Computers in Biology and Medicine 137 (2021) 104825

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Skin lesion image retrieval using transfer learning-based approach for query-driven distance recommendation

Walid Barhoumi [a,b,*], Afifa Khelifa [c]

[a] Université de Tunis El Manar, Institut Supérieur d'Informatique, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance (LIMTIC), 2 Rue Abou Rayhane Bayrouni, 2080, Ariana, Tunisia
[b] Université de Carthage, Ecole Nationale d'Ingénieurs de Carthage, 45 Rue des Entrepreneurs, 2035, Tunis-Carthage, Tunisia
[c] Higher Institute of Technological Studies of Mahdia, 5111, Hiboun, Mahdia, Tunisia

ARTICLE INFO

Keywords:
CBDLR
Deep-learned features
Transfer learning
Skin diseases
Similarity measure recommendation

ABSTRACT

Content-Based Dermatological Lesion Retrieval (CBDLR) systems retrieve similar skin lesion images, with a pathology-confirmed diagnosis, for a given query image of a skin lesion. By producing an intuitive support to both inexperienced and experienced dermatologists, the early diagnosis through CBDLR screening can significantly enhance the patients' survival, while reducing the treatment cost. To deal with this issue, a CBDLR system is proposed in this study. This system integrates a similarity measure recommender which allows a dynamic selection of the adequate distance metric for each query image. The main contributions of this work reside in (i) the adoption of deep-learned features according to their performances for the classification of skin lesions into seven classes; and (ii) the automatic generation of ground truth that was investigated within the framework of transfer learning in order to recommend the most appropriate distance for any new query image. The proposed CBDLR system has been exhaustively evaluated using the challenging ISIC2018 and ISIC2019 datasets, and the obtained results show that the proposed system can provide a useful aided-decision while offering superior performances. Indeed, it outperforms similar CBDLR systems that adopt standard distances by at least 9% in terms of *mAP@K*.

1. Introduction

Medical image repositories have been expanded in quantity, content and dimension thanks to the unprecedented advances in medical imaging devices, computers' performance and network transmission technology [1]. Thus, recent years have witnessed a significant interest in designing automated solutions with the aim of exploring large medical repositories effectively. In fact, Content-Based Medical Image Retrieval (CBMIR) consists in retrieving the most similar past cases, regarding a query image (*i.e.* a new image of an unknown class). By retrieving similar annotated clinical cases, CBMIR systems can assist practitioners in the clinical decision-making process. This can be very beneficial for them since they increase the degree of trust on the prediction made by theses CBMIR systems over time compared to traditional Computer-Aided Diagnosis (CAD) systems. In particular, for image-based skin cancer diagnosis, dermatologists can find out whether the lesion in the query skin image is benign or malign by analyzing some Common Imaging Signs (CISs) within the retrieved images [2]. Skin cancer, which is the most frequent type of cancer in the world, has risen remarkably during the last four decades. Examples of these cancers are melanoma and basal cell carcinoma which are caused by accumulated exposure to ultraviolet radiation, weakened immune systems, harmful chemical elements and sunlight during the winter months [3]. Thus, the development of effective tools for the automated retrieval of dermoscopy images has seen an impressive growth in the past years [4,5]. The common approach consists in extracting some useful features from the query image in order to retrieve images that have a similar set of features. Generally, feature extraction and similarity measure play a very important role in the success of Content-Based Dermatological Lesion Retrieval (CBDLR) systems. This makes feature extraction an extensively investigated topic that deals with the automated RGB-based description of images' contents [6]. Several feature

* Corresponding author. Université de Tunis El Manar, Institut Supérieur d'Informatique, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance (LIMTIC), 2 Rue Abou Rayhane Bayrouni, 2080, Ariana, Tunisia.
E-mail address: walid.barhoumi@enicarthage.rnu.tn (W. Barhoumi).

https://doi.org/10.1016/j.compbiomed.2021.104825
Received 15 April 2021; Received in revised form 11 August 2021; Accepted 29 August 2021
Available online 3 September 2021

extraction techniques have been proposed for indexing, annotation and retrieval purposes. These techniques can be classified into two main groups: hand-crafted features and deep-learned features. In fact, the first attempts, which are often referred to as hand-crafted techniques, have investigated low-level image attributes; such as color, texture, shape and spatial relations [7]. However, many modern techniques, which are often referred to as deep-learned techniques, have focused on deep learning architectures in order to automatically extract relevant features reflecting clinical signs of malignity [8,9]. Deep-learned features outperform hand-crafted ones, but they require large annotated datasets for an accurate training, and sometimes there is a problem of overfitting. On the other hand, the similarity measure is based on certain distance measurements between feature vectors, such that two dermatological images with shorter distance are considered more similar than images that are far away. Despite the large number of studies on similarity assessment [10], no measure can be said to be "perfect". The majority of CBDLR systems adopt one similarity metric to retrieve relevant images. However, while this metric can be suitable for many queries, it is sometimes not convenient for others. Similarly, certain measures that deemed to be inappropriate for most images, proved to be ideal for some individual queries.

These findings led us to investigate the dynamic selection of the adequate similarity measure according to the type of query. The main goal of the proposed method is to automatically select the appropriate similarity measure in a dynamic manner without having to standardize this choice for all queries. More precisely, being inspired by the recent success of deep learning techniques in medical imaging applications [11, 12], we propose to improve CBDLR performance by investigating deep-learned features while integrating a similarity measure recommender that is able to dynamically predict the appropriate distance metric for each query. The designed recommender, which is a model obtained by a transfer learning approach, aims to offer a more personalized distance metric to any query image. Thus, the contribution of this study is threefold. 1) Given that image classification and content-based image retrieval are slightly different variations of the same problem, convolutional neural networks have been adopted for learning the appropriate features for retrieving similar images to a query lesion while encoding higher level of semantics which are present in dermatological images. The selection of these features is conducted according to their performances for classifying skin lesions into 7 classes of dermatological diseases. 2) We construct a challenging ground truth that associates to each skin lesion image (among a dataset of 3, 207 images), the most appropriate distance metric (among 10 standard metrics) for retrieving the most similar cases, diagnostically speaking. 3) In the light of the advantages of transfer learning, an effective model is designed in order to recommend the most adequate distance to better identify clinically similar images for any skin lesion image. The returned images aim to make decision for the query lesion while facilitating the interpretation of results by dermatologists.

The rest of this paper is organized as follows. Section 2 presents a brief synthesis of the related work on CBDLR. The proposed method is described in Section 3. Section 4 summarizes the experimental results while Section 5 provides a summary discussion. Finally, Section 6 outlines a conclusion and some ideas for future research directions.

2. Related work

The performance of a CBDLR system depends on these two steps: feature extraction and image comparison. The first step consists in representing each image by a set of attributes and the second one lies in evaluating the similarity between the query image features and those of images composing the dataset.

2.1. Feature extraction

Features describe the visual properties of a dermoscopic image using low-level descriptors including color, texture, and shape properties of the skin lesion. These features can be extracted either directly from the images, based on deep learning techniques (deep-learned features) [13], or after some low-level procedures involving image pre-processing and lesion segmentation (hand-crafted features) [14]. Within the framework of hand-crafted features, although color features have shown to be important for primary skin efflorescence [15], most of the relevant feature extraction methods are based on the combination of the three types of features. For instance, many studies [15,16] have investigated the fusion of shape, color, and texture features using the Principal Component Analysis (ACP). Differently in Ref. [17], a multi-direction 3D Color-Texture Feature (CTF) has combined color and texture descriptors as a single CTF after processing the lesion component in multiple small windows. Within the framework of deep-learned features, several studies [18–20] have explored deep learning to investigate relevant skin image representations. The proposed tools can exhibit superior sensitivity and specificity in classification as well as in retrieval comparing to board-certified dermatologists. Tschandl et al. [21] compared the diagnosis accuracy obtained by CBDLR to retrieve similar images with corresponding disease labels against predictions made by a neural network. They demonstrated equivalent diagnosis accuracy in addition to the interpretability capability of CBDLR. Likewise, Codella et al. [22] have proposed a CBDLR system based on deep-learned features. This system has been trained in a collaborative way since human feedback about similarity of the lesions was considered. It has improved results both in terms of visual properties and diagnosis. In Ref. [13], several deep features have been combined using transfer learning and meta learning based on Multi-response Linear Regression (MLR). Generally, most recent works have focused on feature extraction using learned features, instead of traditional hand-crafted features. This is mainly due to their ability to effectively capture the intrinsic image patterns without hand-crafted feature design paradigms [13,23]. Deep learning models learn features automatically from raw data thanks to multiple layers that perform non-linear transformations. However, learned features are not yet sufficiently investigated within the context of skin lesion retrieval. This can be explained by the fact that learning discriminative feature representation is not sufficient due to the limited ability of the standard metric distances for matching similar (clinically speaking and not visually) skin lesion images [24] as well to the reduced sizes of current dermatological imaging ground truths. In our case, having seen that the success of many modern approaches slowly led to the belief that image retrieval and classification are just slightly different variations of the same problem [25], along with the availability of larger annotated dataset for skin cancer diagnosis [26], we adopted deep learning. We did this in order to select adequate features according to their performances for classifying skin lesions into 7 classes. In fact, a regularized discriminative method is proposed in order to learn a feature-based representation that allows the distinction between different classes. Since the performance of a CBDLR system is defined according to the similarity of the query-image class and those of the retrieved images, the proposed system can efficiently learn image features optimized for retrieval. The fact that deep-learned features are adopted, let us recall that standard distance metrics are not discriminative enough for measuring similarity between pairs of these learned features, since they are not able to capture the non-linear dependencies within the features.

