Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval

Stefano Allegretti*, Federico Bolelli*, Federico Pollastri*,
Sabrina Longhitano†, Giovanni Pellacani†, and Costantino Grana*

*Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
Email: {*name.surname*}@unimore.it

†Dipartimento Chirurgico, Medico, Odontoiatrico e di Scienze Morfologiche
con Interesse Trapiantologico, Oncologico e di Medicina Rigenerativa,
Università degli Studi di Modena e Reggio Emilia
Email: {*name.surname*}@unimore.it

***Abstract*—In recent years, many attempts have been dedicated to the creation of automated devices that could assist both expert and beginner dermatologists towards fast and early diagnosis of skin lesions. Tasks such as skin lesion classification and segmentation have been extensively addressed with deep learning algorithms, which in some cases reach a diagnostic accuracy comparable to that of expert physicians. However, the general lack of interpretability and reliability severely hinders the ability of those approaches to actually support dermatologists in the diagnosis process.**

In this paper a novel skin image retrieval system is presented, which exploits features extracted by Convolutional Neural Networks to gather similar images from a publicly available dataset, in order to assist the diagnosis process of both expert and novice practitioners. In the proposed framework, ResNet-50 is initially trained for the classification of dermoscopic images; then, the feature extraction part is isolated, and an embedding network is built on top of it. The embedding learns an alternative representation, which allows to check image similarity by means of a distance measure.

Experimental results reveal that the proposed method is able to select meaningful images, which can effectively boost the classification accuracy of human dermatologists.

I. Introduction

Skin cancer is one of the most common forms of human cancer worldwide [1].

Malignant melanoma is less common than basal and squamous cell skin carcinoma (it accounts for only about 3-4% of all skin cancers) but it is responsible for most of skin cancer deaths [1], [2], despite the existence of new therapeutic agents, such as checkpoint and BRAF inhibitors that improve survival of advanced cases [3]. However, Squamous Cell Carcinoma (SCC) can become lethal when it metastasizes, since few standardized and effective therapies for advanced SCC have been established. Although metastatic Basal Cell Carcinoma (BCC) is very rare, any delay in diagnosis may allow tumors to become unresectable [4]. Therefore, early detection of all skin cancers, not limited to melanoma, is required to prevent progression of these cancers to advanced stages and reduce skin cancer-related deaths [4], [5]. The clinical diagnosis of malignant melanoma is still difficult since the morphological characteristics of other pigmented skin lesions may sometimes mimic it. In fact, even in specialized centers, the melanoma diagnosis accuracy achieved with the unaided eye is slightly better than 60% [6], [7], making the early detection very hard to obtain. Nowadays, dermoscopy represents one of the most relevant imaging techniques for melanoma diagnosis. Dermoscopic images are obtained through a non-invasive in vivo examination based on a microscope that exploits an incident light and oil/gel immersion to make skin subsurface structures, dermoepidermal junction, and upper dermis accessible to visual examination [8], [9], [10].

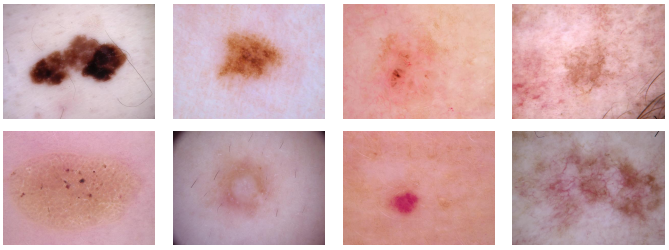

Fig. 1. Example images from the ISIC 2019 dataset, one for each class. From left to right: *melanoma*, *melanocytic nevus*, *basal cell carcinoma*, *actinic keratosis*, *benign keratosis*, *dermatofibroma*, *vascular lesion*, and *squamous cell carcinoma*.

Several approaches have been proposed to improve the diagnostic performance of clinicians: the ABCD rule [11], the CASH algorithm [12], Menzies method [13], 7-point checklist [14] or some other pattern classification methods [15] to distinguish between melanoma and non-melanoma pigmented skin lesions. However, becoming an experienced dermoscopic reader requires a significant time and training investment [4]. Moreover, even after such training, the readings are often complex and subjective.

To make readings more objective and qualitative, as well as support physicians using dermoscopy, many Computer-

Aided Diagnosis (CAD) systems for the automated melanoma recognition have been proposed [16], [17], [18], [19]. Among them, deep learning algorithms have revealed to be the most effective solutions. As a matter of fact, Convolutional Neural Networks (CNNs) are currently the cornerstone of medical image analysis [20], [21], [22], [23], [24].

