Nieniewski *et al.*
EURASIP Journal on Image and Video Processing (2023) 2023:7
https://doi.org/10.1186/s13640-023-00607-y

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Studies in differentiating psoriasis from other dermatoses using small data set and transfer learning

Mariusz Nieniewski[1*], Leszek J. Chmielewski[2], Sebastian Patrzyk[3] and Anna Woźniacka[3]

*Correspondence: mariusz.nieniewski@wmii.uni.lodz.pl

[1] Faculty of Mathematics and Informatics, University of Lodz, ul. Banacha 22, 90-238 Lodz, Poland
[2] Institute of Information Technology, Warsaw University of Life Sciences, SGGW, ul. Nowoursynowska 159, 02-775 Warsaw, Poland
[3] Department of Dermatology and Venereology, Medical University of Lodz, plac Hallera 1, 90-647 Lodz, Poland

Abstract

Psoriasis is a common skin disorder that should be differentiated from other dermatoses if an effective treatment has to be applied. Regions of Interests, or scans for short, of diseased skin are processed by the VGG16 (or VGG19) deep convolutional neural network operating as a feature extractor. 1280 features related to a given scan are passed to the Support Vector Machine (SVM) classifier using Radial Basis Functions (RBF) kernels. The main quality of the described setup is a very small number of 75 psoriasis patients and 75 non-psoriasis patients used in the teaching and testing sets taken together. For each patient, a variable number of clinical images are taken. Then, the scans of size 256 × 256 pixels are cropped from these images. There are 1988 scans of psoriasis patients and 1582 of non-psoriasis patients. The other quality of the described setup is the use of transfer learning for carrying over the neural network's weights from non-medical domain (ImageNet) to clinical images of dermatoses. The next quality is that the input images are obtained with smart phone cameras without any special arrangements or equipment, so there is a great variability in working conditions, which hampers discriminative power of the classifier. The primary classification is carried out on individual scans, and then, majority voting is executed among the scans pertaining to an individual patient. The obtained recall (sensitivity) is 85.33%, and the precision is 82.58%. The 95% confidence interval for the accuracy of 80.08% is [77.14, 83.04]%. These numbers indicate that the described system can be useful for remote diagnosing of psoriasis, particularly in areas where access to dermatological personnel is limited.

Keywords: Psoriasis, Convolutional neural networks, Transfer learning, Papulosquamous skin diseases, Deep learning

1 Introduction

1.1 General remarks

Psoriasis is a common chronic inflammatory skin disease that causes physical and psychological burden to patients. As a reference, one can turn to the paper by Higgins [1]. Psoriasis appears in a variety of forms with distinct visual patterns, such as plaque, guttate, inverse, pustular, and erythroderma. The most common is plaque, which occurs in 80% of cases. Psoriasis is easily and often misdiagnosed, because various

papulosquamous skin diseases have similar clinical presentations. Yet, it is impractical to perform an exhaustive examination or biopsy for each psoriasis patient. For a detailed review of clinical photographs eliciting the appearances of psoriasis and other skin diseases, the reader is directed to the book edited by Morris-Jones [2].

1.2 Related research

Generally speaking, the number of publications devoted to diagnosing psoriasis by means of image analysis is relatively modest when compared with what is available, for example, for melanoma. In the following, most of the relevant papers are mentioned. These papers may be concerned with clinical or dermoscopic images; however, dermoscopic images are easier to classify, since they are taken under more uniform lighting conditions. This paper is devoted exclusively to clinical images.

In one of the early papers on psoriasis image classification, Hashim et al. [3] developed a method differentiating three common types of psoriasis, that is, guttate, plaque, and erythroderma. The authors conducted statistical discrimination of psoriasis varieties by means of chromatic color indices.

Al-Abbadi et al. [4] combined 9 color features and 4 GLDM texture features put into a 13-element feature vector used as an input to the feedforward neural network classifying skin sample images respectively as psoriasis or other dermatoses.

Ballerini et al. [5] proposed a hierarchical classification system based on the K-Nearest Neighbors model and applied it to non-melanoma skin lesion classification. Color and texture features were extracted from skin lesion images obtained with a standard camera. The classification task was decomposed into a set of simpler problems, one for each node of the classification. Feature selection was embedded in the hierarchical framework choosing the most relevant feature subset at each node.

Shrivastava et al. [6] developed a CADx system to automatically classify dermatological images into psoriatic lesion and healthy skin using an online platform. The skin samples were cropped freehand to capture any shape. A total of 270 samples of normal skin and 270 samples of abnormal (psoriasis) skin were obtained from 30 patients. The feature space included data from grayscale space, color space, and aggressiveness of psoriasis disease, such as redness and chaoticness. The classification was carried out by the SVM.

One of the fundamental problems when dealing with the diagnosis of psoriasis is the scarcity of dermatological images for training the neural network. A modern neural network, such as developed by Simonyan et al. [7], requires a data set of millions of training images, for example, ImageNet. This huge data set has no relation to medical images; hence, it appears to be of little use. Nevertheless, Gupta et al. [8] showed that it is possible to accomplish transfer learning, that is carrying over the weights of a neural network from one application domain to another and fine-tuning the new network. In particular, Gupta et al. [8] tested this approach for the development of a classifier of male and female illustrations in a textbook. For this purpose, they modified the final fully connected layers of the original VGG16 network which do the classification and left unchanged the initial layers responsible for feature extraction.

An et al. [9] developed hierarchical deep learning model using transfer learning for glaucoma detection and classification based on a small number of medical images. In

this approach, the transfer learning was implemented from VCC16 trained on ImageNet data set to a proprietary set of glaucoma images. The training set consisted of 156 healthy and 798 glaucoma eyes. The transfer learning included two levels: the low level model for differentiation between glaucoma eye and a healthy eye and the high level model for recognition of four types of glaucoma. In the case of hierarchical transfer learning, the weights were transferred from the original network to the low-level model, and next from the low-level model to the high-level model.

Alzubaidi et al. [10] proposed an intelligent medical system for the diagnosis of diabetic foot ulcer (DFU). Development of such a system faces similar problems to that of psoriasis recognition, that is lack of a large domain specific data set and necessity of using other data set coming from another domain. The authors showed experimentally that the proposed model with the DFU data set achieved an F_1 score of 86.6% with training from scratch, 89.4% with transfer learning from non-medical domain, and 97.6% with transfer learning from the same domain as the target data set.

In another paper, Alzubaidi et al. [11] proposed a transfer learning approach based on training the deep learning model on large unlabeled medical image data set and subsequent transferring the knowledge to train the deep learning model on the small amount of labeled medical images. The idea is that during the training, we are mainly concerned with feature extraction and classification is less important. This means that at this stage, we can use data with random, incorrect labels. This makes sense if we have a large amount of unlabeled data. The next stage uses the weights acquired during the first stage and performs fine-tuning on a small amount of labeled data. The method was tested, among others, on 200 000 unlabeled images of skin cancer. In the second stage, 9 000 labeled benign samples and 584 labeled malignant were used. The method provokes interest but cannot be used for psoriasis directly, since obtaining thousands of psoriasis images is out of the question.

Hogarty et al. [12] presented an introduction to the basic concepts of artificial intelligence in dermatology as well as a review of the main achievements until the year 2020. Several skin disorders were considered, psoriasis among others. The authors' conclusion was that the use of machine learning in smart phones has a great potential for patient care particularly in improving sensitivity and accuracy of the screening of skin lesions, but is also susceptible to the same flaws as classical statistics, and is unlikely to become more than an adjunct to clinical practice for the foreseeable future.