2.2. Image comparison

Most of relevant CBDLR systems have adopted the approach of choosing a standard distance metric in order to retrieve the most similar lesions to the query one. For instance, similar images have been retrieved using a hierarchical multi-scale computation of the Bhattacharyya distance in Ref. [27] and using the Hamming distance in Ref. [28]. Differently in Ref. [29], images have been compared using the Bhattacharyya distance for color covariance-based features and the

Euclidean distance for texture features. In general, existing CBDLR systems choose the appropriate distance metric using prior knowledge of the domain or empirical experiments. For example, in Ref. [30], realized experiments showed that the Cosine distance is more effective than the Euclidean one. Similarly, four popular similarity measures were compared in Ref. [31] and the best results were recorded with the Mahalanobis distance. However, it is often difficult to find a distance metric that is well-suited to various real-world cases. Therefore, there has been, recently, a significant growth in the investigation of the Supervised Distance Metric Learning (SDML) that takes advantage of prior knowledge and experts' experience in order to automatically construct task-specific distance metrics. SDML methods aim to learn a distance metric using a set of equivalence/in-equivalence constraints between data, and most of existing studies have been dedicated to learning a Mahalanobis distance or a kernel function [32]. Major problems with these methods are the scalability respect to the dimension of data and the needed large number of constraints to avoid overfitting in high-dimensional SDML [33]. Although non-metric similarity learning tools have proved to be much more able to capture the different non-linearity features dependencies [24], their design imposes long optimization time as well as the availability of large amounts of annotated datasets [34]. Overall, the chosen distance metric, whether empirically among a set of distances or by means of SDML, can be appropriate for many queries but not necessarily for all the queries. Besides, some distance metrics that deemed inappropriate for most images can be ideal for particular queries. Thus, in order to circumvent the dependency of existing SDML techniques on experts' accuracy in labeling, we propose, in this study, a recommendation model, which omits the sensitivity to prior labels. Our idea consists in learning the most adequate distance metric, among the 10 most used metrics, to optimize the mapping of the selected features with the trained features. In fact, a training phase has been designed in order to minimize the distance between similar points while maximizing the distance between dissimilar points, given the ground truth that we have designed. Then, the proposed method dynamically selects the adequate distance for each query based on a transfer learning process. We self-organize the prior knowledge used for training the machine learning for modeling the selection of the appropriate distance metric for each query image. This allows to learn an appropriate distance metric for each query image instead of adopting a distance for the whole investigated dataset, which may ignore some important properties available in this dataset. The deep models can be trained and transformed effectively between different datasets by sharing and fine-tuning parameters thanks to transfer learning techniques [35], which have proved that the reuse of a pre-trained partial network, by including its structure and connection parameters, can significantly reduce the demand of training data and training time.

3. Proposed method

The flowchart of the proposed CBDLR method is composed of two main phases, namely offline indexing and online retrieval (Fig. 1). In the offline phase, features are extracted from each image, within the investigated large dataset, in order to create a feature database (also referred to as signature image database). In the online phase, features are extracted from the query image, by using the same feature extraction procedure adopted in the offline phase, in order to estimate the similarity between the query image and each image within the dataset. Images having high similarity are ranked and then displayed to the user as relevant retrieval results. The proposed method is completely based on deep learning and particularly on Convolutional Neural Networks (CNN) which are used to extract image features as well as to design a similarity measure recommender. In fact, during the offline phase, a CNN classifier is firstly trained in order to be used in the second step as feature extractor. Moreover, an automatic ground truth generation procedure is performed in order to create a labeled dataset where each lesion image, within the input dataset, is annotated by the best similarity measure dedicated from query simulations and results' evaluation. Given the ground truth, a CNN-based distance recommender is trained thereafter in order to learn the best distance per image. In the online phase, the proposed CBDLR method responds to the user's query by visualizing similar images identified while applying the dynamically predicted distance by the similarity recommender. It is worth mentioning that the typical architecture of CBDLR has been personalized by proposing a generic method, which is an end-to-end solution formed by consecutive steps and divided into the two classical phases (offline

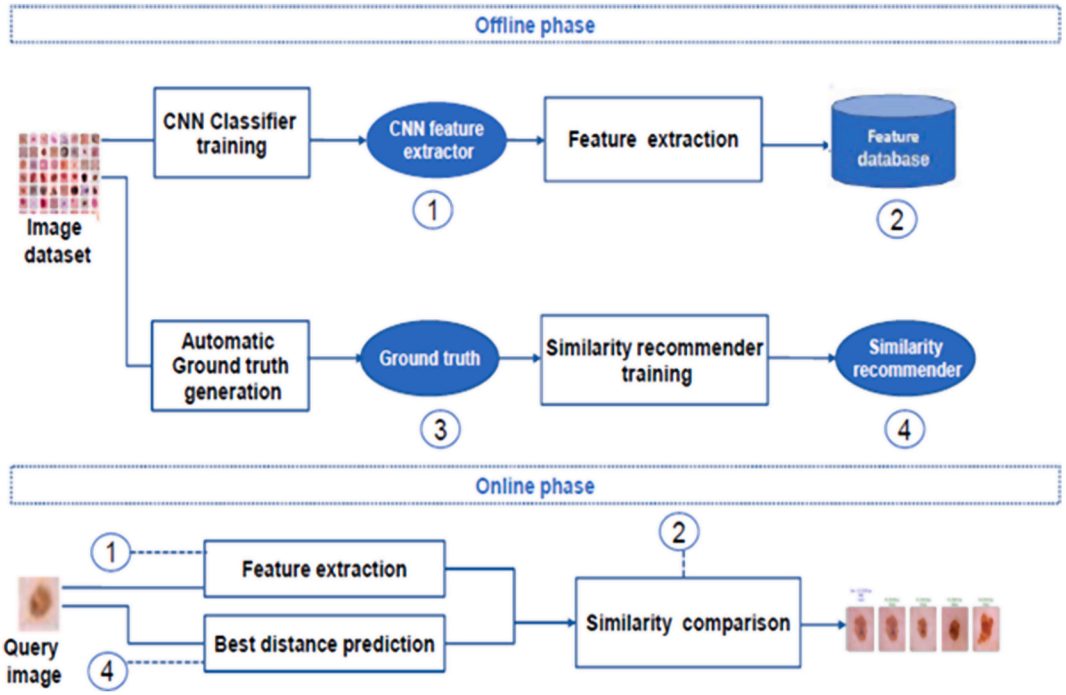

Fig. 1. Flowchart of the proposed CBDLR method.

and online). Indeed, the proposed three-stage architecture is a modified version of the typical scheme of a CBDLR, since it incorporates a distance recommender. In the first stage, the CNN is adopted in order to learn about the appropriate features which allow the retrieval of similar images while encoding higher level of semantics present in dermatological images. The selection of these features is conducted according to their performances in the classification of skin lesions into various classes of dermatological diseases. The output of this stage is the database of the relative features to images composing the investigated dataset. In the second stage, the ground truth associating the best distance metric for retrieving the similar images for each skin lesion image among the ISIC2018 dataset [36,37], is constructed.

In the third stage, the constructed ground truth dataset serves as input to the CNN-based transfer learning procedure to recommend the appropriate distance. In fact, the extracted features are used to dynamically make prediction, for any new image, on the best distance metric allowing to retrieve the most similar images to this query image from a clinical point of view. The proposed architecture is generic enough to be applied to any medical context where images annotated with diagnosis classes are available.