Unfortunately, these approaches require huge amounts of data, which are hard to obtain and particularly expensive to annotate. However, since the first convolutional layers of CNNs learn to recognize simple elements like lines and colors, pre-training neural networks with existing collections of natural images [25] can mitigate the need for large annotated medical datasets [26]. Nevertheless, this approach can introduce biases towards certain characteristics and features. As an example, CNNs trained using ImageNet are strongly biased in recognizing textures rather than shapes [27].

With the aim of mitigating this problem, since 2016 the International Skin Imaging Collaboration (ISIC) has begun to aggregate a large-scale, publicly available dataset of dermoscopic skin lesions images (Fig. 1) and hosting multiple challenges and workshops [28]. The availability of this substantial amount of dermoscopic images allowed to significantly improve the performance of machine learning algorithms. A recent international diagnostic study demonstrates that the accuracy of machine learning algorithms for pigmented skin lesion classification proposed in the last years is comparable with those of expert dermatologists [29]. This is true considering only individual images and ignoring the clinical history of the patient: information that the dermatologist has often available when visiting.

Nevertheless, in the medical diagnosis field the goal should not be to take an action on behalf of an expert practitioner, but rather to assist his/her choice. Indeed, empirical experiments have shown that providing the physician with a second computerized and non-interactive opinion is not well seen by the dermatologists that even tends to change their own diagnoses when presented with a second computerized point of view [30].

In the last years, many attempts have been made to devise AI approaches, which are not only performing well, but are trustworthy, transparent, interpretable and explainable for a human expert [31]. Among others, the Grad-Cam technique [32] has proven to be a valid solution for dermoscopic images analysis: it is able to explain dermatologists on which parts of the input image the model is mostly focused, by simply providing a heat map [33].

This paper mainly aims at providing an interactive approach to effectively support dermatologists in the decision making process. Given a skin lesion, the designed deep learning-based algorithm will retrieve a number of other diagnosed cases that are similar to the original image, thus supporting and driving the dermatologist to a more precise diagnosis, without directly providing a second opinion. We thus propose a Content-Based Image Retrieval (CBIR) tool for searching and retrieving most similar dermoscopic images. A set of quantitative experiments shows that the proposed approach provides a better ranking than existing state-of-the-art solutions on the same topic. Moreover, to test the usefulness of the devised strategy, five dermatologists have been asked to classify different images with and without the output provided by our algorithm: experimental results clearly demonstrate the benefit of the proposal.

The rest of the paper is organized as follows. Section II describes the state-of-the-art dermoscopic image analysis solutions and content-based image retrieval systems available in literature. In Section III the proposed model is described and then it is evaluated from both a quantitative and qualitative point of view in Section IV. Finally, in Section V conclusion are drawn.

II. Related Work

Dermoscopic Images Analysis The extensive usage of dermoscopic images has been causing a wide interest in their computed analysis. Traditional computer vision techniques have focused on several tasks, from basic lesion segmentation [34] to the identification of meaningful patterns for the final classification [35]. The groundbreaking evolution of deep learning pushed the limits of automated dermoscopic images analysis, with great results especially on the segmentation and the complete classification tasks [33], [36]. Esteva *et al.* [23] compared the performance of CNNs to the accuracy of expert dermatologists, across two binary classification tasks of both clinical and dermoscopic images. The results of this study indicate that neural networks are able to classify skin cancer with a level of competence comparable to dermatologists.

The International Skin Imaging Collaboration (ISIC) has been playing a major role in the growth of skin lesion analysis. They have been collecting a large dataset of publicly available dermoscopic images [28], [37], [38], and have been hosting challenges since 2016; the 2019 challenge official training set counts 25 331 images. The work by Tschandl *et al.* [29] thoroughly compares results obtained by CNNs trained using the public ISIC dataset and the diagnosis carried out by numerous expert practitioners, proving that, when trained for the dermoscopic images classification task, neural networks obtain comparable results to human dermatologists. The authors also demonstrated the lack of reliability and resilience to out of distribution samples of these models.

Content-Based Image Retrieval In the work by Ballarini *et al.* [39], authors developed a content-based image retrieval system for a dataset of 533 images, divided into five classes of lesions, including two non-melanoma cancer types. Their system relies on visual features, such and color and composite texture, evolved using genetic algorithms. The similarity matching function is obtained by composing Bhattacharyya distance [40] for color covariance-based features, and Euclidean distance for texture features. In [41], Baldi *et al.* proposed another method based on low-level representative features. They find visually similar images to a query by means of a hierarchical multi-scale computation of the Bhattacharyya distance of all the database images.

