Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16529-w

Fine-tuning pre-trained neural networks for medical image classification in small clinical datasets

Newton Spolaôr, Huei Diana Lee, Ana Isabel Mendes, Conceição Veloso Nogueira, Antonio Rafael Sabino Parmezan, Weber Shoity Resende Takaki, et al. ***[full author details at the end of the article]***

Received: 6 October 2022 / Revised: 19 May 2023 / Accepted: 13 August 2023

Abstract
Convolutional neural networks have been effective in several applications, arising as a promising supporting tool in a relevant Dermatology problem: skin cancer diagnosis. However, generalizing well can be difficult when little training data is available. The fine-tuning transfer learning strategy has been employed to differentiate properly malignant from non-malignant lesions in dermoscopic images. Fine-tuning a pre-trained network allows one to classify data in the target domain, occasionally with few images, using knowledge acquired in another domain. This work proposes eight fine-tuning settings based on convolutional networks previously trained on ImageNet that can be employed mainly in limited data samples to reduce overfitting risk. They differ on the architecture, the learning rate and the number of unfrozen layer blocks. We evaluated the settings in two public datasets with 104 and 200 dermoscopic images. By finding competitive configurations in small datasets, this paper illustrates that deep learning can be effective if one has only a few dozen malignant and non-malignant lesion images to study and differentiate in Dermatology. The proposal is also flexible and potentially useful for other domains. In fact, it performed satisfactorily in an assessment conducted in a larger dataset with 746 computerized tomographic images associated with the coronavirus disease.

Keywords Feature learning · Few-shot learning · RMSprop · Shallow learning · Statistical test · VGG

1 Introduction

In the medical diagnosis process, imaging exams play an increasingly central role, as they allow access to human anatomy and physiology with an increasingly realistic accuracy of pathophysiological processes. These imaging exams, by their nature, enable them to be used as a data source for decision support methods in Medicine, based on automatic learning,

✉ Newton Spolaôr
newtonspolaor@gmail.com

Extended author information available on the last page of the article

Machine Learning (ML) and, more recently, Deep Learning (DL) models. Many Computer-Aided Diagnostic systems (CAD), in turn, rely on ML methods to extract patterns from images that can represent useful knowledge, as illustrated in different medical domains [7, 23, 25].

The use of artificial intelligence models as a tool to support clinical decision is also significant in cases where early diagnosis is crucial for the prognosis [54], such as in the case of cutaneous or colorectal tumors, or the identification of pathognomonic imaging signs associated with known diseases.

In the case of skin cancer, DL has performed well in many ML applications to assist in classifying dermoscopic images. Several techniques associated with DL were exploited in [26] to differentiate Non-malignant (N) from Malignant (M) skin lesions. The network architectures AlexNet and VGG16 achieved competitive results in the PH^2 public dataset with benign lesions and melanoma in [27]. Among the different DL architectures applied to PH^2 in [45], VGG19 performed better. Before classifying 900 dermoscopic images with VGG16, the authors conducted lesion segmentation with the U-Net architecture in [60]. The authors achieved positive results in [61] by applying DL preceded by the Gabor wavelets technique in melanoma and seborrheic keratoses detection.

A reason for DL success involves the possibility of transferring the knowledge learned by a deep neural network in a large image dataset to improve the training of a classifier in a relatively small dataset [22, 50]. Researchers have developed varied strategies, such as Fine-Tuning (FT) a network previously trained in a large set of generic images, to achieve this transference in Dermatology [47, 69], Mastology [23], Ophthalmology [7] and other fields. Such transference may benefit from the crescent availability of benchmark datasets.

These strategies could mitigate limitations that would arise in practical scenarios if a classification model were built directly from a dataset with few images [48, 72], regardless of the medical field. For example, some health institutions often have small sets of images acquired from the same equipment, according to the same environmental conditions, which are not adequately represented by benchmark datasets. In this scenario, DL algorithms trained only with these limited collections might overfit. On the other hand, combining the algorithms with transfer learning strategies could deliver better achievements to the institutions, by reusing models learned from big datasets and tailoring them to classify small sets.

This study can be linked to the few-shot learning paradigm, a maturing subarea of machine learning that focuses on enabling models to rapidly learn and generalize from a limited number of labeled examples [73]. Although researchers have investigated distinct techniques and frameworks to apply few-shot learning in small clinical datasets [39, 55, 65], our proposal differentiates from them by employing transfer learning via several FT settings in two domains. In particular, these settings (1) do not depend on typical few-shot learning procedures for deep learning, such as data augmentation or meta-learning, and (2) have not been experimentally compared in dermoscopic and COVID-19 datasets, in which the smallest one has only a few dozen instances per class for both classes.

This work proposes eight settings to fine-tune VGG-based networks pre-trained in ImageNet and evaluates them in two public datasets with 104 and 200 dermoscopic images. As the datasets are relatively small — they have less than 1,000 images [21, 35, 71] —, the work contributes by illustrating that DL can be effective if an institution has only a few dozen malignant and non-malignant lesion images to study and classify in Dermatology. In fact, our experimental assessment supported this novelty by showing that our best network settings in tiny datasets, supported by trivial pre-processing procedures, (1) can generalize knowledge and (2) achieve results competitive with those reported in recent work that targets larger datasets.

To verify the flexibility of our proposal, an additional evaluation with one of the best VGG-based configurations was carried out in a medical domain different from Dermoscopy. In particular, we fed the network with a public dataset containing 746 Computerized Tomographic (CT) images, from which 349 are associated with the coronavirus disease 2019 (COVID-19). As a result, a classification performance considered satisfactory was found in a group of images larger than the two previous ones.

2 Background

In short, deep neural networks consists of artificial neural networks with a sequence of tens or hundreds of hidden layers [22]. Although many basic concepts underlying DL were already known in 1990s and 2000s, successful and competitive DL applications were found only after the emergence of recent hardware/software improvements. Examples include the use of Graphical Processing Unit (GPU) to support the parallelization of network operations and the advent of algorithms and techniques to better propagate the feedback signal across many network layers.