3.1. Feature extraction

The proposed CBDLR method is based on deep learning and more particularly on CNN-based transfer learning for the feature extraction procedure in order to avoid the drawbacks associated with hand-crafted feature extraction. In fact, hand-crafted features often require expensive human labor, rely on expert knowledge and do not generalize well for large datasets [13]. Thus, using deep learning to automatically extract features from raw data has become a promising approach that attempts to address these drawbacks. The main power of CNN lies in the exploitation of spatial correlation in image data aiming to extract a set of discriminating features at multiple levels of abstraction. In our case, the aspect of transferability of knowledge embedded in the pre-trained CNN [38] has been opted. Indeed, an important alternative to training a CNN from scratch is to fine-tune a CNN that has been trained using a large labeled dataset from a different application. This would allow, above all, to avoid a time-consuming training process [13]. The pre-trained models have been applied magnificently to various applications as a feature extractor or as a starting point for transfer learning from one task to another when the fully-connected layers of a pre-trained CNN are substituted with new specific layers. Then, the labeled data are used to train only the appended layers while keeping the rest of the network unchanged. In fact, transfer learning is adopted in order to modify the pre-trained architectures by adding new specific layers, and then train the model within the target dataset which is annotated using a defined number of skin diagnosis classes. The real challenge resides in the availability of large datasets that help researchers to train their models in order to surmount the visual variability of skin lesions as well as the subtlety of the signs that differentiate benign and malignant cases [39]. Many datasets; such as the Atlas of Dermoscopy [40], the PH2 [41] and the Dermofit Image Library have been proposed in the literature, but the number of offered images remains really limited. Currently, the largest publicly available dataset is the International Skin Imaging Collaboration (ISIC) archive, which provides annotated images from different sources and devices. Since the first ISIC Challenge in 2016 [42], this dataset has been increasing in size and in the amount of information provided for each lesion. This has allowed to aggregate a large-scale publicly available image dataset of dermoscopic skin lesions and to host multiple challenges and workshops [20] which in turn allowed to significantly improve the performance of machine learning algorithms. In our case, the ISIC 2018 challenge dataset, which contains 10, 015 labeled images, has been investigated. All images are annotated with one class of skin lesion from seven types of diseases: Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Keratinocytic tumors (AKIEC), Benign keratosis (BKL), Dermatofibroma (DF), and Vascular lesions (VASC). The disproportion between classes is very high, particularly for the Dermatofibroma and the Vascular lesions classes, and most of the lesions belong to the Melanocytic nevus class (Table 1). In order to reduce overfitting and knowing that two images are considered similar if they possess common features ascribable to a certain class of lesions, downsampling from majority classes; like Melanocytic nevus, Melanoma and Benign keratosis; has been applied. Besides, minority classes; such as Vascular lesion, Dermatofibroma and Keratinocytic tumors; have been upsampled based on data augmentation by random flips and random rotations. The final training dataset is composed of 5057 images that are relatively balanced between the 7 classes (Fig. 2).

Furthermore, training the classifier to be used for feature extraction was performed by ResNet50 pre-trained architecture. This architecture achieved excellent results, within the framework of pigmented skin lesion classification, which are comparable to those of expert dermatologists [30]. The classifier (Fig. 3) is composed of the lower and the mid layers of the ResNet50 architecture. Besides, five layers are appended at the top of the model in order to enhance learning. The last layer is a fully-connected layer composed of 7 nodes corresponding to the skin disease classes. During 20 epochs, the 7-class classifier, which is composed of 26, 742, 663 parameters, was trained by using the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01 and 20% of images for the validation set. The multilayered hierarchical structure of ResNet50 gives it the ability to extract low, mid, and high-level features [43], which can effectively boost the classification accuracy of dermatologists. It is worth noting that we have tested many optimizers (*i.e.* SGD, Adam, RMSProp …) and we have found that SGD improves the overall performance. For instance, the accuracy, the precision and the recall of the ResNet50 classifier trained with Adam optimizer [44] are: 78%, 86% and 66%, respectively (best configuration of hyperparameters: $Lr = 0.001$, learning rate decay $= 0.5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $epochs = 20$). In fact, Adam optimizer converges faster than SGD but it has weaker generalization capacity. Precision (1), Recall (2) and F1-Score (3) metrics; which are based on *TP* (True Positive), *FP* (False Positive), *TN* (True Negative) and *FN* (False Negative) rates; have been used in order to evaluate the performance of the trained model (Table 2). Recorded results show that the trained classifier achieves good performance with an average $F1 - Score$ of 96%, an average precision rate of 95% and an average recall rate of 96%. The highest precision rate has been recorded for the Melanocytic Nevus class (= 98%) whilst the highest recall rate has been obtained for the Vascular lesion class (= 99%). Each layer of the ResNet50 architecture produces an activation for the given image. Besides, first layers capture primitive features like blobs, edges, and colors that are abstracted by the deeper layers to form higher level features and more affluent image representation [45]. Therefore, based on empirical experiments, the optimal layer of the model (the global average pooling layer) has been selected. It permits to obtain a discriminate feature vector ($size = 1024$), while allowing the input image to propagate forward, stopping at optimal layer, and taking the output of that layer as feature vector learned by the ResNet50 model. The feature vector obtained is used as a representation of the input skin lesion image and is stored in the dedicated feature

Table 1
Number (#) of images per class within the investigated ISIC2018 dataset.

Disease	Malignity	# of images
Keratinocytic tumors (AKIEC)	Malign	327
Basal cell carcinoma (BCC)	Malign	514
Benign keratosis (BKL)	Benign	1099
Dermatofibroma (DF)	Benign	115
Melanoma (MEL)	Malign	1113
Melanocytic nevus (NV)	Benign	6705
Vascular lesion (VASC)	Benign	142
Total		10015

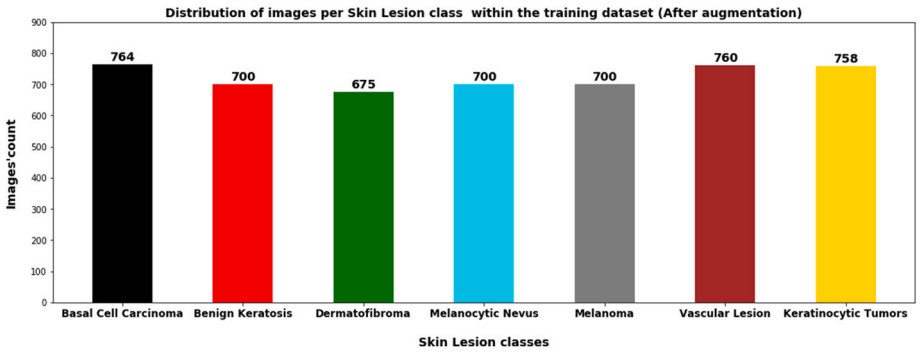

Fig. 2. Distribution of images per class within the training dataset.

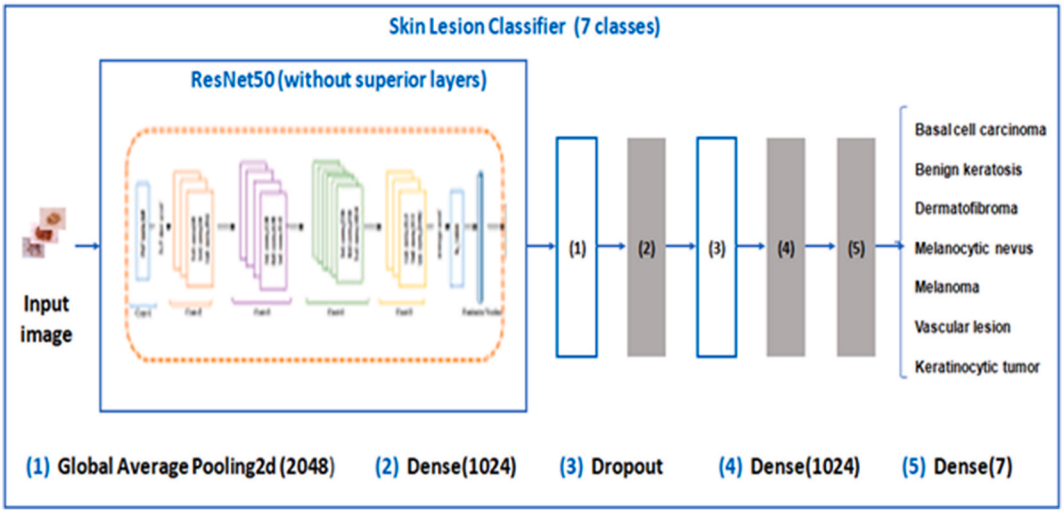

Figure 3. Architecture of the adopted ResNet50 classifier for training the feature extractor: five layers are added including three dense ones (represented by gray color).

Table 2
Classification performances of the trained feature extraction model.