Another CBIR system for skin lesions was developed by Jiji and Raj [42], who exploit features such as color, shape and texture, in order to find the most visually similar images to a query in a dataset composed of 20 different diseases. Their categories, however, do not include melanocytic nevi, and melanomas are not distinguished from other kinds of skin cancer. In 2016, Rahman *et al.* [43] proposed a decision support system based on a fusion between classification and retrieval. Their model considered high-level features such as Non-Subsampled Contourlet Transform (NSCT) and HOG based on Hessian matrix[1]. The model has been validated on the ISIC 2016 dataset, containing two skin lesion categories, melanoma and benign. Belattar *et al.* [44] used a kernelized SVM classification algorithm with an active learning technique and a histogram intersection matching in a retrieval system with relevance feedback for melanoma diagnosis. Their experiments were conducted on a dataset of melanomas and benign lesions. Finally, Pu *et al.* [45] applied the retrieval method proposed in [46] to the skin lesion domain. Their approach consists of training a deep network that learns to map images to hash codes, in a way that preserves similar semantics. Their network minimizes a classification loss function, with additional regulation terms to achieve desirable hash code properties. Then, similar images to a query are ranked by means of the Hamming distance, and clustered with Affinity Propagation (AP). This last proposal improves the measured precision with respect to previous methods based on hand-crafted features, and therefore represents a reference for comparison.

With this paper we present another content-based image retrieval system that requires no hand-crafted feature. Instead, we exploit CNNs' automatically learned filters, and we discuss three variations of a model trained with a loss function specific for the retrieval task, that allows us to overcome the limits of the classification loss and improve the retrieval performance with respect to state-of-the-art.

III. Image Retrieval

As stated, the aim of this paper is to find a model for Content-Based Image Retrieval (CBIR) for skin lesion medical images. Given a query, *i.e.* a new image of an unknown class, the task consists of retrieving k images from a labeled dataset, that are similar to the query. Since the final aim is to provide physicians with a valuable support for the classification task, two images are considered similar if they possess common features ascribable to a certain class of lesions. In order to extract such features, our approach to the problem consists of learning an embedding function f from skin lesion images space I into a compact Euclidean feature space $\mathbb{R}^d$, where distances correspond to a measure of similarity. In feature space, distance between images of the same class should be small, and distance between images of different classes should be large. After such a mapping has been established, the retrieval task can be accomplished by taking the k labeled images with the minimum distance from the query in the embedded space. In order to achieve this task, we built and evaluated four different variations of a deep convolutional neural network.

[1]Eig(Hess)-HOG

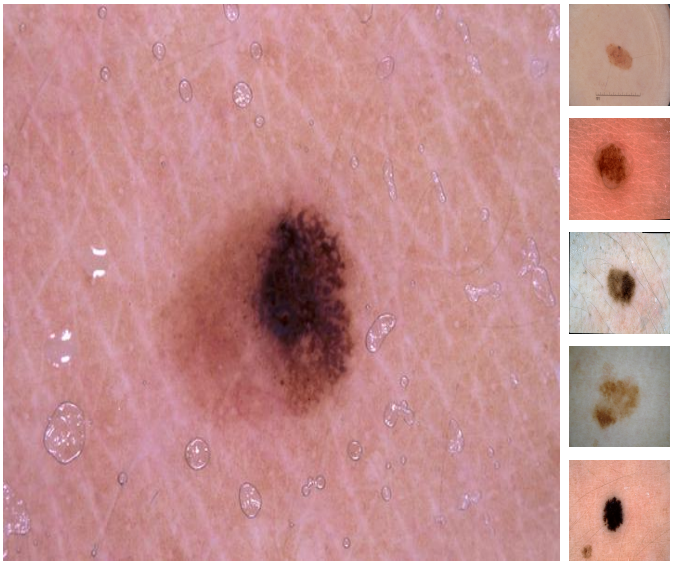

Fig. 2. Example result of the ranking system. On the left an example query image belonging to the melanocytic nevus class. On the left side the first five results in order of score from top to bottom. In this example all the output images are from the same class of the query.

For every model, input images are resized to 512×512, and each channel is normalized by subtracting the mean and dividing by the standard deviation. Moreover, training images undergo augmentation techniques to reduce overfitting, consisting in random flip and random rotation.

A. Classification

The first embedding model is part of a deep neural network trained for classification. We conducted experiments with multiple variations of ResNet (with 18, 34, 50, 101 and 152 layers), which have been proved to be effective solutions for the classification task of the ISIC challenge [28]. The loss used for this network is the cross-entropy loss. The network is trained with Stochastic Gradient Descent (SGD), with an initial learning rate of 0.001, lowered each time the evaluation accuracy reaches a plateau, with initial weights pretrained on the ImageNet dataset. After the training process, the embedding network is obtained by removing the last fully connected layer, which outputs the classification results. In this way, only the feature extraction part of the network is preserved, whose output is a vector of size 512 or 2048, depending on the ResNet variation used. This intermediate feature vector represents high-level visual features, strictly related to the image class, and represents the mapping in the new euclidean space. This simple model is mainly built for comparison with the similar proposal by Pu *et al.* [45], which also employs a classification network, and with the more complex models discussed further on.