The Root Mean Square propagation (RMSprop) optimization scheme is one of the techniques that has contributed to gradient propagation in many deep neural networks settings. In short, RMSprop provides networks with an adaptive Learning Rate (LR) in a mini-batch of data. The scheme to achieve adaptation when defining the weight w considers the division of the setting LR by the moving average of the magnitudes of recent gradients for w [30]. Taking into account previous gradients to estimate the next weight update can improve the convergence speed during training compared with classic stochastic gradient descent.

Convolutional neural networks, a.k.a. convnets, have been one of the most applied DL schemes for computer vision tasks, including image analysis and classification [18, 24, 70]. A reason includes the high performance achieved by convnets in machine learning competitions, such as the ImageNet Large Scale Visual Recognition Challenge[1], which is associated with the large dataset ImageNet with 1.4 million images labeled by 1,000 different classes. Another reason is that, differently from several shallow learning algorithms[2], convnets can automate feature learning. As a result, its user does not need to engineer characteristics that properly describe images.

Convolutional, max pooling and Fully Connected (FC) layers are three components commonly used in convnets. The first one learns local patterns from small windows representing pieces of the input image. In particular, it applies specific filters to an image, yielding a set of feature maps. Each map consists of a tensor reflecting the presence of a filter/feature pattern at distinct input locations. In turn, the Max pooling layer downsamples the feature map received as input by replacing each group of adjacent map tiles with a tile defined by the maximum value of this group [26]. This downsampling procedure is often aggressive and configured by a factor of 2, which helps reduce the number of network parameters. An FC or dense layer, usual in shallow neural networks, fundamentally differs from a convolutional layer by connecting each of its neurons with every neuron from the previous layer. This property promotes the learning of global patterns from the input.

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

[2] Algorithms that transform the input data into one or two successive representations spaces, such as Support Vector Machines (SVM) [22].

The VGG family of neural network architectures [64] illustrates how the previous components can be combined. In particular, VGG starts with a series of convolutional and max pooling layers (the convolutional base) and ends with a sequence of FC layers. The family architectures differ in terms of the number of weight (convolutional) layers, the number of filters and the size of the image windows. Other examples of architecture families include EfficientNet [32] and ResNet [56].

According to the literature, convnets usually perform well when trained on large datasets [20, 39]. However, their effectiveness is contingent upon the availability of a sufficient number of labeled examples in the target domain. When the target domain has few examples for training, convnets cannot rapidly adapt to new target regions due to insufficient network activation variables. In this sense, few-shot classification introduces alternative training schemes that enable models to learn from limited labeled data [39, 55, 65].

Few-shot image classification research is in the early stages of development, often drawing on strategies involving deeper networks to improve model accuracy significantly [76]. Much emphasis is currently given to experiments compared to theoretical studies and practical applications, indicating gaps in knowledge that need to be bridged [62]. Techniques such as data augmentation [19], meta-learning [43, 65], metric learning (similarity-based methods) [40], and transfer learning [6] have been adopted in this machine learning subarea to train models on small image sets, allowing them to adapt quickly and generalize to new, unseen data.

Transfer Learning (TL) underlies many popular convnets' applications. This concept makes it possible to have different feature spaces or data distributions in the source and target domains [50]. For example, TL can use knowledge previously learned from a dataset by a specific classifier to improve the learning ability of another classifier in a different dataset. Inductive transfer learning is particularly useful if both source and target domains are labeled, such that inductive biases taken from the former problem assist instance prediction in the latter one.

Inductive TL can be helpful when the training data available in a target problem is insufficient to build a classification model that generalizes well [48]. In this scenario, a neural network, for example, could be trained in a large dataset and the resulting model, including its weights, could be transferred to the target domain. Two strategies based on inductive TL have been employed to classify small image datasets with convnets [22]: (1) feature extraction with and (2) fine-tuning a pre-trained network. The former strategy replaces the dense layers on the top of a convolutional base with a new classifier, e.g., an SVM model [45] or a custom network. The idea is to use the base only to learn the features that will feed the added classifier, which is trained in the target problem. In turn, the latter strategy extends the architecture defined by the former one by performing two additional steps. First, the highest convolutional base layers are unfrozen, *i.e.*, they are allowed to update their weights. Secondly, the architecture with the new classifier and the unfrozen layers is trained in the target domain. As a result, the highest (more abstract) representations from the transferred model can be slightly adapted to the data at hand.

The next sections detail how this work fine-tunes convnets previously trained in ImageNet and evaluates them with the aim of classifying small datasets with dermoscopic images.

3 Materials and methods

Section 3.1 describes the image datasets and cites the computational software considered in this work. Then, Sections 3.2 and 3.3 address, respectively, the eight deep learning settings defined by us and the experimental setup in which they are evaluated.

3.1 Materials

This work used two datasets with skin images. The first one, described in detail in [38], has 104 dermoscopic images. It should be emphasized that the 58 malignant and 46 non-malignant lesion images in this dataset are RGB true colored (24-bit color) and JPEG compressed with a minimum resolution of 300 dpi. In this work, malignant lesions comprise melanoma and carcinoma, while non-malignant ones include blue nevus and other benign conditions [36]. Image acquisition followed all legal requirements, in accordance with dermoscopic protocols [16]. Each image was cropped at 450× 600 resolution and segmented as reported in [42, 51, 52], yielding the input for the method described in the next section.

The second dataset, PH^2, has 200 dermoscopic examinations [46] and was also used in related papers, as illustrated in [2, 8, 27, 45]. It contains 40 images with melanoma (malignant) lesions and 160 images with non-malignant lesions, from which 80 are common nevi and 80 are atypical nevi. The images are also 24-bit color RGB, but their file format is BMP and the image dimensions vary. Although PH^2 provides, for each image, a binary mask defined by experts containing the lesion, this work did not consider this annotation. In other words, we employed the complete images as inputs for DL.

