18. Giotis, I.; Molders, N.; Land, S.; Biehl, M.; Jonkman, M.; Petkov, N. MED-NODE: A Computer-Assisted Melanoma Diagnosis System using Non-Dermoscopic Images. *Expert Syst. Appl.* **2015**, *42*, [CrossRef]

19. Nasr-Esfahani, E.; Samavi, S.; Karimi, N.; Soroushmehr, S.M.R.; Jafari, M.H.; Ward, K.; Najarian, K. Melanoma detection by analysis of clinical images using convolutional neural network. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 1373–1376.

20. Albert, B.A. Deep Learning From Limited Training Data: Novel Segmentation and Ensemble Algorithms Applied to Automatic Melanoma Diagnosis. *IEEE Access* **2020**, *8*, 31254–31269. [CrossRef]

21. Pereira, P.M.; Fonseca-Pinto, R.; Paiva, R.P.; Assuncao, P.A.; Tavora, L.M.; Thomaz, L.A.; Faria, S.M. Skin lesion classification enhancement using border-line features—The melanoma vs. nevus problem. *Biomed. Signal Process. Control* **2020**, *57*, 101765, [CrossRef]

22. Sultana, N.N.; Puhan, N.B.; Mandal, B. DeepPCA Based Objective Function for Melanoma Detection. In Proceedings of the 2018 International Conference on Information Technology (ICIT), Bhubaneswar, India, 19–21 December 2018; pp. 68–72.

23. Ge, Y.; Li, B.; Zhao, Y.; Guan, E.; Yan, W. Melanoma segmentation and classification in clinical images using deep learning. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, 26–28 February 2018; pp. 252–256.

24. Jafari, M.H.; Samavi, S.; Karimi, N.; Soroushmehr, S.M.R.; Ward, K.; Najarian, K. Automatic detection of melanoma using broad extraction of features from digital images. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 1357–1360.

25. Do, T.; Hoang, T.; Pomponiu, V.; Zhou, Y.; Chen, Z.; Cheung, N.; Koh, D.; Tan, A.; Tan, S. Accessible Melanoma Detection Using Smartphones and Mobile Image Analysis. *IEEE Trans. Multimed.* **2018**, *20*, 2849–2864. [CrossRef]

26. Astorino, A.; Fuduli, A.; Veltri, P.; Vocaturo, E. Melanoma detection by means of Multiple Instance Learning. *Interdiscip. Sci. Comput. Life Sci.* **2020**, *12*, 24–31. [CrossRef] [PubMed]

27. Vocaturo, E.; Zumpano, E.; Giallombardo, G.; Miglionico, G. DC-SMIL: A multiple instance learning solution via spherical separation for automated detection of displastyc nevi. In Proceedings of the 24th Symposium on International Database Engineering & Applications, Incheon (Seoul), South Korea, August 12–18, 2020; pp. 1–9.

28. Fuduli, A.; Veltri, P.; Vocaturo, E.; Zumpano, E. Melanoma detection using color and texture features in computer vision systems. *Adv. Sci. Technol. Eng. Syst. J.* **2019**, *4*, 16–22. [CrossRef]

29. Corinna Cortes, V.V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

30. Kobayashi, T.; Watanabe, K.; Otsu, N. Logistic label propagation. *Pattern Recognit. Lett.* **2012**, *33*, 580–588. [CrossRef]

31. Dasarathy, B.V. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*; IEEE Computer Society Press: Los Alamitos, CA, USA 1991; ISBN 978-0-8186-8930-7

32. Likas, A.; Vlassis, N.; Verbeek, J.J. The global k-means clustering algorithm. *Pattern Recognit.* **2003**, *36*, 451–461. [CrossRef]

33. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]

34. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

35. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]

36. He.; Kaiming.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA , 27–30 June 2016; pp. 770–778.

37. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

38. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]

39. Munteanu, C.; Cooclea, S. Spotmole Melanoma Control System. 2009. https://play.google.com/store/apps/details?id=com.spotmole&hl=en=AU (accessed on 18 November 2020).

40. Zagrouba, E.; Barhoumi, W. A prelimary approach for the automated recognition of malignant melanoma. *Image Anal. Stereol.* **2004**, *23*, 121–135, [CrossRef]

41. Mandal, B.; Sultana, N.; Puhan, N. Deep Residual Network with Regularized Fisher Framework for Detection of Melanoma. *IET Comput. Vis.* **2018**, *12*, [CrossRef]

42. Jafari, M.H.; Samavi, S.; Soroushmehr, S.M.R.; Mohaghegh, H.; Karimi, N.; Najarian, K. Set of descriptors for skin cancer diagnosis using non-dermoscopic color images. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 2638–2642.

43. Amelard, R.; Wong, A.; Clausi, D.A. Extracting high-level intuitive features (HLIF) for classifying skin lesions using standard camera images. In Proceedings of the 2012 Ninth Conference on Computer and Robot Vision, Toronto, ON, Canada, 28–30 May 2012; pp. 396–403.

44. Mendonca, T.; Celebi, M.; Mendonca, T.; Marques, J. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy Image Anal.* **2015**, 419–439, doi:10.1201/b19107-14.

45. Barata, C.; Ruela, M.; Francisco, M.; Mendonça, T.; Marques, J.S. Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Syst. J.* **2013**, *8*, 965–979. [CrossRef]