B. Embedding End-to-End

The second model is a convolutional network trained end-to-end for the embedding task. Like the previous model, it is based on ResNet. However, instead of outputting class probabilities, the last fully connected layer, with output size d, represents the embedding in $\mathbb{R}^d$. The output vector is divided by its L2-norm; this regularization ensures that embedded vectors lie on the surface of the unit $(d-1)$-sphere, *i.e.*, $\|f(x)\|_2 = 1$.

The loss used for training is the triplet loss [47], that directly reflects the specific goal of this work. The triplet loss is calculated over a triplet of input samples composed of an anchor (a), a positive (p) and a negative (n). Sample a and p belong to the same class, while n belongs to a different class. Given a triplet, the loss is computed as:

$$\mathcal{L}(a,p,n) = max(d(f(a),f(p)) - d(f(a),f(n)) + \alpha, 0)$$

In the formula, $d(x,y)$ is the distance between x and y in feature space, and it can be calculated using a chosen metric, while α represents a desired margin between positive and negative pairs. The triplet loss aims at minimizing the distance from anchor to positive, and maximizing the distance from anchor to negative. We performed experiments with the euclidean distance and cosine distance (defined as 1 - cosine similarity), and we verified that, for this task, cosine distance yields better results. It is worth noticing that, if feature vectors lie on the unit $(d-1)$-sphere, cosine similarity is the same as the dot product.

The training process starts with the sampling of balanced minibatches, with n random elements per class, so that each skin lesion category is equally represented. For every minibatch, a network forward yields a corresponding list of embeddings. Then, triplets are chosen in the following way: *(i)* for every class, all possible pairs (a, p), *i.e.* combinations of two samples, are listed. Since there are n samples for each of the 8 classes, the total number of pairs is $\binom{n}{2} \times 8 = 4n(n-1)$; *(ii)* for every pair, a hard negative sample is chosen from the other 7 classes. A hard negative is a sample n for which $\mathcal{L}(a,p,n) > 0$.

Only hard negatives are useful for the training process, because they yield a non-zero loss. We experimented the three most common criteria for the choice of the hard negative:

- *random hard*: a hard negative randomly selected.
- *hardest*: the negative sample yielding the highest loss value.
- *random semihard*: a random negative n for which $d(f(a),f(p)) < d(f(a),f(n)) < d(f(a),f(p)) + \alpha$.

The *hardest* negative criterion leads to the fastest convergence, but it is also known to often lead to local minima [48], and the *random semihard* criterion is useful to reduce that risk. These observations hold true in our case, and we got the best results applying the *random semihard* criterion, even if the *random hard* choice proved to be nearly as effective. After the choice of negatives, the loss is computed for every triplet, and losses are averaged to get an aggregated value. This network is trained with SGD and initial learning rate of 0.01, decreased every 30 epochs.

C. Classification & Embedding

The third model taken into account is composed of both a classification network and a separate embedding network. It builds upon the first model, in the sense that, after the end of the training process, feature vectors extracted by the classification network constitute a new dataset, and they become the input of a fully connected network, that outputs the final embedding in d dimensions. This embedding network is composed of three fully connected layers of size 2048, 1024 and d respectively, randomly initialized. The first two layers are followed by a PReLU activation and a dropout layer; the whole architecture is depicted in Fig. 3. Like the previous case, the loss function is the online triplet loss. We trained this network with SGD, learning rate starting at 0.01 and lowered every 30 epochs, and batches of 30 samples per class.

D. Classification & Embedding End-to-End

The fourth model shares the same architecture of the third one: the difference is that, in this case, the feature extractor and the embedding network are not separated. Instead, after the training of the classification network, its final fully connected layer is substituted with the embedding network, and a new end-to-end training process starts, with online triplet loss. This means that the backward propagation does not stop at the first layer of the embedding network, as in Model-C, but instead reaches the beginning of the feature extractor. Because of the large size of the network, we are strongly limited in the batch size: in our tests, we used batches with 3 examples for each class. The whole architecture, that is the same as Model-C, is depicted in Fig. 3.

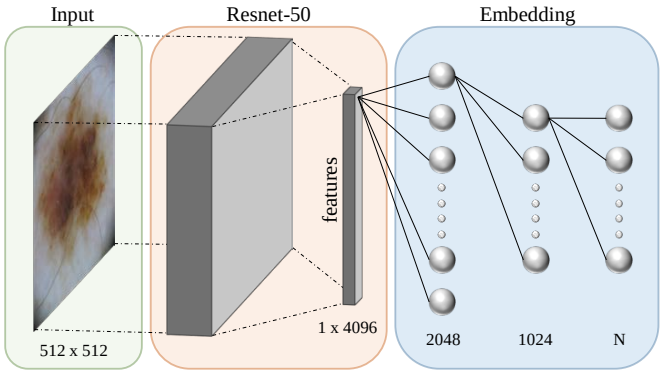

Fig. 3. Architecture of Model-C (Classification & Embedding) and D (Classification & Embedding End-to-End). The difference between the two approaches is that, after the training of the ResNet feature extractor with a classification loss, Model-D is fine-tuned end-to-end with triplet loss, while Model-C only updates weights on the embedding part of the network.