Disease	Precision	Recall	F1-Score	# of images
BCC	0.97	0.97	0.97	514
BKL	0.95	0.95	0.95	700
DF	0.93	0.98	0.96	115
MEL	0.94	0.95	0.94	700
NV	0.98	0.96	0.97	707
VASC	0.95	0.99	0.97	142
AKIEC	0.94	0.92	0.93	329
Average	0.95	0.96	0.96	458.15

database ($size = 3200$).

$$Precision = \frac{TP}{TP + FP}. \quad (1)$$

$$Recall = \frac{TP}{TP + FN}. \quad (2)$$

$$F1 - Score = 2.\frac{Precision.Recall}{Precision + Recall}. \quad (3)$$

3.2. Image comparison

Inspired by the recent success of deep learning techniques in medical imaging context, we propose to integrate a similarity measure recommender. This recommender dynamically chooses the adequate distance to retrieve similar images for any new query image of a skin lesion. Thus, it avoids making an overall choice of the appropriate distance metric for all types of queries. This can be considered as a Case-Based Reasoning (CBR), which is an effective health science paradigm that is inspired by human reasoning. In fact, choosing the adequate measure by query can improve CBDLR performance and user confidence, since dermatologists usually solve a new problem by applying previous experiences to the current situation. The CBR is an adequate approach to explore within the skin lesion retrieval context, where symptoms represent the problem and diagnosis and treatment represent the solution [15]. The similarity recommender is a CNN distance classifier that is trained by using an automatic ground truth that we generated. In fact, since the similarity recommender aims to dynamically select the best distance for each query image, this similarity classifier has to be trained by using a skin lesion dataset annotated by the best distance according to each image. Since there is no work in the literature that provides a study on the similarity measures evaluated in the context of skin diseases, as best as we know, an automated procedure has been designed in order to generate a ground truth for training the similarity recommender. The generated dataset is composed of the couples (image, best distance metric) using the KNN algorithm for retrieving the K most similar images to a query. The result retrieved per distance is evaluated and the best distance per image is identified. Indeed, the procedure of automatic generation of the ground truth is based on two basic steps. In the first step, the execution of query images was simulated using a set of distances, and the results obtained per image and distance were stored before the evaluation of the retrieval accuracy. More precisely, for each

query image and for each distance metric, the algorithm simulates the execution using the trained CNN model to extract features and the KNN algorithm to retrieve the K most similar images. Then, the retrieved result for each pair (image, distance metric) is evaluated using the following score:

$$Score = \frac{1}{5}\sum_{k=1,3,5,8,10}\frac{Precision@k}{k}. \quad (4)$$

By using this score; which is the average of the weighted precisions for the most similar 1, 3, 5, 8 and 10 retrieved images; the result is considered more accurate if it presents good results at first positions. The second step is based on the previous evaluation step and it uses scores already calculated in order to associate the best distance metric for each query image. In the case where an image has the same best score for two distances, the distance with the least instances is selected in order to ensure relatively balanced classes. The output of this step is a dataset composed of images annotated with the most appropriate distance for retrieving similar images, from a clinical point of view. In our implementation, 10 standard similarity metrics have been used. These metrics are listed in Table 3 (where P and Q denote the n-dimensional feature vectors of two skin lesion images), and the numbers of associated images for each distance are shown in Fig. 4. It is clear that each distance, among the investigated ones, is best for at least 211 images.

Once the ground truth is built, the similarity recommender is trained. Given an objectively annotated dataset by the best distance measure, the CNN architecture presented in Fig. 5 is used to train the distance classifier. Thanks to transfer learning, we fine-tuned ResNet50 pre-trained model by removing upper layers and adding five layers in order to boost learning, where the last one is fully-connected with 10 nodes corresponding to similarity metrics studied within the ground truth. During 30 epochs, the 10-class classifier, which is composed of 26, 745, 738 parameters, was trained by using the SGD optimizer with an initial learning rate of 0.01, and 10% of images were used as validation set. Performance metrics of the designed recommender confirm the accuracy of the trained distance recommender (Table 4). During the online phase, the CBDLR system responds to the user query image by visualizing the most similar images. These images are identified by comparing features extracted by the same procedure of the offline phase, while using the best distance predicted by the similarity recommender.

4. Experimental results

Since performances of CBDLR systems are strongly related to the ranking of relevant cases retrieved for each query [2], the proposed method aims to learn the appropriate distance leading to pertinent results. The CNN-based model permits to effectively address the problem of more relevant and less relevant differences among the images in order to obtain human-like performance. In Fig. 6, a sample of retrieval results of the proposed CBDLR system is shown for the 7 classes of diseases within the ISIC2018 dataset (each row illustrates the query image followed by its most 10 similar images). Since the same features (color or differential structure) may be present in multiple disease classes of skin lesions but with different morphology or area covered by that information, misclassification is possible especially for some skin lesions with their known mimics. This occurs particularly for images which are "visually" similar to the query image while they "diagnostically" belong to different lesion classes. For instance, in the first (*resp.* fifth) row of Fig. 6, the proposed CBDLR system visualizes 9 similar images belonging to the BCC (*resp.* MEL) class and one image, which is visually similar, but it corresponds to the NV (*resp.* BKL) class. This could be very useful as an assisting tool for providing an intuitive support to less-experienced dermatologists as well as a learning tool to help students discover common aspects of lesions. Nevertheless, it could be interesting to enhance the accuracy of the feature extractor by adding more samples representing the confusing cases to the training dataset. This would reduce misclassification cases and avoid confusion between lesions having same visual signs. Furthermore, in order to quantitatively evaluate the retrieval performance of the proposed system, the pertinence of the returned images has been considered through two commonly used metrics: $P@K$ (5) and $mAP@K$ (6). In fact, $P@K$ is the precision at K, which corresponds to the number of relevant images among the top K images retrieved, and $mAP@K$ refers to the mean average precision at K, which is sensitive to the entire ranking of returned results and which contains both precision and recall aspects. Thus, mAP is the mean of the Average Precision (AP) over the N evaluated queries, and $mAP@K$ is defined based on $AP@K$ (7) that considers the order in which the returned images are presented. Moreover, the diagnosis performance of the proposed CBDLR system has been evaluated in terms of *Precision* (1), *Recall* (2) and $F1 - Score$ (3).

Table 3
Investigated distance metrics.

Distance	Formula
Euclidean	$d_{Euc}(P,Q) = \sqrt{\sum_{i=1}^{n}\lvert P_i - Q_i\rvert^2}$
Cityblock (Manhattan)	$d_{city}(P,Q) = \sum_{i=1}^{n}\lvert P_i - Q_i\rvert$
Minkowski	$d_{MK}(P,Q) = \sqrt[p]{\sum_{i=1}^{n}(P_i - Q_i)^p}$
Chebyshev	$d_{Cheb}(P,Q) = \max_{i=1}^{n}\lvert P_i - Q_i\rvert$
Canberra	$d_{Canb}(P,Q) = \sum_{i=1}^{n}\frac{\lvert P_i - Q_i\rvert}{\lvert P_i\rvert + \lvert Q_i\rvert}$
Braycurtis	$d_{Bray}(P,Q) = \frac{\sum_{i=1}^{n}\lvert P_i - Q_i\rvert}{\sum_{i=1}^{n}(P_i + Q_i)}$
Hamming	$d_H(P,Q) = \sum_{i=1}^{n}\lvert P_i - Q_i\rvert$
Cosine	$d_{Cos}(P,Q) = \frac{\sum_{i=1}^{n}P_iQ_i}{\sqrt{\sum_{i=1}^{n}(P_i)^2}\sqrt{\sum_{i=1}^{n}(Q_i)^2}}$
Correlation	$d_{Corr}(P,Q) = 1 - \frac{\sum_{i=1}^{n}(P_i - \overline{P_i})(Q_i - \overline{Q_i})}{\sqrt{\sum_{i=1}^{n}(P_i - \overline{P_i})^2(Q_i - \overline{Q_i})^2}}$
Chi2	$d_{Chi2}(P,Q) = \frac{1}{2}\sum_{i=1}^{n}\frac{(P_i - Q_i)^2}{(P_i + Q_i)}$

$$P@K(q) = \frac{\sum_{i=1}^{K}Rel(q,i)}{K}, \quad (5)$$

where, $Rel(q, i)$ denotes the relevancy between the query image q and the ith returned image.

$$mAP@K = \frac{\sum_{i=1}^{N}AP@k(q_i)}{N}. \quad (6)$$

$$AP@K(q) = \frac{\sum_{i=1}^{K}P@i(q) * Rel(q,i)}{Nq}, \quad (7)$$

where, Nq is the number of relevant images among the K retrieved images.