Celebi et al. [13] presented a methodological approach to the classification of dermoscopy images. Their approach involves border detection, feature extraction, and SVM classification with model selection. The system was tested on a large set of images. Promising results were obtained despite the fact that the images came from different sources and there was no control over their acquisition.

Kim et al. [14] developed an equipment together with smart phone software implementing multispectral imaging together with machine learning for discrimination between seborrhoeic dermatitis and psoriasis on the scalp. The two diseases have in fact very similar manifestations and the distinction was made based on the analysis of the skin reflectance properties as a function of wavelength. Three variants of machine learning included: logistic regression, SVM, and multilayer perceptron. The results certainly are interesting, but they are valid for just two specific dermatoses and require special

equipment, so it is hard to assume that this approach might be used in less specialized circumstances.

Hameed et al. [15] conducted classification of dermoscopic images obtained from various sources. The images were put in four classes: healthy, eczema, benign, and malignant. The multiclass multilayer algorithm first differentiated healthy and unhealthy cases, and then classified unhealthy into melanoma and eczema, and finally melanoma into malignant and benign.

Mittal et al. [16] developed an approach for segmentation of skin lesion images. The authors removed the noise from the raw images and there was an increase in the entropy after filtering and segmentation, which suggested an improvement in sharpness and quality of resultant images. The approach was tested on ten types of dermatoses, psoriasis included.

Taur et al. [17] proposed segmentation of psoriasis images by means of a multiresolution-based signature subspace classifier. In this approach, the fuzzy texture spectrum and the two-dimensional fuzzy color histogram in the hue-saturation space were first adopted as the feature vector to locate homogeneous regions in the image. Then, these regions were used to compute the signature matrices for the orthogonal subspace classifier to obtain a more accurate segmentation. In the experiments, the proposed method was quantitatively evaluated using a similarity function and compared with the SVM least-squares method.

Tien et al. [18] developed a multispectral polarized imaging system to capture the image of psoriasis and also used image processing method for evaluating scaly levels. The authors calculated the Psoriasis Area Severity Index for assessing the severity level of psoriasis. Based on this approach, the proposed algorithm automatically segmented scales from the skin surface.

Wei et al. [19] proposed a recognition method for identification of three skin diseases: herpes, dermatitis, and psoriasis. In this approach, the skin images were initially preprocessed to remove noise and irrelevant background via filtering and transformation. Then, the method of Gray Level Cooccurrence Matrices was employed for segmentation of images of skin disease based on texture and color features. Finally, using the SVM classification method, three types of skin diseases were identified.

Udrea et al. [20] developed a system for analysis of images of pigmented and non-pigmented skin lesions obtained by means of a smart phone camera. This system was based on Generative Adversarial Neural Networks implementing image-to-image translation with conditional adversarial nets. The final result was the segmentation of the lesion. This is an initial step, and the next one would be the actual image classification.

Peng et al. [21] developed a deep learning method for classification and diagnosing of psoriasis. Their method included data enhancement and subsequent processing by deep residual network ResNet34. They used 30,000 psoriasis data samples. The net aimed at differentiating psoriasis and healthy persons, as well as identifying four types of psoriasis: psoriasis vulgaris, joint psoriasis, purulent psoriasis, and erythroderma. The authors stressed a strong imbalance of the classes as healthy persons were much more numerous than psoriasis patients.

Yang et al. [22] developed a convolutional neural network for diagnosis of psoriasis by means of analysis of dermoscopic images. They used a proprietary data set of

1156 patients which converted in 7033 scans. The EfficientNet-B4 was pretrained on ImageNet. The original images were of size 1872 × 1053 pixels and were rescaled and cropped to 380 × 380 required by the network. In addition to dermoscopic images, the data set contained clinical images. The network classified dermatoses into four classes: psoriasis, eczema, lichen planus and others. There were some limitations, and images with lesions on scalp, nails and mucosa were removed. These excluded dermatoses do happen in practice, however, and have to be processed somehow.

Fujisawa et al. [23] developed a deep-learning-based classifier of skin tumors using a small data set of clinical images. They claim that their classifier surpassed board-certified dermatologists in skin tumour diagnosis. Although melanoma is outside the interest of the current study a general environment and testing methodologies are similar. It is worth noting that all of the images in the described case were taken with digital cameras, which had at least 6 million pixels and had a macro lens and macro ring flash. No dermoscopic images were included in this study. The small data set in this case included 4867 images obtained from 1542 patients. The images were classified into 14 classes.

Mikołajczyk et al. [24] conducted a clinical test aiming at comparing the diagnostic accuracy of a popular free-to-use web application for automatic dermatosis diagnosis vs. expert diagnosis of selected skin diseases. The authors observed that the probability of a diagnosis repeating for the same patient was below 25%. Furthermore, reliability, sensitivity, and specificity were insufficient for clinical purposes. The described web application might be used for didactic purposes but not in any real-life situation.

The aim of the current paper is an improvement of diagnostic results under the same conditions as in [24] that is of clinical images obtained exclusively via smart phone or hand-held camera. The use of any additional medical equipment or procedures does not come into account, since this diagnostic tool is for the patients in remote areas.

The preceding review of the current literature showed that there are tools for diagnosing psoriasis in a clinical environment where high-quality dermoscopic images are available. This is not enough for some medical specialists, who would like, or maybe dream of, that a patient self-diagnose at home or in a local non-specialized clinic. In particular, the purpose of the current paper is the development of a method for differentiation of psoriasis from other dermatoses based on clinical images which are much easier to acquire than dermoscopic images, but suffer from greater variability of conditions under which they are taken.

2 Materials and methods

Our images of psoriasis and non-psoriasis cases were collected in the Department of Dermatology and Venereology, Medical University of Lodz, Lodz, Poland during a period of about 2 years. This psoriasis study was approved by the Medical Ethics Committee of the Medical University of Lodz.

A record of diagnosed non-psoriasis diseases is presented in Table 1. One dermatosis, that is eczema, was diagnosed 7 times; herpes zoster was diagnosed 5 times; and pemfigoid was diagnosed 4 times. Other diseases were represented by one to three cases. This simply reflects the fact the certain dermatoses were encountered in a particular clinic during certain time. Depending upon a geographical area, one can expect that the list of dermatoses and their relative frequencies may vary to some degree.

Table 1 Diagnosed non-psoriatic diseases

Disease	No of patients	Disease	No of patients
Granuloma annulare	3	Urticaria	2
Quincke's vascular lesion	1	Drug urticaria	1
Eczema	7	Pemfigoid	4
Atopic dermatitis	3	Systemic sclerosis	1
Pityriasis rosea Gibert	1	Eczema of lower leg	1
Head lice bites	1	Impetigo contagiosa	1
Seborrhoeic dermatitis	2	Mycosis fungoides	1
Lichen planous	2	Secondary early syphilis (man)	1
Systemic sclerosis	2	Carcinoma basocellulare	1
Bed bug bites	1	Lymphoma	3
Disseminated lupus erythematosus	2	Lupus erythematosus	3
Herpes zoster	5	Actinic keratosis	1
Subacute cutaneous lupus erythematosus	2	Parapsoriasis	3
Papular lesions in the course of lupus	1	Lichen sclerosus	1
Vascular purpura	1	Inflammatory mole	1
Contact dermatitis	1	Tinea circinate	1
Erythema induratum	1	Erythema nodosum	2
Drug-induced changes	1	Pemphigus	1
Discoloration in hypothyrodism	1	Allergic reaction to bites	1
Mycosis	3	Birthmark	1
Porokeratosis	2	Insect bites	1

The data set under consideration was arranged in the following way. There were 75 patients with psoriasis and 75 patients with other dermatoses. From these patients, 1988 scans of psoriasis and 1582 scans of other diseases were obtained. The medical staff taking images were mostly unaware of the future use of images. In particular, it turned out that the number of images for an individual patient was highly variable. As a result, the number of scans for each patient was also variable. All the scans were of the size 256×256 pixels. The scans were cropped by an IT specialist from the set of images available for a given patient in a random manner. However, care was taken to have some representative variety of scans for a given patient. The entire set of scans was tested using an approach similar to 5-fold split; however, it was impossible to split this set into five equal parts, since this would involve assigning some scans from an individual patient to the training set and some to the testing set. The point is that the scans of a given patient tend to be highly correlated and using the scans from the same patient in both the training set and the testing set would bias the results.