This work includes a third dataset, COVID-CT [78], with 746 CT slices, to verify the flexibility of our proposal. Altogether, it consists of 349 images containing clinical findings of COVID-19 from 216 patients, as well as 397 slices without those findings from 171 people. The media was collected by COVID-CT authors directly from scientific papers. A senior radiologist, who worked with several coronavirus patients, has confirmed its utility. Furthermore, the dataset was used in scientific work [28, 63].

We employed convolutional neural networks implemented in the Keras high-level library with TensorFlow backend [1]. Additional image pre-processing procedures and part of the network assessment was also performed with Keras functions. This work applied the GraphPad Prism software to support the conduction of statistical tests and some functions of the Multi-label Exploratory Data Analysis (ML-EDA) tool [17] to plot a few graphics.

3.2 Deep neural network settings

The method underlying all the settings defined in this work consists of three steps, which are described in the following sections.

3.2.1 Step 1 — Image pre-processing

Step 1 transforms the medical images in a dataset into batches of pre-processed tensors (multidimensional arrays). In particular, the images are converted into floating-point tensors with normalized values and grouped into batches. The images are also resized to fixed dimensions during this process, as some convolutional networks including dense layers (e.g., VGG [64]) expect fixed-size images as input.

3.2.2 Step 2 — Image classification

Step 2 performs image classification based on deep neural networks that received the batches generated from the previous step. In this work, we considered two network architectures from the VGG family previously trained on ImageNet: VGG16 and VGG19. Figure 1 shows

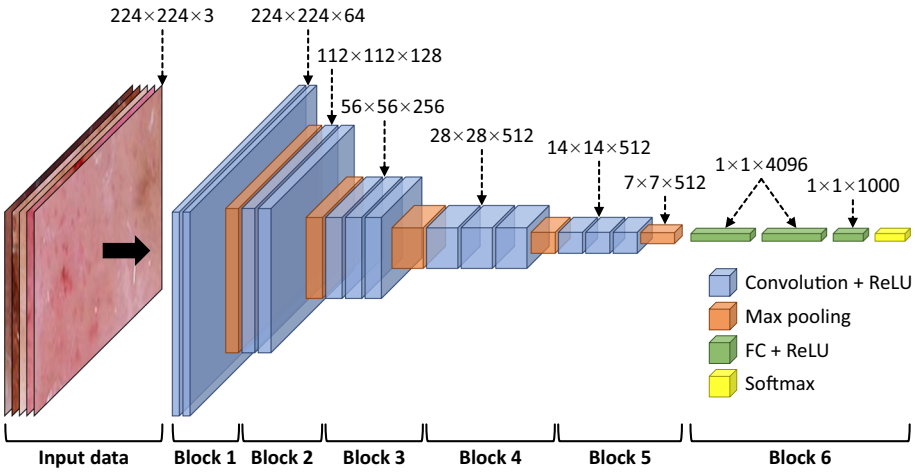

Fig. 1 VGG16 architecture

the first architecture with six-layer blocks and the use of Rectified Linear Unit (ReLU) and Softmax activation functions [22, 67].

VGG19 is similar to VGG16 but includes one convolutional layer before max pooling in Blocks 3, 4 and 5 to yield 19 weight layers. Note that the sequence starting from Block 1 to Block 5, which occurs before the fully connected layers, constitutes the VGG convolutional base.

To deal with a small dataset, we fine-tuned pre-trained VGG16 and VGG19 by following six procedures:

1. Remove Block 6, as it was designed to classify images into 1000 ImageNet classes, which are not relevant in some domains, such as Dermoscopy.
2. Insert a new Block 6 on top of the network with two fully connected layers (custom network).
3. Freeze from Block 1 to Block 5 (convolutional base).
4. Train the custom network in images from the small dataset.
5. Unfreeze from Block B to Block 5 after defining the index B according to the chosen network setting.
6. Train the custom network together with the unfrozen blocks from the convolutional base in images from the small dataset.

Altogether, this work defined eight network fine-tuning settings by combining unique variations of three variables described in what follows. Table 1 displays all the variations.

$V1$. Base architecture.
$V2$. Block index B.
$V3$. Learning rate used in the 4^{th} and 6^{th} procedures.

3.2.3 Step 3 — Classifier evaluation

Each network setting was evaluated individually to assess its ability to differentiate malignant from non-malignant lesion images in the dermoscopic datasets described in Section 3.1. After estimating the settings' performance, we chose the best ones to compare them with recently published methods and applied one of them to classify CT scans.

Table 1 Eight deep neural network settings defined and evaluated in this work

	V1		V2	
	VGG16	VGG19	B value	
			3	4
FT1	x		x	
FT2	x		x	
FT3	x			x
FT4		x	x	
FT5	x			x
FT6		x	x	
FT7		x		x
FT8		x		x
		V3		
	constant	higher rate in the 4^{th} procedure and smaller rate in the 6^{th} one		
FT1	x			
FT2			x	
FT3	x			
FT4	x			
FT5			x	
FT6			x	
FT7	x			
FT8			x	

3.3 Experimental setup

All the settings evaluated in this work have two dense layers in Block 6 — Fig. 1. The lowest layer contains 256 hidden units, while the highest one has one hidden unit used for classification purposes. The corresponding activation functions are, respectively, ReLU and sigmoid, while RMSprop was adopted as the optimizer.

We set the learning rate as 1×10^{-5} for settings FT1, FT3, FT4 and FT7. In turn, the remaining settings employ LR=2×10^{-5} and LR=1×10^{-5}, respectively, in the 4^{th} and 6^{th} training procedures reported in Section 3.2. Regardless of the conducted training procedure, the number of steps per epoch was the same: 100. As is the case with the previous parameter values, this number was defined based on examples described in [22].