The used test dataset for the performance assessment is composed of 210 images randomly selected form the ISIC2018 dataset and equitably distributed among the 7 classes. The obtained results (Tables 5 and 6 and Fig. 7) prove that the integration of the similarity recommender provides higher retrieval performance compared to all other distances. In fact, the proposed method allows to predict the best similarity measure for an input image that leads to the improvement of the retrieval performance. For instance, the $P@K$ metric reaches 96.6%, 86.6%, 77.4%, 68.5% and 64% for 1, 3, 5, 8 and 10 retrieved images, respectively (Table 5). The Cityblock distance has the highest precision in the case of the 5 retrieved images ($P@5 = 70.1\%$) whilst the Correlation distance has the highest precision in the case of the 10 images ($P@10 = 58.8\%$). Meanwhile, the lowest precision rates were recorded with the Hamming distance: 71.4%, 49.9%, 44.4%, 40.5% and 39% for $P@1$, $P@3$, $P@5$, $P@8$ and $P@10$, respectively. This same distance records the highest precision for query images illustrating the Melanocytic Nevus class with 100%, 88.8%, 88.6%, 83.75% and 83% for $P@1$, $P@3$, $P@5$, $P@8$ and $P@10$, respectively (Fig. 7). This empirical finding supports our idea of adopting a dynamic selection of the similarity measure according to the

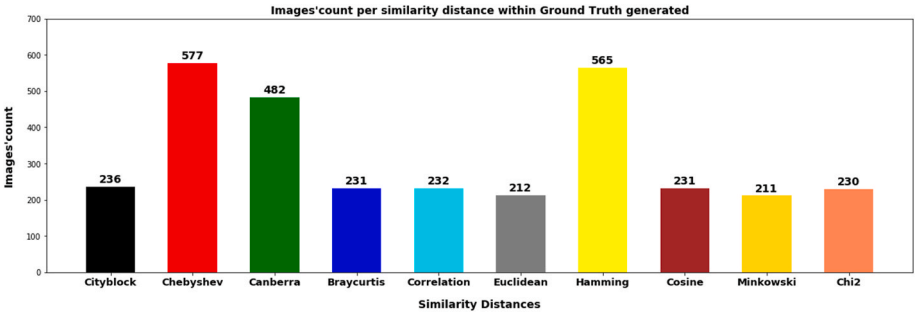

Fig. 4. Number of images per distance within the generated ground truth.

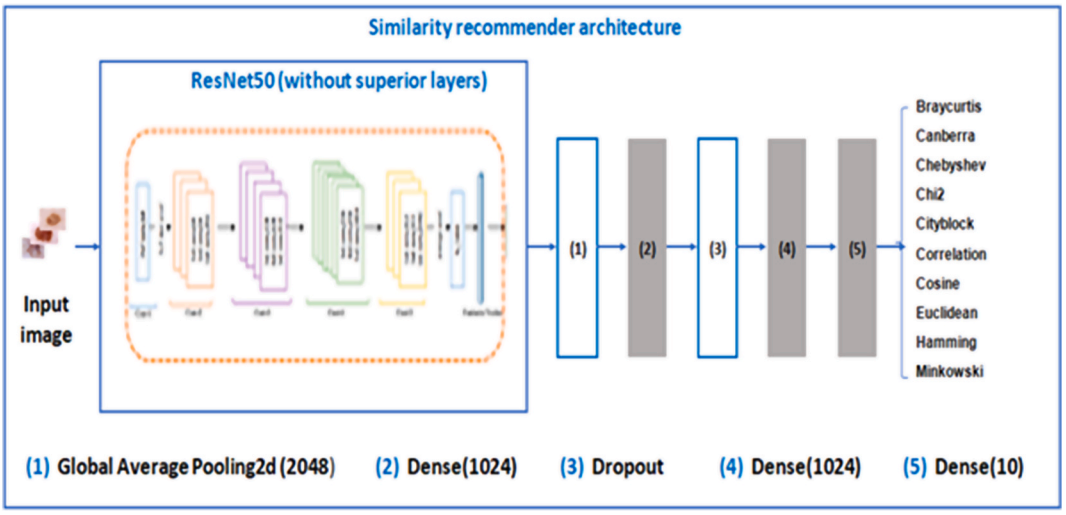

Fig. 5. Similarity recommender architecture: five layers, including three dense ones (represented by gray color), are added.

Table 4
Classification performances of the trained similarity recommender model.

Distance	Precision	Recall	F1-Score	# of images
Braycurtis	0.97	0.92	0.94	231
Canberra	0.88	0.93	0.91	482
Chebyshev	0.93	0.92	0.92	577
Chi2	0.92	0.90	0.91	230
Cityblock	0.94	0.93	0.93	236
Correlation	0.94	0.91	0.92	232
Cosine	0.98	0.90	0.94	231
Euclidean	0.93	0.92	0.93	212
Hamming	0.90	0.95	0.92	565
Minkowski	0.94	0.91	0.93	211
Average	0.93	0.92	0.93	320.7

query image. Fig. 7 shows the *P@K* rate, for various values of *K*, of the proposed CBDLR method against the ones of the investigated distance metrics, while using the same procedure for feature extraction. Theses rates were recorded for each of the 7 classes of dermatological diseases. Best retrieval performances were achieved for Melanocytic Nevus and Vascular lesion, where the *P*@5 rate exceeds 90%. Most images returned belong to the same class of malignancy (Table 1), even if they do not belong to the same class of disease. Considering that a retrieved image is relevant if it belongs to the same malignancy type of the query image, the retrieval accuracy increases and *mAP*@10 exceeds 80% for all classes. Table 6 confirms the same finding and shows the superiority of the Proposed Method (PM) that performs the highest *mAP@K* values, when compared to all other distances. In fact, PM overtakes Cityblock, which provides the best *mAP*@5 (61.9%) and *mAP*@10 (48.2%) rates, by at least 9%. Furthermore, since the main goal of a CBDLR system is to assist practitioners in the diagnosis process through similar images provided in relation to a case, an evaluation of CBDLR diagnosis performance (through the three metrics of *Precision*, *Recall* and *F*1 − *Score*) has been performed based on frequencies classes of the five first images retrieved. This is because clinicians pay most attention on the top 5 images retrieved for the diagnosis task [31]. The obtained results (Table 7) confirm again that the proposed CBDLR method provides dermatologists with a valuable support for the diagnosis stage. Besides, in order to compare the classification model with and without CBDLR, the diagnosis performance of the ResNet50 classifier, which was already trained in the first step of the proposed method, was evaluated by using the same dataset composed of 210 images (Table 8). The automated classification provided by the ResNet50 classifier is slightly more accurate than the CBDLR results. However, contrary to automated classification tools, which are usually considered as black boxes with no explanation of how they form their predictions, CBDLR systems can retrieve a number of other diagnosed cases that are similar to the query image. This supports the dermatologist to perform a more precise diagnosis without directly providing a second opinion. Overall, the performed evaluations show clearly that the retrieval performance increases significantly by integrating the designed similarity measure recommender that dynamically selects the adequate distance corresponding to a query image. To explain the obtained results, it is enough to take the case of the Hamming distance which has the minimum performance in most evaluations shown in Fig. 7. However, it is the best distance in 565 cases according to the automatically generated ground truth (Fig. 4). Thus, the proposed ResNet50-based similarity learning approach permits to avoid a global choice of a similarity measure that can penalize individual cases and therefore affect the overall

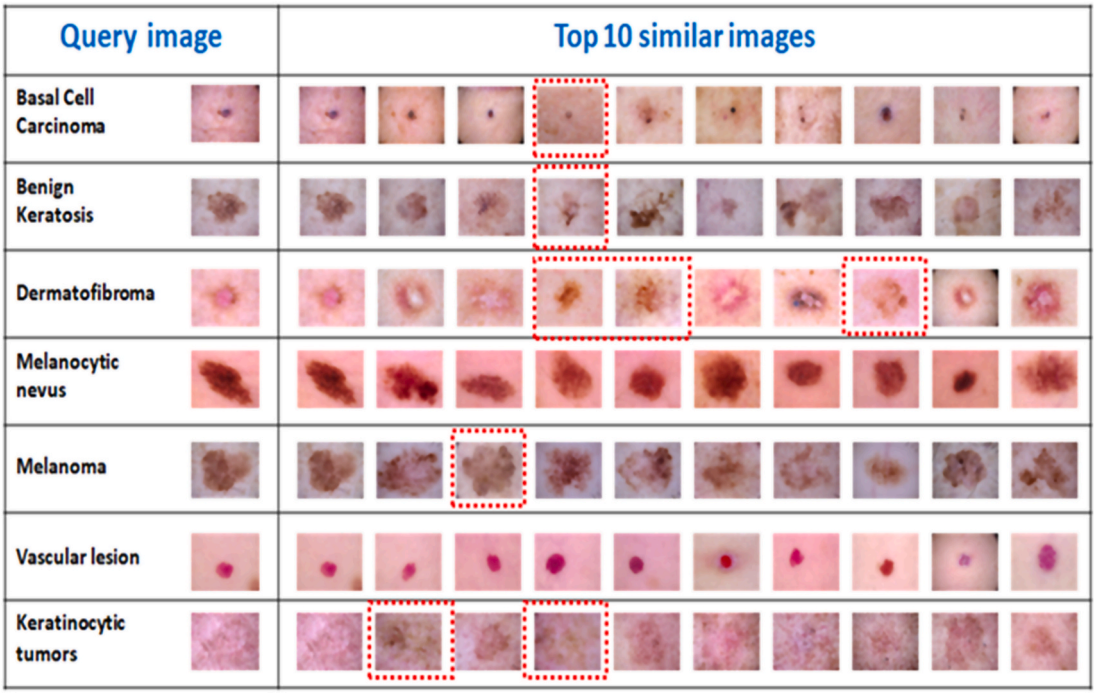

Fig. 6. A sample of the retrieval results for 7 query images (one query per class of diseases): first row: Basal cell carcinoma (*Precision* = 90%); second row: Benign keratosis (*Precision* = 90%); third row: Dermatofibroma (*Precision* = 70%); fourth row: Melanocytic nevus (*Precision* = 100%); fifth row: Melanoma (*Precision* = 90%); sixth row: Vascular lesion (*Precision* = 100%); seventh row: Keratinocytic tumors (*Precision* = 80%).