IV. Experimental Results

A. Quantitative

The dataset used for training and evaluating the proposed models is the ISIC 2019 archive [28], [37], [38], an international repository of dermoscopic images built for both clinical

TABLE I

AVERAGE PRECISION AT K MEASURED FOR EVERY MODEL ANALYZED, FOR THREE VALUES OF K, 1, 5, AND 10. CENTRAL COLUMNS REPORT AVERAGE VALUES SEPARATED FOR EACH CLASS, AND THE LAST COLUMN REPORTS THE BALANCED AVERAGE. NOVEL PROPOSALS ARE IDENTIFIED BY *.

Model	Cut-Off k	Per class P@k								AP@k
		MEL	NV	BCC	AK	BKL	DF	VASC	SCC	
Hash-AP [45]	-	0.4786	0.6111	0.5730	0.1896	0.1984	0.1505	0.3842	0.1375	0.3404
Hash-AP ResNet*	-	0.8176	0.7558	0.8509	0.7417	0.6256	0.7604	0.8271	0.6851	0.7580
Classification*	1	0.7840	0.9369	0.9347	0.7400	0.8300	0.7733	0.8667	0.7133	0.8224
	5	0.7262	0.9111	0.9029	0.7190	0.7724	0.7333	0.8373	0.6853	0.7859
	10	0.7040	0.9038	0.8957	0.7160	0.7470	0.7213	0.8307	0.6787	0.7746
Embedding End-to-End*	1	0.7400	0.9018	0.9133	0.6600	0.7520	0.7333	0.8267	0.7000	0.7784
	5	0.7314	0.8923	0.9005	0.6820	0.7576	0.7387	0.8240	0.7027	0.7786
	10	0.7322	0.8905	0.8993	0.6855	0.7572	0.7440	0.8253	0.7093	0.7804
Class & Embedding*	1	0.7490	0.9347	0.8973	0.7150	0.7700	0.8400	0.9067	0.7133	0.8157
	5	0.7542	0.9347	0.9013	0.7170	0.7768	0.8400	0.9013	0.7093	0.8168
	10	0.7531	0.9331	0.9032	0.7170	0.7814	0.8453	0.9040	0.7147	0.8190
Class & Embedding End-to-End*	1	0.7560	0.9022	0.9027	0.6600	0.7600	0.7733	0.8267	0.7133	0.7867
	5	0.7458	0.9030	0.8992	0.6770	0.7588	0.7707	0.8293	0.7213	0.7881
	10	0.7436	0.9012	0.9009	0.6830	0.7668	0.7760	0.8373	0.7273	0.7920

training and for supporting technical research by means of the international challenge of the same name, introduced in Section II. The dataset includes 25 331 images, divided into 8 categories: melanoma (MEL), melanocytic nevus (MN), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC) and squamous cell carcinoma (SCC). A priori probability of each class is reported in Table III. We split the dataset into training, validation and test sets, composed of 19 331, 1 000 and 5 000 images respectively. Every model is trained with the aim of minimizing the corresponding loss function on the training set. After each epoch, a partial evaluation of the model is performed by retrieving the k nearest neighbors taken from the training set of every sample in validation set, for three values of k (1, 5, and 10), and computing precision at k ($P@k$), *i.e.*, the number of neighbors that share the same class of the query, divided by k. The values for k have been chosen in a way that reflects the use case, that is showing a physician similar lesions to a new image, which needs to be classified. $P@k$ values computed on the validation set are used for hyperparameters optimization and early stopping of the training process. Then, final values of $P@k$ are computed by substituting validation set with test set.

In literature, another common metric used to evaluate the performance of retrieval systems is the recall, *i.e.* the ratio of relevant retrieved images to the total number of relevant images in the dataset. However, the recall is not meaningful for this use case, because the amount of retrieved images is orders of magnitude lower than the total number of dataset images belonging to each class.

As regards the choice of hyperparameters, experiments revealed that, independently of the model, the best value for α is 0.2, cosine distance is more effective than euclidean distance, and the L2 regularization actually improves the results. Model-dependent hyperparameters, instead, are reported in Table II. Feature space dimensionality is fixed for Model-A (*Classification*): it is equal to the size of the second to last layer of ResNet, which precedes the fully connected layer for classification. For the other models, we experimented values 3, 10, 30, 100, 300 and 1 000.