As an evidence of variability of conditions under which images were acquired, one can consider the images' resolution. Strictly speaking, all images for a given patient were obtained with the same resolution. This did not help too much, however. The most common image resolution was 4032×3024 pixels for 41 psoriasis patients and for 57 non-psoriasis patients. The second most common resolution was 4272×2848 pixels for 17 psoriasis patients. Overall, there were more than 12 different resolutions. The smallest one was 980×552 pixels for one psoriasis patient. One relatively common variation for a single patient was the exchange of the width and height of the image.

Our system for differentiation of psoriasis from other dermatoses consists of two main parts (comp. Figure 1): the feature generation part, which is based on the deep convolutional neural network (CNN), and the feature classifier which can be implemented in several common versions, among which the best one was chosen. The CNN selected was VGG16, and VGG19 as a possible alternative. Both networks are fully described in original publication by Simonyan et al. [7], and here, only details indispensable for our presentation are given. As shown in Fig. 1 the network consists of five convolutional blocks. Each block contains two or three convolution layers and a maximum pooling layer. The usual information flow path in the CNN follows thin vertical arrows, as shown in Fig. 1, and the signals from the last pooling layer are passed to the flattening layer and subsequently to the dense layer. The shape of the information of the 5th block pooling layer is $8 \times 8 \times 512$. This results in $32{,}768 + 1$ weights, which have to be learned. In the original VGG network, the output layer has 1000 outputs, each output delivering logical value of

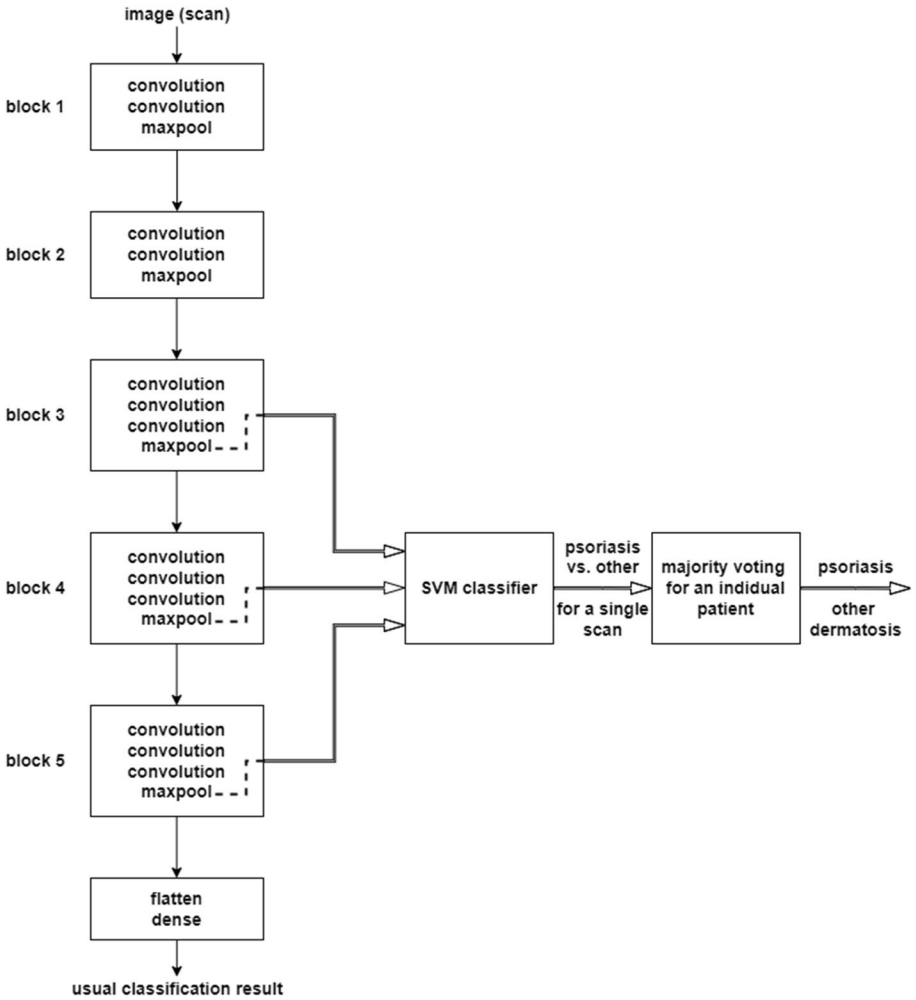

Fig. 1 Block diagram of the modification of the VGG16/19 network. The usual information flow from input image to output classification is indicated by thin vertical arrows. The information flow in the case of psoriasis classification follows the horizontal double arrows connecting three convolutional blocks in the VGG16 with the SVM classifier. The output of the SVM classifier can be used both for individual scans as well for majority voting for a single patient

one when an object of a particular class is detected. A typical transfer learning consists in using the original weights obtained on the ImageNet data set for all the layers except the top (dense) one and fine-tuning the weights of the dense layer in the training process in which a limited number of domain specific examples are shown to the network. In our case, the output layer would have two outputs—for two classes: psoriasis and non-psoriasis. In fact, an extensive experimentation with several variants of this approach was conducted, and the results of fine-tuning were unacceptable. The ImageNet data set contains about 1.2 million images of all kinds of objects, such as cars, animals, buildings, etc.; however, they bear no resemblance to medical images and the available number of our dermatosis samples was way too small for fine-tuning even with the extensive use of data augmentation for increasing the number of samples.

In this paper, we take another approach to transfer learning. The idea of transfer learning has become quite popular in recent time and the reader is referred to the review paper by Kora et al. [25]. In our particular case, we use that part of the VVG16 (or VGG19) consisting of five convolutional blocks, as illustrated in Fig. 1, and take the outputs of pooling layers of the convolutional blocks 3, 4, 5, as shown by wide horizontal arrows in Fig. 1. These outputs are passed to an external classifier that can be of any nature. In this way, the VGG16 simply plays a role of a feature extractor, and we treat the outputs from the VGG as features, similar to, say, morphological features considered in [4] or any other relevant publication. The original weights of blocks 1, ..., 5 obtained from the ImageNet data set are still used without any change. The number of features obtained from blocks 3, 4, and 5 are, respectively: 256, 512, and 512. These features are simply concatenated, so we have 1280 features per scan. There are a number of classifiers possible, and we tested SVM, decision tree, and random forest. There is an abundant literature on the classifiers, and it is not necessary to get into details here. The reader is referred to the classical book by Hastie et al. [26].

All the calculations described in this paper were performed using Google Colab from the personal computer but could be carried over to any portable device with appropriate operating system, for example, to the smart phone, if needed.