Four evaluation measures were adopted in this work. Accuracy (Acc) is defined as the number of correct classifications obtained by the classifier under assessment; Sensitivity (Sen) or true positive rate measures the proportion of malignant lesions correctly identified as such; Specificity (Spe) or true negative rate measures the proportion of non-malignant lesions that are identified correctly as such; and F1 score (F1) corresponds to the harmonic mean of precision — number of true positives divided by the number of lesions predicted as malignant — and sensitivity. These measures are respectively defined by (1), (2), (3) and (4) in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) predicted by a classifier C. As the Keras tool did not directly support Sen,

Spe and F1 measures, we implemented them in this work.

$$Acc(C) = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

$$Sen(C) = \frac{TP}{TP + FN}. \tag{2}$$

$$Spe(C) = \frac{TN}{TN + FP}. \tag{3}$$

$$F1(C) = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{4}$$

Cross-Validation (CV) supported the assessment of the DL settings. In particular, each dataset of dermoscopic images was split into five pairs of training and testing folds. In turn, each training fold was split into three pairs of training and validation folds. Figures 2 and 3 depict this process for each dataset and indicates the number of malignant and non-malignant lesion images, as well as the total number of images per fold. Note that each fold approximates the class distribution from the input dataset, qualifying our CV one [72].

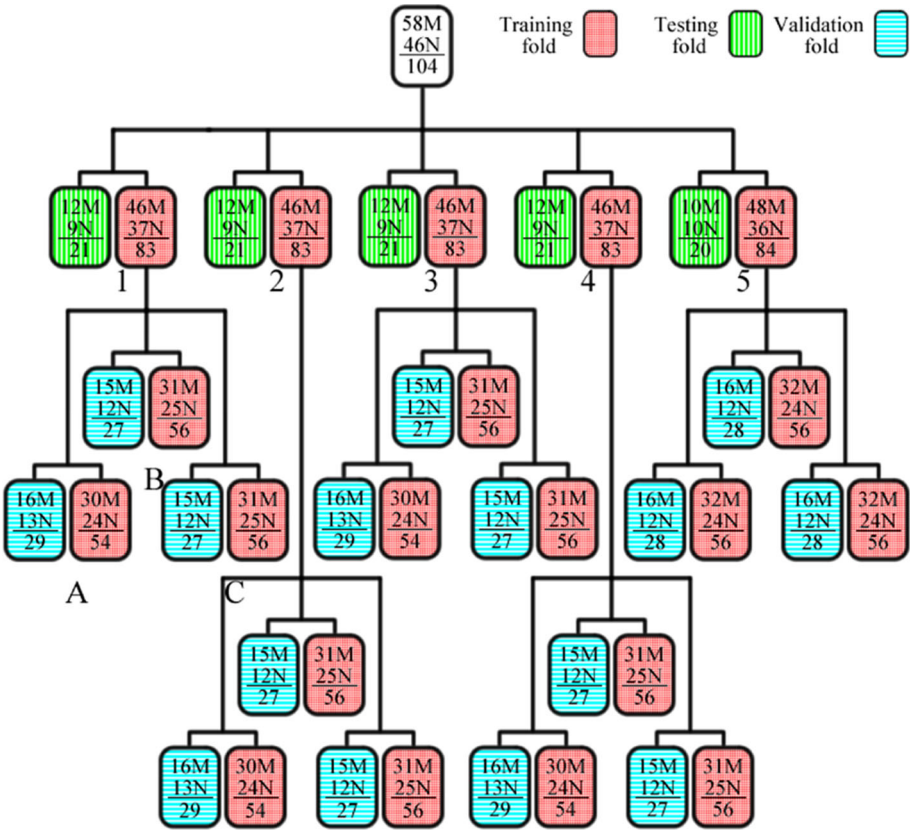

Fig. 2 Number of malignant and non-malignant lesion images in each cross-validation fold from the first dataset

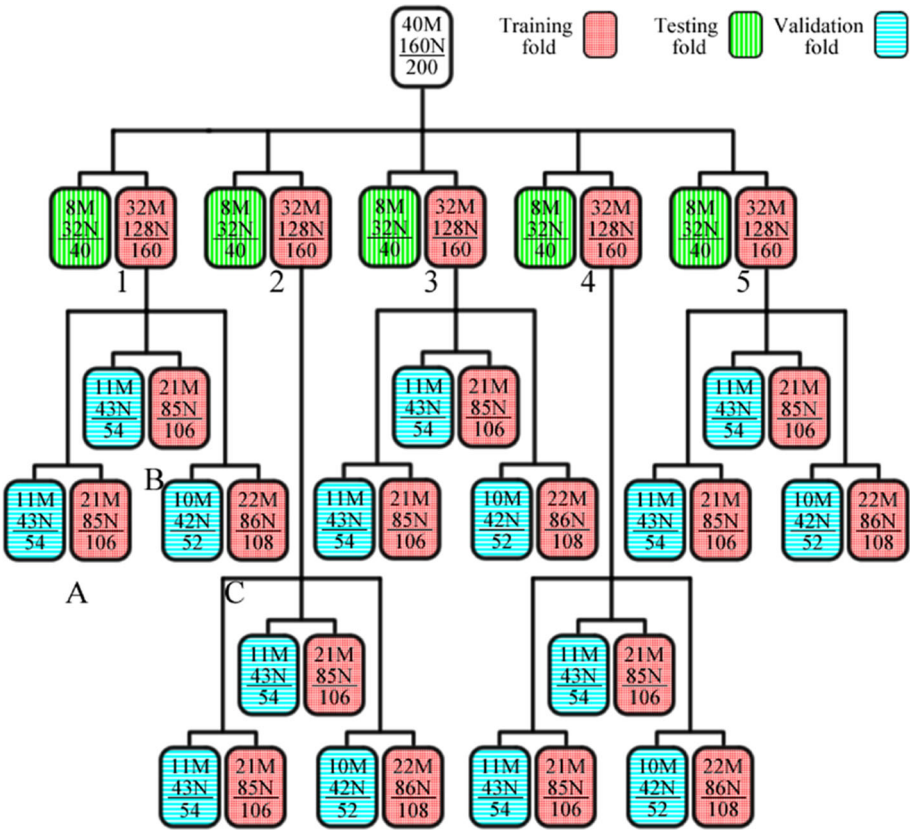

Fig. 3 Number of malignant and non-malignant lesion images in each cross-validation fold from the second dataset – PH^2