TABLE II

OPTIMAL HYPERPARAMETERS USED FOR EACH MODEL, FOUND WITH GRID SEARCH. FE, DIM, LOSS, LR AND BS RESPECTIVELY STAND FOR FEATURE EXTRACTOR, DIMENSIONS OF FEATURE SPACE, LOSS FUNCTION, LEARNING RATE AND BATCH SIZE.

Model	FE	Dim	Loss	Lr	BS
Classification	ResNet-50	2048	Cross-entropy	0.001	16
Embedding E2E	ResNet-34	300	Triplet	0.001	48
Class & Emb	ResNet-50	30	Triplet	0.01	240
Class & Emb E2E	ResNet-50	30	Triplet	0.01	24

It can be observed that the feature extractors used in the final models, ResNet-34 and ResNet-50, include a relatively small number of layers. We verified that deeper versions of ResNet, with 101 or 152 layers, are more effective for the classification task, *i.e.*, lead to higher accuracy; however, when measuring $P@k$, the results are not satisfactory. As regards models employing triplet loss, one of the reasons is that implementation details of the loss itself make large batches more efficient [48], and, under the same hardware, smaller networks allow for larger batches. Another consequence is that Model-C (*Classification & Embedding End-to-End*) reaches lower $AP@k$ than Model-D (*Classification & Embedding*): the latter, despite being a simpler model, can exploit a batch size larger by an order of magnitude.

For comparative reasons, we have implemented the method proposed in [45] and known as Hash-AP. To the best of our knowledge, it is the only skin lesion CBIR system that makes use of deep learning to extract visual features, and it represents the current state-of-the-art in the field. Hash-AP is

a classification network with an added hidden layer that learns a binary embedding, or hash, and retrieves similar images to a query measuring the Hamming distance of hash codes, then clustering the results and the query itself using affinity propagation. The original model is based on AlexNet [49]. In order to provide a fair comparison with our proposals, we also built a variation of Hash-AP based on ResNet, and we verified that, at least on the ISIC dataset, the retrieval performance notably improves w.r.t. the original model.

Final results are reported in Table I: $P@k$ is calculated for each class, and the last column reports the average precision at k ($AP@k$), which represents a summary of the model performance. $AP@k$ is the average value of per class $P@k$: in this way, every class is given the same weight, independently of its representativeness. Models Hash-AP and Hash-AP ResNet retrieve a variable number of images, which depends on the clustering result, and therefore only one value is reported for them. Novel proposals are marked by a star.

The precision measured for the original Hash-AP model is quite low. This result can be attributed to the classification network used, AlexNet, which fails to compete with more modern architectures, and to the difference between the ISIC dataset used in this work and the smaller dataset considered in the original paper. However, the ResNet-50 variation we implemented alongside the original one reaches an average precision of almost 0.76, proving the effectiveness of the idea. As regards models discussed in this paper, the *Classification* one yields the highest $AP@1$, being the only one to exceed 0.82. Nevertheless, for higher values of k, its performance is less satisfactory and, indeed, this network was not directly trained for image retrieval. The best performing model for higher values of k is Model-C, *Classification & Embedding*. As stated before, this result is partially related to the batch size used in training: Model-C has been trained with a batch size respectively 5 times and 10 times larger than that used for Model-B and D (Table II). A larger batch size means that, in the triplet selection process, more negatives are evaluated for the same positive pair, and thus there is a higher probability to find a *random semihard* negative. The improvement with respect to Hash-AP ResNet demonstrates the effectiveness of the triplet loss function, that is more suitable than a classification loss for the retrieval task.

B. Qualitative

In order to understand if our model can be effectively useful in supporting medical diagnosis, we asked five dermatologists to undergo a test, composed of two tasks. The first task simply consisted in classifying 100 dermoscopic images, randomly sampled from the ISIC test set, without any additional help. In the second task, the physicians were asked to classify the same images, but we also showed them the 5 nearest neighbors of each sample, labeled with their ground truth classes. The neighbors were found in the training set using Model-C. The tests were performed using exactly the same 100 images, with the aim of reducing the probability of a possible improvement being only imputable to the dataset variance. In order to mitigate bias in the second task, each image was rotated by 180°, and the list was shuffled. We measured the accuracy of the physician classification for both tasks. Average result is 67.4% for the first one (images only) and 76.6% for the second one (images and 5 nearest neighbors). Results obtained by individual dermatologists are shown in Table IV. Although a set of five dermatologists constitutes a relatively small statistical sample, an average improvement of 9.2% on the very same images proves the usefulness of our CBIR model as a support in the decision making process.

TABLE III
A PRIORI PROBABILITY OF EACH SKIN LESION CLASS.