3 Results and discussion

3.1 Precision and recall

Because of a limited number of disease cases, it is customary to split them into several folds of the same size and to use one fraction of cases in each fold for training and the rest for testing. Then, one might do the averaging of the results over the folds. This approach cannot be followed exactly in our case mainly because the number of patients is quite small. At the same time, the number of images is at least several times larger, and the number of scans even higher. Following the rule of splitting the collection of patients into individual folds, we obtained the distribution of the scans, as shown in Table 2. This gave us some imbalance of the two classes (psoriasis vs. others) when considered in terms of scans, but we avoided putting scans from the same patient into the training set and the testing set.

Confusion matrices obtained for various classifiers and all 5-folds in the case of the VGG16 network are shown in Fig. A1 in the Appendix to avoid clutter. TP is a number of True Positive scans, TN—True Negatives, FP—False Positives, FN—False Negatives.

Table 2 No of scans analyzed in each fold

Fold	1	2	3	4	5
Psoriasis, Training set	1554	1570	1741	1601	1477
Others, Training set	1262	1265	1315	1239	1247
Psoriasis, Testing set	434	409	247	387	511
Others, Testing set	320	317	267	343	335

The classifiers used were selected from the scikit-learn python library [27] and [28] and include the following: SVM with Radial Basis Function (RBF), SVM with Polynomial Basis Function, Decision Tree (DT), and Random Forest (RF). The particular parameters for the RBF and other classifiers are specified in the caption of Fig. 11. The meaning of these parameters as well as their tuning are described in detail in [27]. In particular, the RBF requires two parameters: γ and C. Basically, γ defines an overall scale factor for the SVM's notion of distance between two points; this in turn defines how a support vector shapes the decision boundary in its nearby neighborhood. In addition, C controls the trade-off between the slack variable penalty (misclassifications) and width of the margin between the classes. In the literature, the ranges of values $0.0001 < \gamma < 10$ and $0.1 < C < 10$ are mentioned as reasonable. Finding the optimal values might be implemented by setting up a grid of points representing selected pairs of γ and C, and analyzing what confusion matrices are obtained for all of these points. This would involve lengthy calculations with a vast majority of confusion matrices being quite far from ideal diagonal ones. The calculations can be greatly reduced by assuming the default "scale" value for the parameter γ, or we can search for a nearby numerical value possibly giving a better result. Similarly, the C parameter may have a default value of 1 or can be set to some number. In our experiments, the values of the parameters were obtained by experimentation. The use of "scale" for γ has the disadvantage that various values of γ may be used for various folds in Fig. 11. The underlined entry in the caption to Fig. 11 indicates the best results obtained. They were generated for RBF with $\gamma = 10^{-7}$ and $C = 15$. Inspection of Fig. 11 reveals that the value of C can be changed in some range without significant influence. The confusion matrices for $\gamma =$ "scale" are quite similar to those for underlined entry. Further cases in Fig. 11 show the confusion matrices for polynomial kernels of the second and third degrees. The results for the first degree are not shown, since they were significantly worse. The parameters used with the Decision Trees are: criterion, random−state, max−depth, and min−samples−leaf [27]. The last two cases show the confusion matrices for the Random Forest with the number of estimators equal to 100 or 200. In fact, the last two cases in Fig. 11 (and similarly in Fig. 12) are exemplary ones since repeating the calculations one obtains the confusion matrices in which individual entries can go several units up or down. In any case the Random Forest was not the best classifier in these experiments, and besides it was significantly slower in comparison with other classifiers in Fig. 11.

Figure 11 gives us a feeling of what happens in terms of actual numbers. For comparisons, however, it is more convenient to use the standard evaluation metrics, that is, the precision

$$\mathrm{Pre} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}),$$

which defines the fraction of relevant instances among the retrieved instances, and recall

$\mathrm{Rec} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$,

which defines the fraction of relevant instances that were retrieved.

The precision and recall calculated in percent for the confusion matrices of Fig. 11 are shown in Fig. 2.

Performing majority voting on scans belonging to individual patients, we obtain the confusion matrices for VGG16 and selected SVM RBF classifier, as shown in Fig. 3.

The folds in Fig. 3 (and Figs. 11 and 12) each contain 15 psoriasis cases and 15 other dermatoses. By summing the individual confusion matrices elementwise, we obtain the final average confusion matrix for the VGG16 network shown in Table 3. The corresponding precision is $\mathrm{Pre} = 82.581\,\%$, and the recall is $\mathrm{Rec} = 85.333\,\%$.

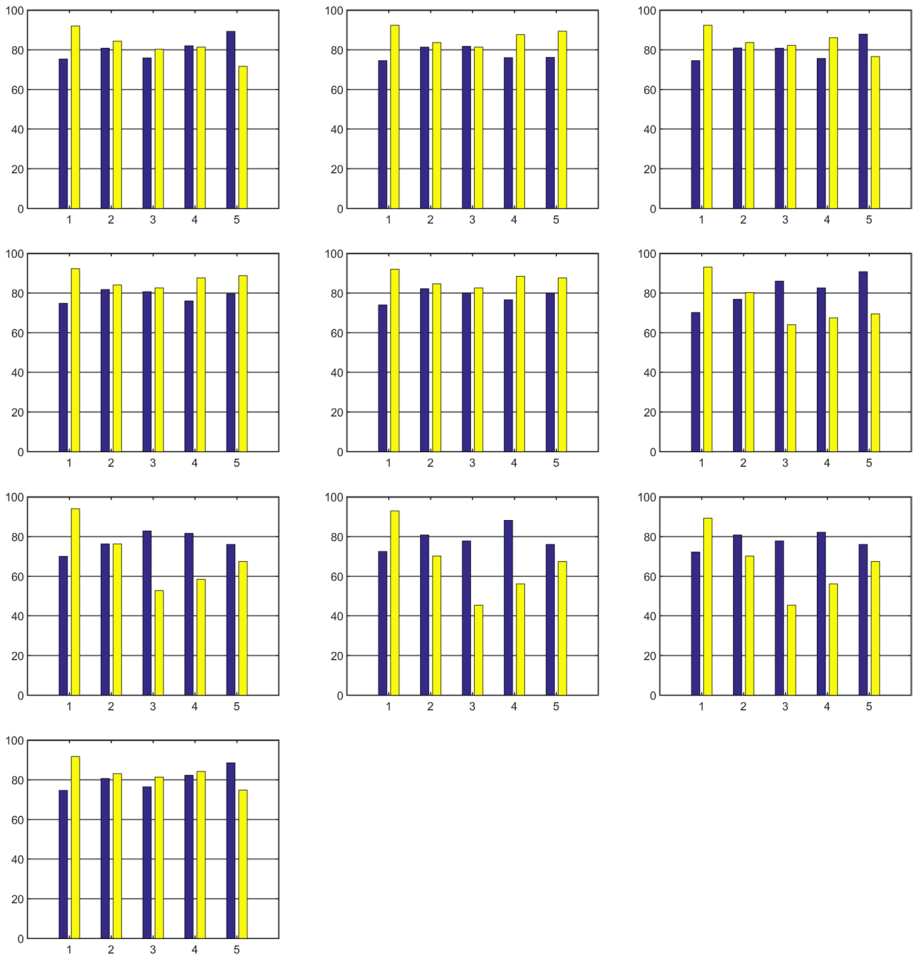

Fig. 2 Precision and recall for the VGG16 network with various classifiers. The color bars represent Pre and Rec in this order. The classifiers are arranged as follows: SVM, RBF, γ ="scale", C=15; SVM, RBF, $\gamma = 10^{-7}$, C=10; SVM, RBF, $\gamma = 10^{-7}$, C = 15; SVM, RBF, $\gamma = 10^{-7}$, C=20; SVM, polynomial of deg = 2; SVM, polynomial of deg = 3; DT, min−samples−leaf = 2; DT, min−samples−leaf = 3; RF, n−estimators = 100; RF, n−estimators = 200. The extra parameters of the DT are: criterion = "gini", random−state = 42, and max−depth =3