CV also allowed us to select the number of epochs to train each DL setting network to be evaluated in the testing folds. To do so, for each DL setting, we conducted a process illustrated by the folds numbered in Figs. 2 and 3. First, we built a network from training Fold A and assessed its accuracy in the corresponding validation fold during 100 epochs. We applied the same procedure for Folds B and C. We then plotted the Acc averaged across the validation Folds A, B and C. Afterward, the number of epochs that led to the highest average Acc was chosen to train a network in Fold 1, which in turn was assessed in the corresponding testing fold. The same process described for Fold 1 was employed in Folds 2, 3, 4 and 5. Section 4 considers the accuracy, specificity, sensitivity and F1 score values found in these five folds.

The batch size used to evaluate a network in each testing fold in the first dataset equals the number of images in this fold, as the greatest common divisor between the different testing fold sizes (20 and 21) is 1. We followed the same reasoning for PH^2, which resulted in a batch size of 40 images in the second dataset — Fig. 3.

We applied the CV process three times for each DL setting, as the inherent training procedure is not deterministic. Then, this work applied the Shapiro-Wilk normality test implemented in GraphPad Prism. Depending on the result from this test and the number of compared columns (algorithms), a specific parametric or non-parametric test was conducted

in the same software. We chose tests for unpaired groups, as a classifier compared with this work dealt with different folds in the first dataset.

The Holdout Validation (HV) strategy was applied three times specifically to assess the proposal in COVID-CT. Thus, we could use the original dataset split — Table 2 — and consider literature results [28] as experimental references for this work. Similarly to each execution of the CV-based procedure, the number of epochs that led to the highest Acc in the validation set, inherent to a holdout process run, was chosen to build the network assessed in the testing set. Section 4 considers the accuracy, specificity, sensitivity and F1 score values found in the testing set. A batch size of 1 was employed in this dataset.

4 Results

Tables 3 and 4 show the performance reached by the eight neural networks settings evaluated in this work, respectively, in the first and second datasets. For each setting, the table presents the average (avg) and the corresponding standard deviation (sd) and coefficient of variation (cv) in terms of accuracy, sensitivity, specificity and F1 score. In particular, the avg, sd and cv values were calculated across the results obtained from three runs in the five testing folds represented in Figs. 2 and 3. The best results in each table column are highlighted in gray. Figures 4 and 5, in turn, highlight the average of each evaluation measure by associating a measure with a spider chart axis in each dataset. Section 5.1 discusses these results and reports findings from the corresponding statistical test of significance.

Table 5 shows the average performance and the corresponding standard deviation and coefficient of variation achieved by the best setting in this work (Table 3) and TSL20 NN, a shallow learning model based on 1 Nearest Neighbor and selected handcrafted features that stood out in the previous paper [38]. This work considered three runs of an algorithm in five pairs of training/testing folds from stratified CV. In turn, the previous one evaluated a single run of a deterministic classifier in 10-folds stratified cross-validation. Both classifiers were compared in the first dataset with 104 dermoscopic images according to the evaluation measures applied in [38]. The best results in each column in Table 5 are highlighted in gray.

Table 6 shows the average performance achieved by the best setting in this work (Table 4) and DL proposals from the literature [2, 8, 27, 45]. All the classifiers were compared in the second dataset with 200 dermoscopic images. We also included FT2 performance in PH^2 in the table, as this setting was the best in the first dataset. It should be emphasized that this work considered three runs of every DL setting in five pairs of training/testing folds from CV, while the previous ones evaluated their models in a single run. Note that [2, 8, 27, 45] did not publish the standard deviations reached by their algorithms. Thus, we did not include this information or the coefficient of variation in Table 6. The best results in each column are highlighted in gray.

Table 7 allows one to compare the best deep neural network setting found in this work in the first dataset with related References (Ref) in terms of average accuracy, sensitivity, specificity

Table 2 Data split in COVID-CT: number of images for each class and set of images

Set	COVID-19	NonCOVID-19	Total
training	191	234	425
validation	60	58	118
testing	98	105	203

Table 3 Average performance and the corresponding standard deviation and coefficient of variation achieved by eight deep neural networks settings in the first dataset

	Acc		
	avg	sd	cv (%)
FT1	0.8951	0.1067	11.9201
FT2	0.9202	0.0930	10.1018
FT3	0.9013	0.0975	10.8211
FT4	0.8911	0.1378	15.4652
FT5	0.9079	0.1010	11.1199
FT6	0.8984	0.1483	16.5038
FT7	0.9111	0.1381	15.1564
FT8	0.8794	0.1504	17.1079
	Sen		
	avg	sd	cv (%)
FT1	0.9378	0.0920	9.8144
FT2	0.9311	0.1059	11.3786
FT3	0.9256	0.0769	8.3042
FT4	0.9389	0.1067	11.3599
FT5	0.9278	0.0825	8.8961
FT6	0.9389	0.1239	13.1933
FT7	0.9611	0.0883	9.1916
FT8	0.8944	0.1558	17.4182
	Spe		
	avg	sd	cv (%)
FT1	0.8370	0.1918	22.9187
FT2	0.9044	0.1319	14.5788
FT3	0.8667	0.1840	21.2328
FT4	0.8341	0.2196	26.3342
FT5	0.8815	0.1900	21.5535
FT6	0.8444	0.2214	26.2216
FT7	0.8444	0.2715	32.1533
FT8	0.8593	0.1805	21.0029
	F1		
	avg	sd	cv (%)
FT1	0.9125	0.0877	9.6071
FT2	0.9338	0.0837	8.9642
FT3	0.9163	0.0800	8.7268
FT4	0.9068	0.1125	12.4069
FT5	0.9224	0.0823	8.9198
FT6	0.9066	0.1401	15.4582
FT7	0.9304	0.1042	11.1961
FT8	0.8925	0.1352	15.1462

and F1 score. The literature results were achieved in distinct datasets with malignant and non-malignant lesion images. The datasets differ, for example, in terms of the number of images and the Majority Class Accuracy (MCA), *i.e.*, the complement of the majority class