Table 5
Performance evaluation of the retrieval results using the *P@K* metric.

Distance	P@1	P@3	P@5	P@8	P@10
Braycurtis	0.871	0.719	0.656	0.595	0.571
Canberra	0.814	0.704	0.652	0.582	0.553
Chebyshev	0.809	0.646	0.589	0.531	0.503
Chi2	0.871	0.760	0.679	0.615	0.582
Cityblock	0.876	0.753	0.701	0.622	0.583
Correlation	0.852	0.734	0.673	0.610	0.588
Cosine	0.852	0.722	0.671	0.607	0.585
Euclidean	0.857	0.723	0.672	0.605	0.574
Hamming	0.714	0.499	0.444	0.405	0.390
Minkowski	0.857	0.723	0.672	0.605	0.574
PM	**0.966**	**0.866**	**0.774**	**0.685**	**0.64**

Table 6
Performance evaluation of the retrieval results using the *mAP@K* metric.

Distance	mAP@1	mAP@3	mAP@5	mAP@8	mAP@10
Braycurtis	0.871	0.678	0.582	0.497	0.461
Canberra	0.814	0.658	0.577	0.489	0.451
Chebyshev	0.809	0.594	0.505	0.426	0.391
Chi2	0.871	0.713	0.606	0.519	0.476
Cityblock	0.876	0.707	0.619	0.527	0.482
Correlation	0.852	0.689	0.598	0.509	0.477
Cosine	0.852	0.677	0.594	0.506	0.473
Euclidean	0.857	0.675	0.588	0.504	0.465
Hamming	0.714	0.473	0.397	0.338	0.314
Minkowski	0.857	0.675	0.588	0.504	0.465
PM	**0.966**	**0.842**	**0.733**	**0.624**	**0.571**

performance of the retrieval and consequently its diagnostic capability. To evaluate the generalization ability of the proposed CBDLR method, cross dataset validations have been performed on the new ISIC2019 dataset [46]. This dataset includes 25, 331 dermoscopic images labeled with 8 skin lesion classes (the same classes within ISIC2018 except the AKIEC class that has been divided into two sub-classes: Actinic Keratosis (AK) and Squamous Cell Carcinoma (SCC)). When compared to ISIC 2018, ISIC 2019 is found to be more challenging as many images are uncropped, with lesions present in difficult and uncommon locations. Preliminary retrieval tests have been conducted on a sample of 360 images randomly selected from ISIC2019, while being representative of the 8 classes, without any additional training for both feature extractor and similarity recommender. The obtained results (Table 9) show that the *P@K* value has increased by at least 4% for 10 retrieved images when compared to all other distances. This finding confirms the effectiveness of the dynamic recommendation of distance and the robustness of the proposed similarity recommender. Finally, diagnosis performance evaluation was conducted in order to assess the diagnostic generalization capability of the proposed method on the ISIC2019 dataset (Table 10). Although the ISIC2019 data are much more challenging than the ISIC2018 ones, the proposed model, trained by using only the ISIC 2018 dataset, has been able to achieve a significantly good performance (Average Precision = 76%, Average Recall = 72%). Nevertheless, retraining the feature extractor and the similarity recommender using the ISIC2019 datatset should improve retrieval as well as diagnosis results.

5. Discussion

The main objective of this study is to improve the CBDLR performance by investigating deep-learned features while integrating a similarity measure recommender that dynamically predicts the appropriate distance metric for each query. CNN has been adopted to learn effective features aiming to retrieve similar images to a query image based on the performance of these features to classify skin lesions into 7 classes of diseases. In fact, a ResNet50-based transfer learning has been performed using the ISIC2018 dataset in order to deeply craft a set of features, relying on common signs of skin lesion classes. This dataset is the largest publicly available collection of quality controlled dermoscopic images which includes a representative collection of all important diagnostic categories in the realm of pigmented lesions. It is obvious that the deep-

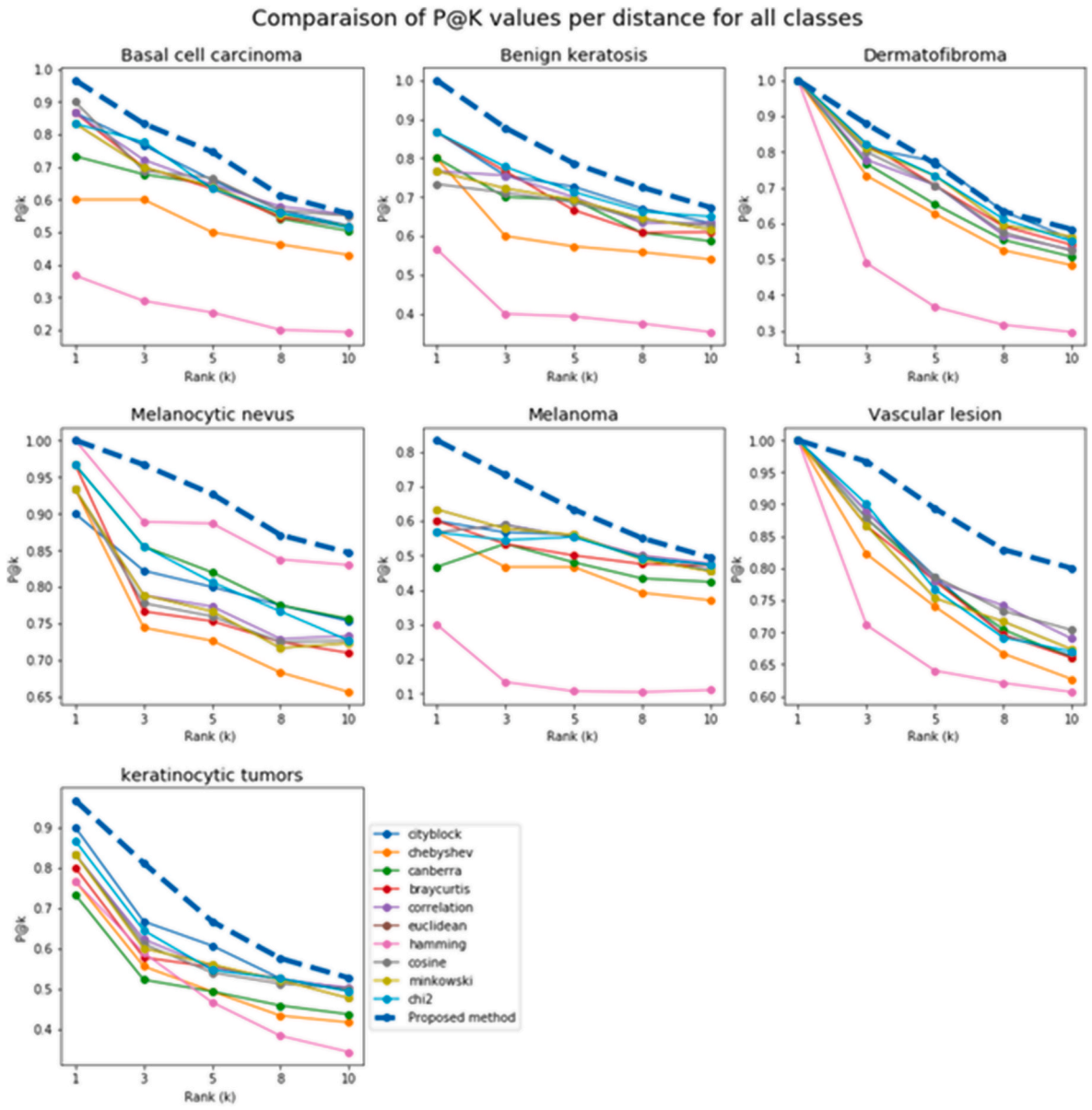

Fig. 7. Comparison of the *P@K* rate of the proposed CBDLR method, for the 7 classes of diseases of the ISIC2018 dataset, with the ones of the 10 investigated distance metrics.