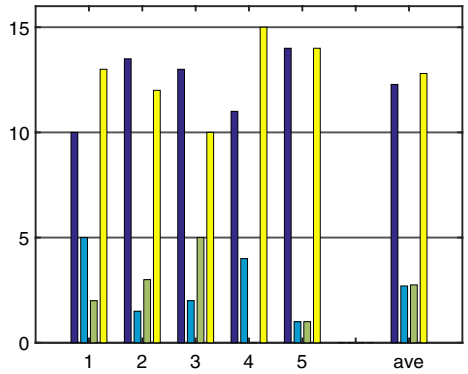

Fig. 3 Confusion matrices for 5-folds in the case of VGG16 network and selected classifier. The last confusion matrix is the average over all the folds. The classifier parameters are: SVM, RBF, $\gamma = 10^{-7}$, C=15

Table 3 Average confusion matrix for the VGG16 network and selected SVM classifier (comp. Figure 3)

Classifier; parameters	Confusion matrix	
SVM; RBF,	61.5	13.5
$\gamma = 10^{-7}$, C=15	11	64

It is often desirable to have a single measure instead of two, that is, recall and precision. For this purpose, one can use F_β score, which assumes that recall is β times more important than precision, that is

$$F_\beta = (1 + \beta^2) \times \frac{\text{Pre} \times \text{Rec}}{\beta^2 \times \text{Pre} + \text{Rec}} \tag{1}$$

A decision as to the relative importance of recall with respect to precision is somewhat arbitrary. As an example, we will give just two values, that is F_β for $\beta = 1$, for which both measures are equally important; and for $\beta = 2$, for which recall is two times more important than precision, that is

$F_1 = 83.93\,\%$ and $F_2 = 84.77\,\%$.

Experiments were also conducted for the VGG19 network replacing VGG16. The confusion matrices obtained for this case are shown in Fig. 12 in the Appendix. Comparing Figs. 11 and 12 one comes to the conclusion that a bigger and apparently more advanced network does not give any better results in the case under consideration. Hence it is recommended to use the VGG16.

3.2 Calculation of accuracy confidence intervals

The basic underlying process describing the operation of the classifier under consideration is the Bernoulli trial with 0 or 1 outcome. The probability of getting 1 is p, and probability of 0 is $q = 1 - p$. Performing Bernoulli trial n times, we obtain binomial distribution with unknown mean value np and unknown variance npq. Based on experimental results, we then want to estimate the mean and the variance or standard deviation. Dealing with binomial distribution is not the easiest one, and one tends to use the normal distribution as an

approximation. An advanced description of the theory involved is given in Wallis [29]. In the following, a detailed outline of practical calculations is described.

These calculations would be straightforward if we just had folds consisting of cases, but we also have patients. As a result, we can consider two approaches to the calculation of the confidence intervals. We will present them in the sequence.

The first approach is as follows. Consider a single fold, Fold 1 for example. We have classification results for 15 positive cases and 15 negative cases. For each positive case, there is a certain number of TP and FN observations, and for each negative case, there is a certain number of TN and FP observations. A popular metrics used in machine learning for evaluation of the classifiers based on neural networks is the accuracy, defined as

$\mathrm{Acc} = (\mathrm{TN} + \mathrm{TP}) / (\mathrm{TN} + \mathrm{FP} + \mathrm{FN} + \mathrm{TP})$.

The values of TN, FP, FN, TP can be calculated for every patient separately, so we get 30 numbers. In order to be able to do the calculations, we obtain the averages TN, FP, FN, TP. Obviously, this is an approximation. However, when acquiring the samples, we can tend to have a more or less even distribution of samples despite an uneven distribution of images per patient. It would not be practical to require that every patient be represented by a fixed number of images, since for one patient a single image "tells all" and for another, one needs a large number of images. The same problem is with scans, which can be quite different or almost the same.

An exemplary distribution of accuracies for positive patients in Fold 1 is shown in the left diagram of Fig. 4. An analogous diagram would be for negative patients. Obviously, we can calculate the average accuracy. The calculation of the 95% confidence interval CI follows the equations often used in machine learning

$$\mathrm{Range} = 1.96 \times \sqrt{\frac{\mathrm{Acc}(1-\mathrm{Acc})}{n}} \tag{2}$$

and

$$\mathrm{CI} = [\mathrm{Acc} - \mathrm{Range} < \mathrm{Acc} < \mathrm{Acc} + \mathrm{Range}], \tag{3}$$

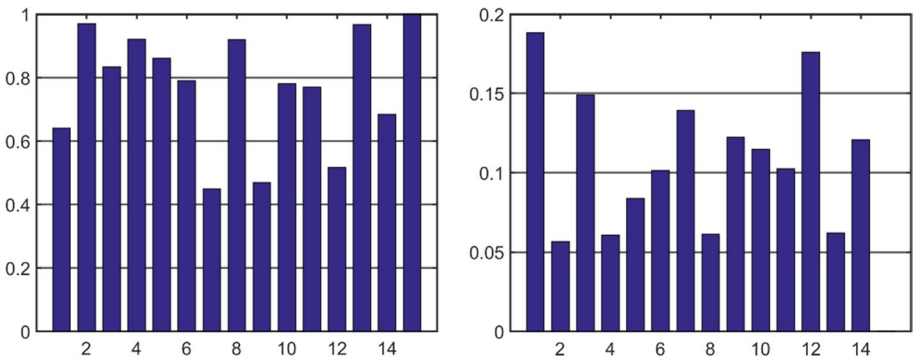

Fig. 4 Example of accuracies (left) and ranges (right) obtained for exemplary Fold 1 in the case of the VGG16-based classifier. The classifier parameters are: SVM, RBF, $\gamma = 10^{-7}$, $C = 15$

where auxiliary variable Range is used for convenience of presentation of the results below.

Equations 2 and 3 are valid for sampling from normal distribution. It is typically assumed that normal distribution can be used as an approximation to the binomial distribution if the number of samples is 30 or more, and this condition is satisfied since each fold contains 30 patients.

If we repeated the experiment over and over, each time drawing new examples, we would find that for approximately 95% of these experiments, the calculated confidence interval would contain the true accuracy.

The diagram on the right in Fig. 4 shows exemplary ranges corresponding to accuracies on the left in the same figure.

Figure 5 shows a bar diagram illustrating confidence interval lower boundary, accuracy, and upper boundary for all the folds of the VGG16-based classifier as well as the average confidence interval for all the folds.

The 95% confidence interval for the accuracy of 79.80% is [70.54, 89.05]%. The operation of averaging confidence intervals cannot dramatically improve the results, since the average is always restricted by maximum and minimum of the processed numbers.

The second approach to the calculation of confidence intervals gives better results. In this method, each case (scan) is treated separately and not considered as belonging to any particular patient. As a result, according to Table 2, we have, for example, $434 + 320 = 754$ cases in Fold 1. Significantly bigger number of cases gives a narrow confidence interval.

Figure 6 shows a bar diagram illustrating confidence interval lower boundary, accuracy, and upper boundary for 10 tested VGG16-based classifiers.