Table 4 Average performance and the corresponding standard deviation and coefficient of variation achieved by eight deep neural networks settings in the second dataset – PH^2

	Acc		
	avg	sd	cv (%)
FT1	0.9750	0.0366	3.7535
FT2	0.9650	0.0461	4.7770
FT3	0.9700	0.0465	4.7921
FT4	0.9683	0.0623	6.4335
FT5	0.9533	0.0566	5.9360
FT6	0.9583	0.0572	5.9705
FT7	0.9600	0.0761	7.9233
FT8	0.9533	0.0633	6.6391
	Sen		
	avg	sd	cv (%)
FT1	0.9167	0.1543	16.8331
FT2	0.8667	0.2239	25.8391
FT3	0.8917	0.2052	23.0147
FT4	0.9167	0.1747	19.0538
FT5	0.8250	0.2400	29.0882
FT6	0.9000	0.1581	17.5682
FT7	0.9000	0.2123	23.5936
FT8	0.8833	0.1858	21.0346
	Spe		
	avg	sd	cv (%)
FT1	0.9896	0.0152	1.5409
FT2	0.9896	0.0152	1.5409
FT3	0.9896	0.0152	1.5409
FT4	0.9813	0.0369	3.7638
FT5	0.9854	0.0161	1.6376
FT6	0.9729	0.0352	3.6150
FT7	0.9750	0.0490	5.0241
FT8	0.9708	0.0382	3.9360
	F1		
	avg	sd	cv (%)
FT1	0.9303	0.1064	11.4371
FT2	0.8922	0.1553	17.4097
FT3	0.9103	0.1466	16.1098
FT4	0.9178	0.1618	17.6289
FT5	0.8578	0.1907	22.2355
FT6	0.8947	0.1443	16.1248
FT7	0.8961	0.1973	22.0135
FT8	0.8796	0.1643	18.6745

error – the error in the case of new examples being classified as belonging to the most frequent class. In Table 7, a '?' character indicates that the number of M and N images is unknown, which prevents us from calculating the corresponding MCA. The best results in each evaluation measure column are highlighted in gray.

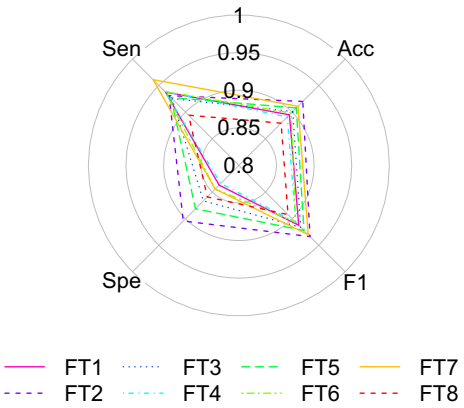

Fig. 4 Average performance achieved by eight deep neural networks settings in the first dataset

Table 8 shows the average performance and the corresponding standard deviation and coefficient of variation achieved in COVID-CT by the best setting in this work (Table 3). The dataset MCA is 0.53.

5 Discussion

First, we discuss the results achieved by the eight classification settings described in this work. Afterward, we compare the best settings from this work with a shallow learning method and four deep learning proposals from related work evaluated in the same datasets we used. This section also compares the best setting from this work in the first dataset with several deep and shallow learning methods from the literature. The application of this setting in images associated with COVID is then addressed.

5.1 Deep neural network settings described in this work

As Table 3 indicates, the eight settings performed similarly. In fact, the Kruskal-Wallis statistical test did not find any significant difference in this scenario. To select the best

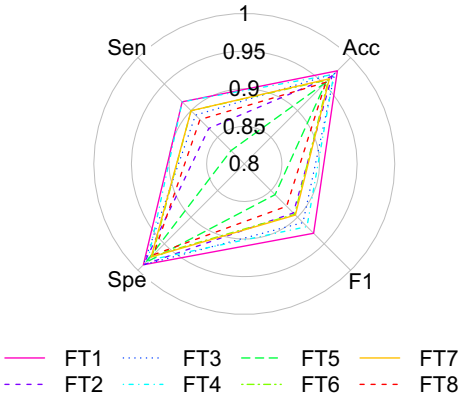

Fig. 5 Average performance achieved by eight deep neural networks settings in the second dataset – PH^2

Table 5 Best results from FT2 (this work) and TSL20 NN (the previous one) in the first dataset

	Acc		
	avg	sd	cv (%)
FT2	0.9202	0.0930	10.1018
Lee et al. [38]	0.8091	0.1036	12.8000
	Sen		
	avg	sd	cv (%)
FT2	0.9311	0.1059	11.3786
Lee et al. [38]	0.8600	0.1554	18.0694
	Spe		
	avg	sd	cv (%)
FT2	0.9044	0.1319	14.5788
Lee et al. [38]	0.7300	0.1814	24.8429
	F1		
	avg	sd	cv (%)
FT2	0.9338	0.0837	8.9642
Lee et al. [38]	–	–	–

network generated in the first dataset, we considered only the average performance from this table and depicted it in Fig. 4. Note that FT2 outperformed the remaining deep neural networks settings in terms of three out of the four evaluation measures. In two measures, FT2 reached the best standard deviation and coefficient of variation. One of them is accuracy, a criterion that takes into account correct classifications regarding both classes equally.

Another advantage of FT2 involves its architecture, simpler than the ones adopted by FT4, FT6, FT7 and FT8 due to the lower number of weight layers (16 *vs* 19). Different from FT1 and FT3, the highlighted setting uses a higher learning rate (LR=2×10^{-5}) to train the custom network, added on top of the base one, than to fine-tune the custom network and the unfrozen blocks (LR=1×10^{-5}). As a result, the first training process can perform larger modifications in its data representations and potentially achieve faster convergence [22]. Finally, FT2 unfreezes one block more than FT5 — Fig. 1 —, releasing more layers to be specialized to the dermoscopic images used in this work.