Table 7
Diagnosis performance of the proposed CBDLR system.

Skin Lesion	Precision	Recall	F1-score
BCC	0.91	1.0	0.95
BKL	0.94	1.0	0.97
DF	0.94	0.97	0.95
NV	0.91	1.0	0.95
MEL	1.00	0.77	0.87
VASC	0.97	1.0	0.98
AKIEC	1.0	0.9	0.95
Average	0.95	0.95	0.95

Table 8
Diagnosis performance of the ResNet50 Classifier.

Skin Lesion	Precision	Recall	F1-score
BCC	0.97	0.97	0.97
BKL	1.0	0.97	0.98
DF	1.0	1.0	1.0
NV	0.97	1.0	0.98
MEL	0.9	0.93	0.92
VASC	1.0	1.0	1.0
AKIEC	0.93	0.9	0.92
Average	0.97	0.97	0.97

learned features should be correlated with diagnosis and/or expert's observation. In fact, more than 50% of lesions have been confirmed by pathology; while the ground truth for the rest of the cases was either follow-up, expert consensus, or confirmation by in-vivo confocal microscopy [36]. The obtained results confirm the ability of CNN to extract relevant features, illustrating non-linear correlations, from raw data without any additional feature extraction techniques while allowing to learn complex representations at different levels of abstraction. Indeed, the quantitative evaluation demonstrates the ability of the obtained model to extract discriminative features through the distinction between the 7 classes, given the true labels mainly vetted by recognized melanoma experts. Furthermore, a ground truth has been automatically built

Table 9
Performance evaluation of the retrieval results on the ISIC2019 dataset.

Distance	P@1	P@3	P@5	P@8	P@10
Braycurtis	**0.82**	0.61	0.54	0.50	0.48
Canberra	0.81	0.59	0.53	0.50	0.48
Chebyshev	0.78	0.58	0.52	0.47	0.44
Chi2	0.81	0.60	0.54	0.49	0.48
Cityblock	0.80	0.60	0.53	0.49	0.48
Correlation	0.80	0.59	0.54	0.50	0.48
Cosine	0.80	0.60	0.54	0.50	0.48
Euclidean	0.81	0.60	0.54	0.48	0.47
Hamming	0.77	0.48	0.42	0.38	0.36
Minkowski	0.81	0.60	0.54	0.48	0.47
PM	0.80	**0.63**	**0.57**	**0.53**	**0.52**

Table 10
Diagnosis performance evaluation using the ISIC2019 dataset.

Skin Lesion Class	Precision	Recall	F1-score
BCC	0.65	0.68	0.67
BKL	0.62	0.72	0.67
DF	0.95	0.63	0.76
NV	0.81	0.78	0.80
MEL	0.56	0.80	0.66
VASC	0.93	0.83	0.88
AKIEC	0.78	0.48	0.60
Average	0.76	0.70	0.72

to be investigated within the context of transfer learning. The obtained model allows to dynamically recommend the most adequate distance metric for any new query image without being obliged to make this choice in a global manner for all queries. Indeed, thanks to the queries' simulations and evaluation results based on labeled data by expert dermatologists, the constructed ground truth has identified the most adequate distance for each image within the ISIC2018 dataset. The experimental study has confirmed that the identified ground truth is correlated with expert's diagnosis for the 7 skin disease classes, and that all distances are suitable (*i.e.* selected as best distance for 211 images at least). Nevertheless, it would be interesting to couple the automatic generation of the ground truth with manual analysis of simulation results in order to examine the correlation between the visual signs of skin lesions and the selected distances. Such analysis can be useful to refine the generated ground truth, by minimizing the number of the investigated metrics for example, and therefore to improve the efficiency of the similarity recommender. Then, the task of subject-dependency distance recommendation for retrieving similar cases to any query image has been fulfilled using a dynamic query-driven distance selection mechanism based on a CNN-based transfer learning model. The designed recommender proved to be able, when used jointly with KNN, to retain the best distance for evaluating similarity between each image and the query one. That is to say, each subject has its own distance. The choice of the KNN algorithm has been motivated by the fact that it does not make any underlying assumptions about the distributions of labeled data. However, it depends on the quality of the data since the prediction stage might be slow and sensitive to the scale of the data and to irrelevant features in the case of large data. For larger datasets, PCA and hashing-based indexing techniques can be adopted. Overall, the proposed method has proved to be effective for assisting dermatologists by integrating the expert's knowledge in automatic systems. Indeed, the used datasets for the quantitative evaluation of the proposed CBDLR system are mainly labeled by experts. It is worth noting that the same method has been implemented in order to retrieve relevant images according to the malignancy signs (*i.e.* 2 classes of malignity: Malign and Benign) by fine tuning VGG16 pre-trained model as feature extractor and training similarity recommender that uses the automatically generated ground truth. The designed model shows higher retrieval performance comparatively to the use of one distance, which confirms the effectiveness of the proposed similarity recommender. Furthermore, to the best of our knowledge, this is the first study that investigates the ISIC2018 dataset within the framework of CBDLR, which justifies the absence of quantitative comparisons of the proposed method, specially if we consider that the number of lesion classes differs from one study to another. For instance, in Ref. [47] (*resp.* [9]), 2 (*resp.* 10) classes and 30 (*resp.* 325) queries were assessed while recording 70.19% (*resp.* 85%) as precision. Nonetheless, it is observed that the time complexity of the proposed CBDLR system is quite low. Table 11 illustrates the CPU time using Google Colab environment (GPU Runtime with NVIDIA Tesla K80, 12 GB RAM, 2496 CUDA cores). Training steps of feature extractor and similarity recommender have relatively low computational cost thanks to transfer learning mechanism. The most expensive step is the automatic ground truth generation that costs approximately 6 hours. This can be justified by the number of similarity measures investigated and the size of feature database. However, the online phase takes only 0.9 second for retrieving 10 similar images. Finally, although it has been implemented within the framework of skin diseases, the proposed method can be effectively validated for the retrieval of any kind of medical images, provided that an annotated dataset is available. In fact, the proposed method is generic enough to be applied in different fields, even outside the medical domain.

6. Conclusion and perspectives

During the last years, the massive increase of skin lesion images has precipitated the challenge of mining specific images among huge collections, which explains the exhibition of CBDLR as an active research topic. In fact, unlike automatic classification tools that present results as 'black box', CBDLR provides a sorted set of similar images, with a confirmed diagnosis, relative to a given skin lesion image. This can serve to support dermatologists in the diagnosis process, without directly providing a second opinion. It can be used for educational purposes in order to provide more explainable results. Within this context, the image comparison issue, which is a key point for designing successful CBDLR systems, has been particularly addressed. The proposed solution aims to enhance the retrieval performance by integrating the similarity measure recommender that is able to predict the best distance corresponding to each query image. This recommender is a CNN classifier that learns, from an automatically generated ground truth, how to identify the adequate distance metric for any query image. Experimental results prove that the proposed system is effective to retrieve clinically similar lesions from a feature database generated from the challenging ISIC2018 dataset. This system can provide dermatologists with a diagnostic aid by automatically retrieving similar images of different skin disease types. This can significantly improve the patients' survival rate as well as reduce their suffering and the treatment cost. In future research, we aim to further improve the retrieval results by integrating relevance feedback mechanisms. They allow the user to specify the degree of relevance of the produced CBDLR results. Besides, using the provided metadata within the ISIC2018 dataset (*e.g.* demographic information, lesion localisation …) can refine the retrieval results. Furthermore, since deep-learning-based CBDLR requires a balanced and complete training set

Table 11
Computational cost of the proposed CBDLR method.

Step	Computational cost
Offline phase	
Feature extractor training	19 minutes
Feature extraction	≈ 2 hours 30 minutes
Ground truth generation	≈ 6 hours
Similarity recommender training	30 minutes
Online phase	
Retrieval query	0.9 second

with annotated data, which is hard to gather in medical domain, learning accurate models from imbalanced and incomplete training data is more and more crucial for operational CBDLR systems.

Declaration of competing interest

There is no conflict of interest.