Class	**Probability**
Melanoma	0.1785
Melanocytic Nevus	0.5083
Basal Cell Carcinoma	0.1312
Actinic Keratosis	0.0342
Benign Keratosis	0.1036
Dermatofibra	0.0094
Vascular Lesion	0.0100
Squamous Cell Carcinoma	0.0248

TABLE IV
ACCURACY OF 5 DERMATOLOGISTS ON TASK 1 AND TASK 2, CONSISTING IN CLASSIFYING 100 SKIN LESION IMAGES WITH AND WITHOUT THE SUPPORT OF OUR RETRIEVAL NETWORK.

	Task 1	**Task 2**
Dermatologist #1	75%	79%
Dermatologist #2	64%	80%
Dermatologist #3	69%	71%
Dermatologist #4	68%	82%
Dermatologist #5	61%	71%

V. CONCLUSION

In this paper we introduced a novel content-based image retrieval system for dermoscopic image analysis. The proposed system is based on learned features obtained through state-of-the-art convolutional neural networks.

An exhaustive experimental evaluation proved that the proposed solution outperforms competitors on the well known ISIC dataset. Moreover, the effectiveness and the usefulness of the devised solution have been demonstrated by a set of qualitative experiments carried out by a group of dermatologists.

These results prove that, although the existing knowledge of dermatologists is invaluable for the diagnosis of skin cancer, the proposed CBIR model can assist physicians with dermatologist-grade decision support.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] D. S. Rigel and J. A. Carucci, "Malignant melanoma: Prevention, early detection, and treatment in the 21st century," *CA: A Cancer Journal for Clinicians*, vol. 50, no. 4, pp. 215–236, 2000.

[3] K. K. Broman, L. A. Dossett, J. Sun, Z. Eroglu, and J. S. Zager, "Update on BRAF and MEK inhibition for treatment of melanoma in metastatic, unresectable, and adjuvant settings," *Expert Opinion on Drug Safety*, vol. 18, no. 5, pp. 381–392, 2019.

[4] Y. Fujisawa, S. Inoue, and Y. Nakamura, "The Possibility of Deep Learning-Based, Computer-Aided Skin Tumor Classifiers," in *Frontiers in Medicine*, 2019.

[5] M. E. Celebi, Q. Wen, H. Iyatomi, K. Shimizu, H. Zhou, and G. Schaefer, "A State-of-the-Art Survey on Lesion Border Detection in Dermoscopy Images," *Dermoscopy Image Analysis*, pp. 97–129, 2015.

[6] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The Lancet Oncology*, vol. 3, no. 3, pp. 159–165, 2002.

[7] H. Kittler, "Diagnostic accuracy of dermoscopy/dermatoscopy," *An Atlas of Dermoscopy*, p. 12, 2004.

[8] G. Argenziano, S. Puig, Z. Iris, F. Sera, R. Corona, M. Alsina, F. Barbato, C. Carrera, G. Ferrara, A. Guilabert *et al.*, "Dermoscopy Improves Accuracy of Primary Care Physicians to Triage Lesions Suggestive of Skin Cancer," *Journal of Clinical Oncology*, vol. 24, no. 12, pp. 1877–1882, 2006.

[9] M.-L. Bafounta, A. Beauchet, P. Aegerter, and P. Saiag, "Is Dermoscopy (Epiluminescence Microscopy) Useful for the Diagnosis of Melanoma? Results of a Meta-analysis Using Techniques Adapted to the Evaluation of Diagnostic Tests," *Archives of Dermatology*, vol. 137, no. 10, pp. 1343–1350, 2001.

[10] M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," *British Journal of Dermatology*, vol. 159, no. 3, pp. 669–676, 2008.

[11] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig, "The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions," *Journal of the American Academy of Dermatology*, vol. 30, no. 4, pp. 551–559, 1994.

[12] J. S. Henning, S. W. Dusza, S. Q. Wang, A. A. Marghoob, H. S. Rabinovitz, D. Polsky, and A. W. Kopf, "The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy," *Journal of the American Academy of Dermatology*, vol. 56, no. 1, pp. 45–52, 2007.

[13] S. W. Menzies, C. Ingvar, K. A. Crotty, and W. H. McCarthy, "Frequency and Morphologic Characteristics of Invasive Melanomas Lacking Specific Surface Microscopic Features," *Archives of Dermatology*, vol. 132, no. 10, pp. 1178–1182, 1996.

[14] R. H. Johr, "Dermoscopy: alternative melanocytic algorithms—the ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist," *Clinics in Dermatology*, vol. 20, no. 3, pp. 240–247, 2002.

[15] B. Rao and C. Ahn, "Dermatoscopy for Melanoma and Pigmented Lesions," *Dermatologic Clinics*, vol. 30, pp. 413–434, 2012.