Table 4 and Fig. 5 show a partially different scenario, as the FT1 setting outperformed FT2 and other alternatives in terms of accuracy, sensitivity and F1 score in PH^2. In fact, it reached the largest average value and the smallest standard deviation and coefficient of variation in

Table 6 Best average performances from FT1 and FT2 (this work) and DL methods (the previous ones) in the second dataset – PH^2

	Acc	Sen	Spe	F1
FT1	0.9750	0.9167	0.9896	0.9303
FT2	0.9650	0.8667	0.9896	0.8922
Bansal et al. [8]	0.9800	0.9750	0.9810	0.9500
Abayomi-Alli et al. [2]	0.9218	0.8077	0.9510	0.8084
Gulati and Bhogal [27]	0.9750	1.0000	0.9687	–
Maia et al. [45]	0.9250	0.7500	0.9688	0.8000

Table 7 Average performance achieved by the best setting in this work (first dataset) and several literature methods. The Learning Approach (LA) — deep learning or Shallow Learning (SL) —, the number of images and the majority class accuracy in the target dataset submitted for classification are also shown for each method

LA	Ref	Acc	Sen	Spe	F1	Number of images	MCA
DL	FT2	0.9202	0.9311	0.9044	0.9338	58M + 46N = 104	0.56
	[8]	0.9800	0.9750	0.9810	0.9500	40M + 160N = 200	0.80
	[39]	0.8500	0.9310	–	0.8090	?M + ?N = 9025	0.74
	[2]	0.9218	0.8077	0.9510	0.8084	40M + 160N = 200	0.80
	[34]	0.9067	0.9070	–	–	?M + ?N = 10015	0.67
	[65]	0.8425	–	–	–	?M + ?N = 10015	0.67
	[11]	–	0.5910	0.9460	–	?M + ?N = 11527	-
	[26]	0.7960	0.8780	0.7140	–	1100M + 12500N = 13600	0.92
	[27]	0.9750	1.0000	0.9687	–	40M + 160N = 200	0.80
	[60]	0.8318	0.9553	0.9622	–	?M + ?N = 900	-
	[61]	0.8300	0.1300	1.0000	–	374M + 1626N = 2000	0.81
	[45]	0.9250	0.7500	0.9688	0.8000	40M + 160N = 200	0.80
	[37]	0.8120	0.9500	0.6800	–	770M + 9400N = 10170	0.92
	[47]	0.7920	0.4760	0.8810	–	388M + 891N = 1279	0.70
	[69]	0.8770	0.6440	0.9180	–	?M + ?N = 2600	-
SL	[4]	0.9993	0.9975	0.9998	–	40M + 160N = 200	0.80
	[75]	0.9970	1.0000	0.9910	0.9973	184M + 113N = 297	0.62
	[58]	0.7760	–	–	–	83M + 167N = 250	0.67
	[59]	0.6851	–	–	–	185M + 377N = 562	0.67
	[3]	0.9650	0.9760	0.9050	–	40M + 160N = 200	0.80
	[10]	0.8430	0.9250	0.7630	–	241M + 241N = 482	0.5
	[33]	0.8790	–	–	–	128M + 128N = 256	0.5
	[57]	–	0.9846	0.7000	–	90M + 5040N = 5130	0.98
	[12]	0.8170	0.9170	0.7450	–	40M + 160N = 200	0.80
	[9]	–	0.9600	0.8000	–	25M + 151N = 176	0.86
	[14]	–	0.9800	0.8600	–	25M + 151N = 176	0.86
	[15]	0.8700	0.8500	0.8700	–	25M + 151N = 176	0.86
	[13]	–	0.9300	0.8500	–	25M + 151N = 176	0.86

these evaluation measures. Together with FT2 and FT3, FT1 yielded the best performance in terms of specificity. However, regardless of the evaluation measure, Kruskal-Wallis did not find any significant difference among the eight settings. In this context, FT1 was considered the best setting in PH^2, as it was competitive in all evaluation measures and used the same higher rate in the 4^{th} and 6^{th} fine-tuning procedures — Section 3.2.

The fact that no statistical difference was found among the eight settings in Tables 3 and 4 favors FT2 and FT1, highlighted respectively in the first and second datasets, as well as FT3 and FT5. In particular, these four settings are the only ones derived from VGG16, which has three layers fewer than VGG19, and the VGG computational complexity depends on the number of layers [29, 77]. Thus, FT1, FT2, FT3 and FT5 can lead to similar performance with a smaller cost than the alternatives.

It should be emphasized that all the settings employed in this work freeze two blocks at least. Thus, generic and reusable features learned from a large dataset (ImageNet) are kept,

Table 8 Average performance and the corresponding standard deviation and coefficient of variation achieved in COVID-CT by the best setting in this work (first dataset)

	Acc		
	avg	sd	cv (%)
FT2	0.7800	0.0413	5.2967
	Sen		
	avg	sd	cv (%)
FT2	0.7381	0.0920	12.4680
	Spe		
	avg	sd	cv (%)
FT2	0.8190	0.0530	6.4741
	F1		
	avg	sd	cv (%)
FT2	0.7624	0.0534	6.9999

supporting the classification of the relatively small sets of medical images considered in this work.

5.2 Direct comparison with a shallow learning and four deep learning methods

The first dataset used in this work was also considered in our previous work [38], easing a direct comparison between the best approaches from each work. As Table 5 indicates, FT2 stood out in terms of average performance, standard deviation and coefficient of variation, regardless of the evaluation measure. In other words, the deep learning setting classified correctly malignant and non-malignant lesions more often, with lower dispersion. The Mann-Whitney statistical test strengthens these findings, as it found that the DL setting significantly outperformed the method based on Nearest Neighbor and selected handcrafted features in terms of accuracy and specificity.