According to Fig. 6, the best classifier is represented by bar No 3 (one of the SVM's with RBF). The accuracy for this bar is 80.08% and the 95% confidence interval is [77.14, 83.04]%. It is clear from Fig. 6 that all SVM variants have practically the same accuracy and the same confidence intervals. This means that the corresponding maximum is relatively flat and the parameters can be changed in some range without major effect. The second best option is the Random Tree, which is almost as good as the SVM with RBF. The Random Tree is not recommended, however, because generating 100 or 200 trees takes longer time and there is no computational advantage to compensate for this

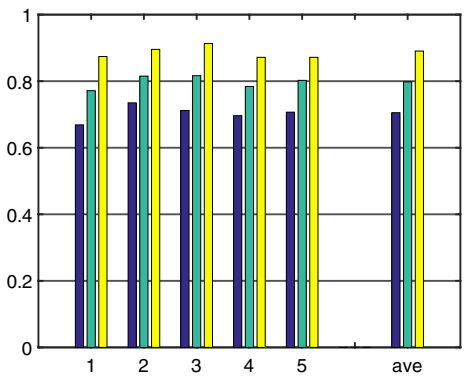

Fig. 5 95% confidence intervals of accuracy for the VGG16-based classifier. The last confidence interval is the average over all the folds. The classifier parameters are: SVM, RBF, $\gamma = 10^{-7}$, $C = 15$

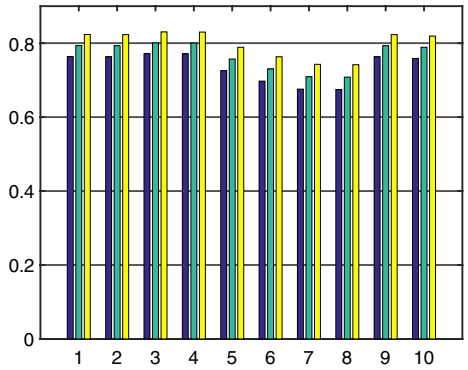

Fig. 6 95% confidence intervals of accuracy for tested classifiers. The classifiers, indexed 1–10 are arranged as follows: SVM, RBF, γ ="scale", C = 15; SVM, RBF, $\gamma = 10^{-7}$, C = 10; SVM, RBF, $\gamma = 10^{-7}$, C = 15; SVM, RBF, $\gamma = 10^{-7}$, C = 20; SVM, polynomial of deg = 2; SVM, polynomial of deg = 3; DT, min−samples−leaf = 2; DT, min−samples−leaf = 3; RF, n−estimators = 100; RF, n−estimators = 200. The extra parameters of the DT are: criterion = "gini", random−state = 42, and max−depth=3

drawback. The remaining classifiers, for example, this represented by bar No 8 (Decision Tree), have the confidence interval below that of the best classifier, and there is little or no overlap between respective confidence intervals.

Comparing the accuracies obtained for the best classifiers in our two approaches: 79.80 vs. 80.08% we see that they are close enough. The second approach to the calculation of the confidence intervals has another important advantage not reflected in the numbers considered above. The first approach requires testing all the cases one by one, that means we have to process all the data set using batch of size one, and this takes a lot of work. In the second approach the batches contain several hundred cases each, and this greatly speeds up computations.

3.3 Execution time

The execution time of the described calculations in Google Colab is hard to reliably estimate, because it varies in a wide range. Since the weights of the network are imported from keras and fixed, a substantial portion of calculations consists in predicting the features at appropriate maximum pooling layers of the network. These features are subsequently used in the external classifier that operates in a usual way and processes the training data as well as testing data. It was observed that processing a single maximum pooling layer in the case of VGG16 took 20–60 min when using the GPU option. This time tended to be shortest for processing the features of maximum pooling layer 3 and longest for layer 5.

3.4 Examples of classification

Several examples of correctly classified psoriasis (TP) and non-psoriasis (TN) scans are shown in Figs. 7 and 8, respectively. It is impossible to draw any definite conclusions as to the visual differences between psoriasis and non-psoriasis perceived by a human based on these sample images, since one could easily provide other samples which would look quite different.

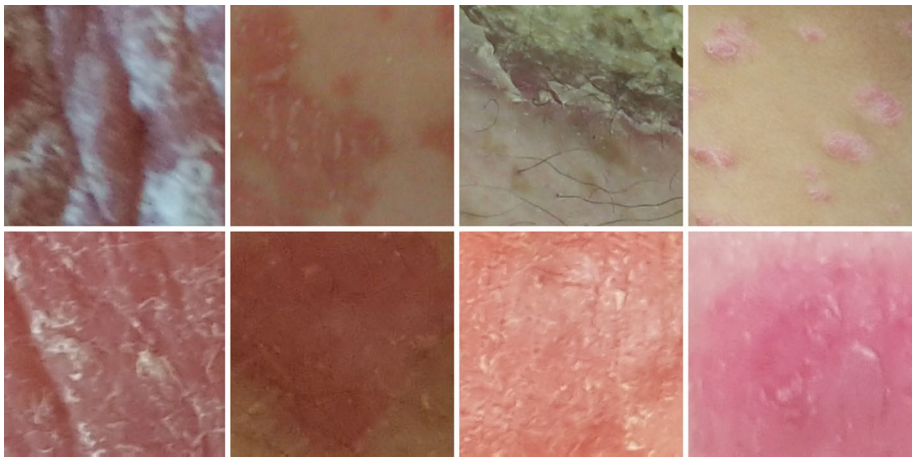

Fig. 7 Examples of correctly classified psoriasis scans (TP)

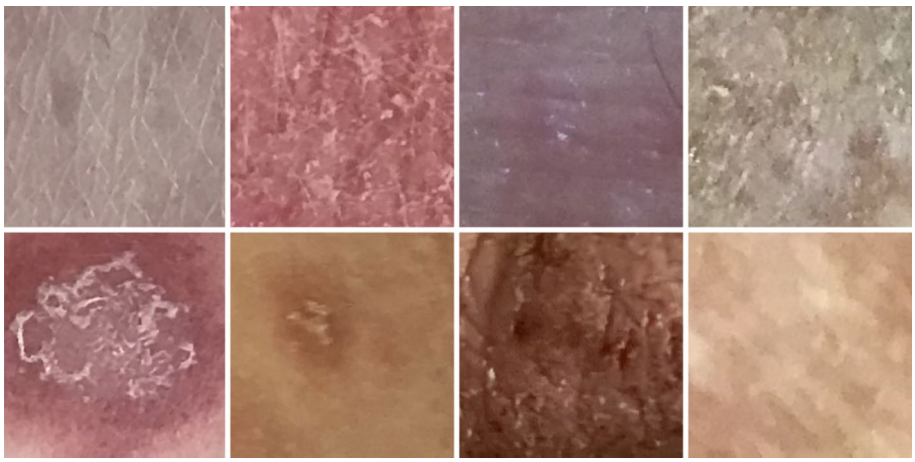

Fig. 8 Examples of correctly classified non-psoriasis scans (TN). The images are shown in the following order: granuloma annulare, head lice, mycosis fungoides, lichen sclerosus; subac. cuta. lupus erythematosus, parapsoriasis, lymphoma, mycosis

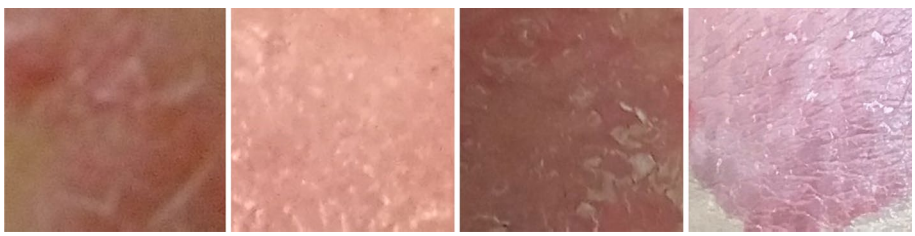

Fig. 9 Examples of psoriasis scans classified incorrectly (FN)

Of great interest are cases, where the majority voting gave an incorrect result. According to Table 3, there are 11 FN psoriasis patients. Examples of misclassified psoriasis scans of these patients are shown in Fig. 9. Then, there are 16.5 FP non-psoriasis patients (one half corresponding to a tie in voting). The misclassified scans

were selected from deeply misclassified patients. In particular, for seborrhoeic dermatitis the number of scans TN = 0, FP = 17; for herpes zoster (1) TN = 6, FP = 24; for herpes zoster (2) TN = 5, FP = 16; for pemphigus TN = 1, FP = 24. This means that future investigations should pay attention to particularly hard cases as exemplified by Figs. 9 and 10.