[16] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007.

[17] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Transactions on Medical Imaging*, vol. 20, no. 3, pp. 233–239, 2001.

[18] P. Schmid-Saugeona, J. Guillodb, and J.-P. Thirana, "Towards a computer-aided diagnosis system for pigmented skin lesions," *Computerized Medical Imaging and Graphics*, vol. 27, no. 1, pp. 65–78, 2003.

[19] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Improving Skin Lesion Segmentation with Generative Adversarial Networks," in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2018, pp. 442–443.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[21] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Augmenting Data with GANs to Segment Melanoma Skin Lesions," *Multimedia Tools and Applications Journal*, vol. 79, no. 21-22, pp. 15 575–15 592, 2019.

[22] G. Ligabue, F. Pollastri, F. Fontana, M. Leonelli, L. Furci, S. Giovanella, G. Alfano, G. Cappelli, F. Testa, F. Bolelli, C. Grana, and R. Magistroni, "Evaluation of the Classification Accuracy of the Kidney Biopsy Direct Immunofluorescence through Convolutional Neural Networks," *Clinical Journal of the American Society of Nephrology*, vol. 15, no. 10, pp. 1445–1454, 2020.

[23] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[24] F. Pollastri, J. Maroñas, F. Bolelli, G. Ligabue, R. Paredes, R. Magistroni, and C. Grana, "Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[26] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. E. Hopcroft, "Convergent Learning: Do different neural networks learn the same representations?" in *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, 2015, pp. 196–212.

[27] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.

[28] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[29] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof *et al.*, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *The Lancet Oncology*, vol. 20, no. 7, pp. 938–947, 2019.

[30] S. Dreiseitl and M. Binder, "Do physicians value decision support? A look at the effect of decision support systems on physician opinion," *Artificial Intelligence in Medicine*, vol. 33, no. 1, pp. 25–30, 2005.

[31] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[33] L. Canalini, F. Pollastri, F. Bolelli, M. Cancilla, S. Allegretti, and C. Grana, "Skin Lesion Segmentation Ensemble with Diverse Training Strategies," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2019, pp. 89–101.

[34] R. Cucchiara and C. Grana, "Exploiting color and topological features for region segmentation with recursive fuzzy C-means," *Machine Graphics & Vision International Journal*, 2002.

[35] G. Pellacani, C. Grana, and S. Seidenari, "Algorithmic reproduction of asymmetry and border cut-off parameters according to the abcd rule for dermoscopy," *Journal of the European Academy of Dermatology and Venereology*, vol. 20, no. 10, pp. 1214–1219, 2006.

[36] A. R. Lopez, X. Giro-i Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *2017 13th IASTED international conference on biomedical engineering (BioMed)*. IEEE, 2017, pp. 49–54.

[37] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions," *Scientific Data*, vol. 5, 2018.

[38] M. Combalia, N. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, A. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic Lesions in the Wild," *arXiv preprint arXiv:1908.02288*, 2019.

[39] L. Ballerini, X. Li, R. B. Fisher, and J. Rees, "A Query-by-Example Content-Based Image Retrieval System of Non-melanoma Skin Lesions," in *MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support*. Springer, 2009, pp. 31–38.

[40] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.

[41] A. Baldi, R. Murace, E. Dragonetti, M. Manganaro, O. Guerra, S. Bizzi, and L. Galli, "Definition of an automated Content-Based Image Retrieval (CBIR) system for the comparison of dermoscopic images of pigmented skin lesions," *BioMedical Engineering OnLine*, vol. 8, no. 1, p. 18, 2009.

[42] W. Jiji and J. Durai Raj, "Content-based image retrieval in dermatology using intelligent technique," *IET Image Processing*, vol. 9, pp. 306–317, 2014.

[43] M. Rahman, N. Alpaslan, and P. Bhattacharya, "Developing a retrieval based diagnostic aid for automated melanoma recognition of dermoscopic images," in *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2016, pp. 1–7.

[44] K. Belattar, S. Mostefai, and A. Draa, "Intelligent Content-Based Dermoscopic Image Retrieval with Relevance Feedback for Computer-Aided Melanoma Diagnosis," *Journal of Information Technology Research*, vol. 10, pp. 85–108, 2017.

[45] X. Pu, Y. Li, H. Qiu, and Y. Sun, "Deep Semantics-Preserving Hashing Based Skin Lesion Image Retrieval," in *Advances in Neural Networks - ISNN 2017*. Springer, 2017, pp. 282–289.

[46] H. Yang, K. Lin, and C. Chen, "Supervised Learning of Semantics-Preserving Hash via Deep Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 437–451, 2018.

[47] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research - JMLR*, vol. 11, pp. 11–14, 2009.

[48] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[49] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, 2012.