References

[1] Z. Li, X. Zhang, H. Müller, S. Zhang, Large-scale retrieval for medical image analytics: a comprehensive review, Med. Image Anal. 43 (2018) 66–84.
[2] M. Kashif, G. Raja, F. Shaukat, An efficient content-based image retrieval system for the diagnosis of lung diseases, J. Digit. Imag. 33 (2020) 971–987.
[3] I. Bakkouri, K. Afdel, Computer-aided diagnosis (CAD) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images, Multimed. Tool. Appl. 79 (2020) 20483–20518.
[4] M. Goyal, T. Knackstedt, S. Yan, S. Hassanpour, Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities, Comput. Biol. Med. 127 (2020) 104065.
[5] A. Maiti, B. Chatterjee, A. Ashour, N. Dey, Computer-aided diagnosis of melanoma: a review of existing knowledge and strategies, Curr. Med. Imag. 16 (2020) 835–854.
[6] F.A. Damian, S. Moldovanu, N. Dey, A. Ashour, L. Moraru, Feature selection of non-dermoscopic skin lesion images for nevus and melanoma classification, Computation 8 (2020) 41.
[7] J.M. Patel, N.C. Gamit, A review on feature extraction techniques in Content Based Image Retrieval, in: International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 2259–2263.
[8] C. Barata, E. Celebi, J. Marques, Explainable skin lesion diagnosis using taxonomies, Pattern Recogn. 110 (2021) 107413.
[9] O. Layode, T. Alam, M.M. Rahman, Deep learning based integrated classification and image retrieval system for early skin cancer detection, in: IEEE Applied Imagery Pattern Recognition Workshop, 2019, pp. 1–7.
[10] N. Pasumarthi, L. Malleswari, An empirical study and comparative analysis of Content Based Image Retrieval (CBIR) techniques with various similarity measures, in: International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS 2016), 2016, pp. 1–6.
[11] M. Owais, M. Arsalan, J. Choi, K.R. Park, Effective diagnosis and treatment through Content-Based Medical Image Retrieval (CBMIR) by using artificial intelligence, J. Clin. Med. 8 (4) (2019) 462.
[12] A. Soudani, W. Barhoumi, An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction, Expert Syst. Appl. 118 (2019) 400–410.
[13] M.M. Rahman, A Decision support system for skin cancer recognition with deep feature extraction and multi response linear regression (MLR)-based meta learning, in: International Conference on Health Informatics & Medical Systems, 2019, pp. 10–15.
[14] N. Dey, V. Rajinikanth, A. Ashour, J.M. Tavares, Social group optimization supported segmentation and evaluation of skin melanoma images, Symmetry 10 (2018) 51.
[15] G.W. Jiji, A. Rajesh, P.J.D. Raj, CBI+R: a fusion approach to assist dermatological diagnoses, Int. J. Image Graph. 21 (2021) 2150005.
[16] G.W. Jiji, P.J.D. Raj, A retrieval system to analyse dermatological lesions using feature ortho-normalisation, J. Exp. Theor. Artif. Intell. **31** (2019) 41–55.
[17] F. Warsi, R. Khanam, S. Kamya, C.P. Suarez-Araujo, An efficient 3D color-texture feature and neural network technique for melanoma detection, Inf. Med. Unlocked 17 (2019) 100176.
[18] J. Kawahara, A. BenTaieb, G. Hamarneh, Deep features to classify skin lesions, in: IEEE International Symposium on Biomedical Imaging, 2016, pp. 1397–1400.
[19] V. Pomponiu, H. Nejati, N. Cheung, Deepmole: deep neural networks for skin mole lesion classification, in: IEEE International Conference on Image Processing, 2016, pp. 2623–2627.
[20] C. Barata, M.E. Celebi, J.S. Marques, A survey of feature extraction in dermoscopy image analysis of skin cancer, IEEE J. Biomed. Health Inf. 23 (2019) 1096–1109.
[21] P. Tschandl, G. Argenziano, M. Razmara, J. Yap, Diagnostic accuracy of content based dermatoscopic image retrieval with deep classification deatures, CoRR **abs/1810.**09487 (2018).
[22] C.F.N. Codella, C.C. Lin, Collaborative Human-AI (CHAI): Evidence-based Interpretable Melanoma Classification in Dermoscopic Images. MLCN/DLF/iMIMIC@MICCAI, 2018, pp. 97–105.
[23] A. Ashour, R.M. Nagieb, H. El-Khobby, M.A. Elnaby, N. Dey, Genetic algorithm-based initial contour optimization for skin lesion border detection, Multimed. Tool. Appl. 80 (2021) 2583–2597.
[24] Y. Boualleg, M. Farah, I.R. Farah, TLDCNN: a triplet low dimensional convolutional neural networks for high-resolution remote sensing image retrieval, in: Mediterranean and Middle-East Geoscience and Remote Sensing Symposium, M2GARSS, 2020.
[25] N. Passalis, A. Iosidis, M. Gabbouj, A. Tefas, Variance-preserving deep metric learning for content-based image retrieval, Pattern Recogn. Lett. 131 (2020) 8–14.
[26] N.C.F. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S.W. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M.A. Marchetti, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC), 2019. CoRR abs/1902.03368.
[27] A. Baldi, R. Murace, E. Dragonetti, M. Manganaro, O. Guerra, S. Bizzi, L. Galli, Definition of an automated Content-Based Image Retrieval (CBIR) system for the comparison of dermoscopic images of pigmented skin lesions, Biomed. Eng. Online 8 (2009) 264–279.
[28] X. Pu, Y. Li, H. Qiu, Y. Sun, Deep semantics-preserving hashing based skin lesion image retrieval, in: F. Cong, A. Leung, Q. Wei (Eds.), Advances in Neural Networks - ISNN 2017 vol. 10262, 2017, pp. 282–289.
[29] L. Ballerini, X. Li, R.B. Fisher, 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, 2009, pp. 31–38.
[30] S. Allegretti, F. Bolelli, F. Pollastri, S. Longhitano, G. Pellacani, C. Grana, Supporting skin lesion diagnosis with content-based image retrieval, in: International Conference on Pattern Recognition, 2020, pp. 1–8.
[31] W. Barhoumi, S. Dhahbi, E. Zagrouba, A collaborative system for pigmented skin lesions malignancy tracking, in: IEEE International Workshop on Imaging Systems and Techniques, 2007, pp. 1–6.
[32] X. Gao, T. Mu, J. Goulermas, J. Thiyagalingam, M. Wang, An interpretable deep architecture for similarity learning built upon hierarchical concepts, IEEE Trans. Image Process. 29 (2020) 3911–3926.
[33] A.S. Rasheed, D. Zabihzadeh, S.A.R. Abdulhussien, Large-scale multi-modal distance metric learning with application to content-based information retrieval and image classification, Int. J. Pattern Recogn. Artif. Intell. 34 (2020) 2050034.
[34] W. Wu, D. Tao, H. Li, Z. Yang, J. Cheng, Deep features for person re-identification on metric learning, Pattern Recogn. 110 (2021) 107424.
[35] K.R. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3 (2016) 1–40.
[36] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (2018) 180161.
[37] N.C.F. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection: a challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC), in: IEEE International Symposium on Biomedical Imaging, 2018, pp. 168–172.
[38] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imag. 35 (2016) 1299–1312.
[39] A. Bissoto, M. Fornaciali, E. Valle, S. Avila, (De)Constructing bias on skin lesion datasets, in: IEEE Computer Vision and Pattern Recognition Workshop (CVPRW), 2019, pp. 2766–2774.
[40] P. Lio, P. Nghiem, Interactive atlas of dermoscopy, J. Am. Acad. Dermatol. 50 (2004) 807–808.
[41] T. Mendonca, P. Ferreira, J. Marques, A. Marçal, J. Rozeira, PH2 - a dermoscopic image database for research and benchmarking, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2013, pp. 5437–5440.
[42] M. Marchetti, N. Codella, S. Dusza, D. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, E. Celebi, J. DeFazio, N. Jaimes, A. Marghoob, E. Quigley, A. Scope, O. Yelamos, A.C. Halpern, Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images, J. Am. Acad. Dermatol. 78 (2018) 270–277.
[43] A. Khan, A. Sohail, U. Zahoora, A.S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, 2019. CoRR abs/1901.06032.
[44] N.Y. Ali, G. Sarowar, L. Rahman, J. Chaki, N. Dey, J.M. Tavares, Adam deep learning with SOM for human sentiment classification, Int. J. Ambient Comput. Intell. (IJACI) 10 (2019) 92–116.
[45] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, CoRR **abs/** 1311 (2013) 2901.
[46] M. Combalia, N. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, A. Halpern, S. Puig, J. Malvehy, BCN20000: Dermoscopic Lesions in the Wild, ArXiv **abs/1908.02288**, 2019.
[47] K. Belattar, S. Mostefai, A. Draa, Intelligent content-based dermoscopic image retrieval with relevance feedback for computer-aided melanoma diagnosis, J. Inf. Technol. Res. 10 (2017) 85–108.