It is known that due to auxiliary equipment, dermoscopic images represent a more or less standardized way of acquiring an image of a skin lesion. In contrast, clinical images use out-of-shelf conventional cameras often embedded in smart phones. On top of this, there are common problems due to variations in the environmental conditions while capturing an image, such as varying light sources, and inhomogeneous illumination additionally changing from one image to another, which gives rise to varying shadows or reflections/specularities. Other problems that can appear are: irrelevant lesions of any possible kind, marks or objects, lesions captured incompletely, low degree of focusing, widely varying acquisition distance and angle. When using digital images as input to a psoriasis classification algorithm, one should tend to follow an acquisition procedure that would be the same for all of the images. In particular the acquisition distance and angle preferably be within certain ranges, and illumination requirements should be defined.

4 Conclusions

The obtained results confirm that the proposed method based on the use of the deep convolutional network for feature generation together with the SVM classifier is suitable for differentiating psoriasis from other dermatoses. The achieved precision is 82.58% and recall (sensitivity) is 85.33% for clinical images. The 95% confidence interval for the accuracy of 80.08% is [77.14, 83.04]%.

The advantage of the proposed approach is that by necessity it is well suited to local conditions, in particular to the local frequency distribution of various dermatoses and can be easily modified if new cases are collected.

Further research might be conducted aiming at the differentiation of psoriasis from particular dermatoses, such as seborrhoeic dermatitis, herpes zoster, and pemphigus, for which the skin samples might appear quite similar to psoriasis; however, this necessitates collecting much more examples of the relevant images. Keeping in mind that presented results were obtained on images acquired by the staff using smart phones without following any particular recommendations regarding image acquisition procedure, one concludes that the proposed approach turned out quite resistant to unfriendly

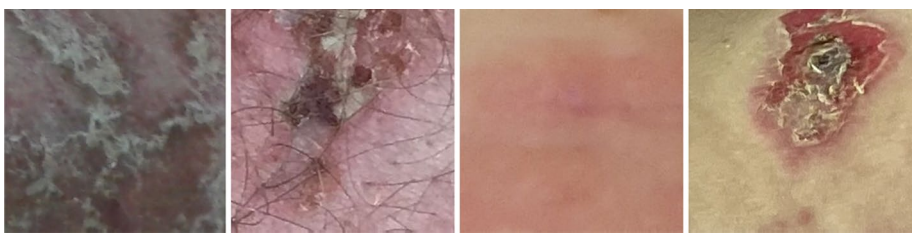

Fig. 10 Examples of non-psoriasis scans classified incorrectly (FP). The images are shown in the following order: seborrhoeic dermatitis, herpes zoster, herpes zoster, pemphigus

conditions, and for this reason, it may be useful for telemedicine applications, for example, in geographical areas, where dermatology specialists may be scarce or unreachable.

Appendix

See Figs. 11, 12.

Fig. 11 Confusion matrices for the VGG16 network with various classifiers. The color bars represent TN, FP, FN, and TP in this order. The classifiers are arranged as follows: SVM, RBF, γ ="scale", C = 15; SVM, RBF, $\gamma = 10^{-7}$, C = 10; SVM, RBF, $\gamma = 10^{-7}$, C = 15; SVM, RBF, $\gamma = 10^{-7}$, C = 20; SVM, polynomial of deg = 2; SVM, polynomial of deg = 3; DT, min−samples−leaf = 2; DT, min−samples−leaf = 3; RF, n−estimators = 100; RF, n−estimators = 200. The extra parameters of the DT are: criterion = "gini", random−state = 42, and max−depth=3

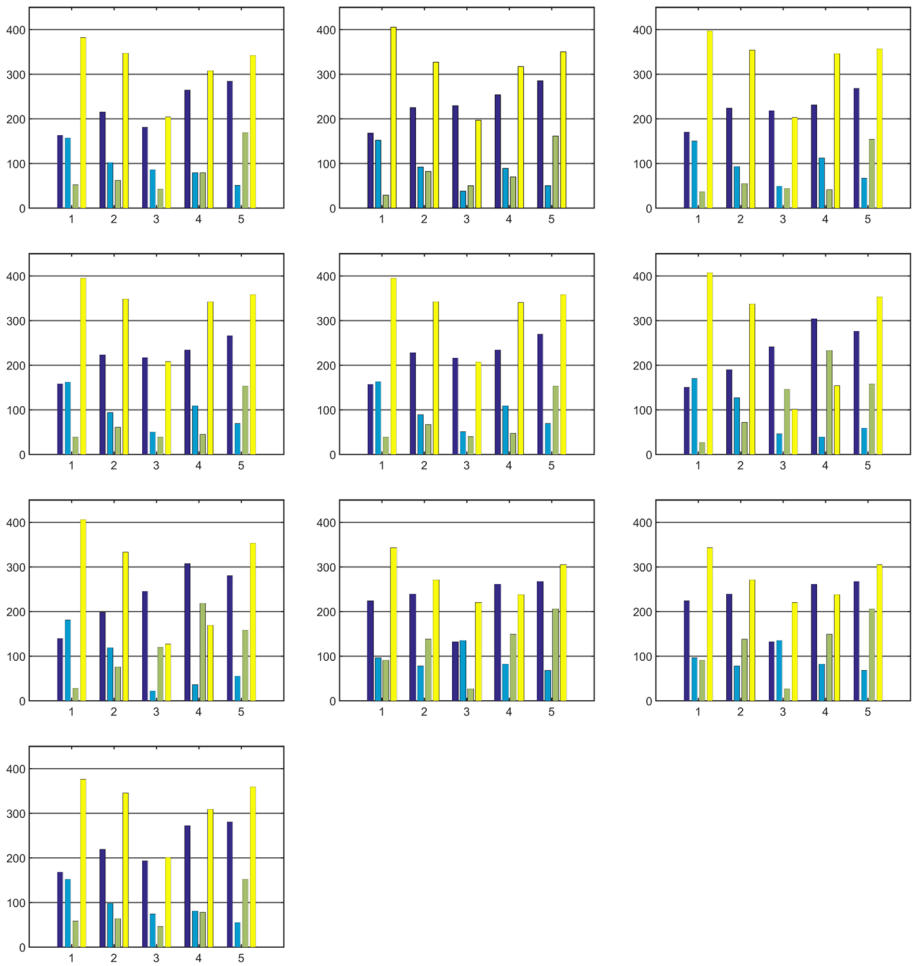

Fig. 12 Confusion matrices for the VGG19 network with various classifiers. The color bars represent TN, FP, FN, and TP in this order. The classifiers are arranged as follows: SVM, RBF, γ ="scale", C = 15; SVM, RBF, $\gamma = 10^{-7}$, C = 10; SVM, RBF, $\gamma = 10^{-7}$, C = 15; SVM, RBF, $\gamma = 10^{-7}$, C=20; SVM, polynomial of deg = 2; SVM, polynomial of deg = 3; DT, min−samples−leaf = 2; DT, min−samples−leaf = 3; RF, n−estimators = 100; RF, n−estimators = 200. The extra parameters of the DT are: criterion = "gini", random−state = 42, and max−depth=3

Abbreviations

DT	Decision tree
FN	False negative
FP	False positive
RF	Random forest
RBF	Radial basis function
SVM	Support vector machine
TN	True negative
TP	Truepositive

Acknowledgements
The authors are grateful to the Faculty of Mathematics and Informatics of the University of Lodz for providing the facilities to carry out the research.

Author contributions
MN: conceptualization, software, validation, writing—original draft/review. LJC: data curation, software, writing—review. SP: image acqusition, data curation, writing—review. AW: image acquisition, data curation, clinical studies, writing—review. All authors read and approved the final manuscript.

Funding
No special funding for the research described in this paper was provided.

Availability of data and materials
Reasonable requests for the data set of used images will be considered by the authors and should be accepted by the Medical University of Lodz, in Lodz, Poland.

Declarations

Ethics approval and consent to participate
This study was approved by the Medical Ethics Committee of the Medical University of Lodz (RNN/186/17/KE, June 13, 2017).

Consent of publication
Not applicable.

Competing interests
The authors have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 2 June 2022 Accepted: 30 March 2023
Published online: 15 May 2023

References

1. E. Higgins, Psoriasis. Medicine **49**(6), 361–69 (2021)
2. R. Morris-Jones (ed.), ABC of Dermatology. (Wiley, Hoboken, 2019)
3. H. Hashim, M.N. Taib, N.S.Z. Abidin, E.A. Akmar, Statistically discrimination of psoriasis lesions with chromatic color indices. In: International Federation for Medical and Biological Engineering Proc. of 4th Kuala Lumpur International Conference on Biomedical Engineering, pp. 619–23 (2008)
4. N.K. Al-Abbadi, N.S. Dahir, M.A. Al-Dhalimi, H. Restom, Psoriasis detection using skin color and texture features. J. Comput. Sci. **6**(6), 648–52 (2010)
5. L. Ballerini, R.B. Fisher, B. Aldridge, J. Rees, A color and texture based hierarchical k-NN approach to the classification of non-melanoma skin lesions, in *Color Medical Image Analysis*. ed. by M.E. Celebi, G. Schaefer (Springer, Dordrecht, 2013), pp.63–86
6. V.K. Shrivastava, N.D. Londhe, S. Sonawane, J.S. Suri, Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. Expert Syst. Appl. **42**(15–16), 6184–95 (2015)
7. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2015). arXiv:1409.1556
8. A. Gupta, M. Gupta, Transfer learning for small and different datasets: fine-tuning a pre-trained model affects performance. J. Emerg. Res. **3**, 5 (2020)
9. G. An, M. Akiba, K. Omodaka, T. Nakazawa, H. Yokota, Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images. Sci Rep Nat **11**(4250), 9 (2021)
10. L. Alzubaidi, M.A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, S.R. Oleiwi, Towards a better understanding of transfer learning for medical imaging: a case study. Appl. Sci. **10**(13) (2020)
11. L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A.J. Humaidi, O. Al-Shamma, M.A. Fadhel, J. Zhang, J. Santamaría, Y. Duan, Novel transfer learning approach for medical imaging with limited labeled data. Cancers **13**(7) (2021)
12. D.T. Hogarty, J.C. Su, K. Phan, M. Attia, M. Hossny, S. Nahavandi, P. Lenane, F.J. Moloney, A. Yazdabadi, Artificial intelligence in dermatology-where we are and the way to the future: a review. Am. J. Clin. Dermatol. **21**(1), 41–47 (2020)
13. M.E. Celebi, H.A. Kingravi, B. Uddin, H. Iyatomi, Y.A. Aslandogan, W.V. Stoecker, R.H. Moss, A methodological approach to the classification of dermoscopy images. Comput. Med. Imaging Graph. **31**(6), 362–73 (2007)
14. S. Kim, J. Kim, M. Hawng, M. Kim, S.J. Jo, M. Je, J.E. Jang, D.H. Lee, J.Y. Hwang, Smartphone-based multispectral imaging and machine-learning based analysis for discrimination between seborrheic dermatitis and psoriasis on the scalp. Biomed. Opt. Express **10**(2), 879–91 (2019)
15. N. Hameed, F. Hameed, A. Shabut, S. Khan, S. Cirstea, A. Hossain, An intelligent computer-aided scheme for classifying multiple skin lesions. Computers **8**(3), 12 (2019)
16. N. Mittal, S. Tanwar, S.K. Khatri, Identification & enhancement of different skin lesion images by segmentation techniques. In: Proc. of 6th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), pp. 609–14 (2017)
17. J.S. Taur, G.H. Lee, C.W. Tao, C.C. Chen, C.W. Yang, Segmentation of psoriasis vulgaris images using multiresolution-based orthogonal subspace techniques. IEEE Trans. Syst. Man. Cybern. Part B Cybern. **36**(2), 390–402 (2019)
18. T.V. Tien, N.H. Phuc, L.Q. Nhien, T.T. Trang, D.S. Hieu, P.N. Cat, P.T. Mien, H.Q. Linh, Evaluation of scaly levels in psoriasis multispectral polarized imaging. In: 6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6) In IFMBE Proceedings 63, pp. 97–101 (2018)
19. L.S. Wei, Q. Gan, T. Ji, Skin disease recognition method based on image color and texture features. Comput. Math. Methods Med. (2018)
20. A. Udrea, G.D. Mitra, Generative adversarial neural networks for pigmented and non-pigmented skin lesions detection in clinical images. In: (IEEE) 2017 21st International Conference on Control Systems and Computer Science, pp. 364–68 (2017)
21. L. Peng, Y. Na, D. Changsong, L.I. Sheng, M. Hui, Research on classification diagnosis model of psoriasis based on deep residual network. Digital Chin. Med. **4**(2), 92–101 (2021)

22. Y. Yang, J. Wang, F. Xie, J. Liu, C. Shu, Y. Wang, Y. Zheng, H. Zhang, A convolutional neural network trained with dermoscopic images of psoriasis performed on par with 230 dermatologists. Comput. Biol. Med. **139**, 104924 (2021)
23. Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, M. Fujimoto, Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br. J. Dermatol. **180**, 373–81 (2019)
24. M. Mikołajczyk, S. Patrzyk, M. Nieniewski, A. Woźniacka, Evaluation of a smartphone application for diagnosis of skin diseases. Adv. Dermatol .Allergol **5**, 761–66 (2021)
25. P. Kora, C.P. Ooi, O. Faust, U. Raghavendra, A. Gudigar, W.Y. Chan, K. Meenakshi, K. Swarajaa, P. Plawiak, U.R. Acharya, Transfer learning techniques for medical image analysis: a review. Biocybern. Biomed. Eng. **42**, 79–107 (2022)
26. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, Holland, 2009)
27. F. Pedregosa, G. Varoquaux, A. Gramfort, scikit-learn, API Reference. https:*//scikit-learn.org/stable/modules/classes.html
28. F. Pedregosa, G. Varoquaux, A. Gramfort et al., Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–30 (2011)
29. S. Wallis, Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. J. Quant. Linguist. **20**(3), 178–208 (2013)

Publisher's Note