Nevertheless, the shallow learning method is still considered competitive, as no statistical difference was found regarding sensitivity or true positive rate, a relevant measure in the medical domain. Also, TSL20 NN complexity is lower than FT2 one, which needs to be taken into account in scenarios with limited computational resources.

These methods can also be compared in terms of the input submitted for learning. FT2, like many deep learning methods used in computer vision, receives images as input. They learn representations directly from these images to label them [22]. On the other hand, TSL20 NN requires handcrafted features describing images as input. In particular, it uses features based on texture, shape and local binary patterns [38]. The feature selection algorithm ReliefF is then employed to find a subset of 33 relevant features that are submitted to Nearest Neighbor to classify the inherent images. In this scenario, deep neural networks are superior by automating feature engineering and, consequently, easing the building of a competitive learner.

Fine-tuning a pre-trained network built from a large dataset (ImageNet) also supported FT2 to outperform TSL20 NN. While FT2 employs generic features learned from thousands of images and specific features acquired from a small target dataset, TSL20 NN takes into account 33 handcrafted features and applies them in the target dataset. Despite of the differences between ImageNet and the dermoscopic images used in this study, the convolutional network inherent to FT2 can combine local, portable patterns learned by the lower layers

into particular patterns in the higher layers [22]. Consequently, one could achieve a proper dermoscopic image description.

The second dataset investigated in this work was also studied in related work on deep learning, as illustrated in [2, 8, 27, 45], simplifying a direct comparison. Table 6 includes the performance of our best setting in PH^2 (FT1), the four references and FT2, which was considered our best setting in the first dataset.

The approach described in [8] reached the best average accuracy in PH^2, followed by FT1 and [27] — Table 6. The maximum average F1 score was also reported in [8]. FT1, together with FT2, achieved the best average specificity. The DL method evaluated in [27], in turn, was highlighted in terms of sensitivity.

The approach introduced in [8] achieved the highest accuracy and F1 score using the EfficientNet-B0 architecture. As the name suggests, this scheme has a reduced number of parameters (5.3 million) compared with other alternatives. Despite the remarkable results, the authors employed a proper image pre-processing method not conducted by us to benefit the input of their network: hair removal. Moreover, they included effective handcrafted features based on shape, color and texture, supplementing characteristics acquired automatically from the images by their deep network.

A noticeable performance was obtained in [27] by using the same architecture inherent to FT1 and FT2 (VGG16 [64]). However, the authors evaluated their network with a single run of the holdout validation strategy, which divides the image dataset into training and testing folds only once before assessing the learner. As in the previous work, the authors also performed hair removal.

The application of pre-trained VGG16, VGG19 and other architectures for feature extraction, yielding the input for shallow classification algorithms with parameter values empirically optimized, led to relatively high accuracy and specificity in [45]. The authors ran cross-validation only once, while we did it three times due to the non-deterministic training procedure of the mentioned convnets. CV splits the dataset a few times before classifier assessment, potentially mitigating biases derived from the choice of a particular sample by conventional HV [72] used in [8, 27]. It should also be emphasized that Regions of Interest (ROI) extracted from the Dermoscopy examinations by experts fed the classifiers evaluated in [45]. ROI centering and the inclusion of vertical and horizontal lines around each lesion to standardize its size are other pre-processing techniques not used in this work.

Table 6 also allows one to note that FT2, which was considered the best setting in the first dataset evaluated in this study, is also competitive in the second one (PH^2), especially in terms of specificity.

As mentioned, the related methods [8, 27, 45] performed a single validation run. Thus, an additional comparison between our settings and them considers the results of our best CV run. In that evaluation, FT1 led to the following average values: 0.975 (Acc), 0.925 (Sen), 0.9875 (Spe) and 0.9311 (F1). In turn, FT2 reached the following average values: 0.975 (Acc), 0.9000 (Sen), 0.9938 (Spe) and 0.9216 (F1). Based on these results and Table 6, one can note that FT1 would be more competitive in sensitivity and F1 score, while FT2 would do so in all measures.

5.3 Indirect comparison with literature methods

Table 7 shows the results of FT2 in the first dataset chosen in this work and results published by several references. Any discussion on the table results should take into account that the image datasets, the learning approach, the classification algorithms and/or the evaluation

strategy chosen vary across the publications. Also, in many cases, the standard deviation and even the average performance in terms of specific measures were not reported by the related literature. Thus, a more in-depth comparison between the results published in the references and this work was not possible. In any case, the indirect comparison presented in this section allows one to feel how promising our results are.

In this scenario, one can note that FT2 achieved one of the best performances in terms of accuracy, specificity and F1 score among the deep learning methods. In turn, papers [2, 8, 27] and [45] were already compared with this setting in Section 5.2. It should be emphasized that, besides the validation strategy — Section 5.2 —, this work differentiates from them by conducting fewer image pre-processing procedures. We believe that FT2 could be even more competitive in Table 7 if these procedures preceded learning.

It should also be emphasized that a few shallow learning algorithms were competitive, such as genetic programming [4], an ad hoc classification algorithm [75] and support vector machines [3]. Nevertheless, the proposal described in [4] needs to extract handcrafted features to construct new ones before learning, while this work can automatically and directly acquire features from the input images. The learner built in [75] was designed to identify melanoma based on specific image patterns in acral areas, while the settings evaluated in this work do not depend on these patterns. In turn, single run HV supported learner assessment in [3]. The authors reported Acc, Sen and Spe from their best setting, which conducted a two-step classification. First, the setting classified images as benign or abnormal. Then, it concentrated on the abnormal images, differentiating melanoma from atypical lesions. An issue with their approach is that label errors from the first step propagate to the second one.

FT2 performance is less promising according to sensitivity, as could also be inferred from Section 5.2. The U-net convolutional neural networks algorithm was used in [60] to segment melanoma and non-melanoma lesions and, consequently, support VGG16 to achieve higher performance. This finding indicates that employing other segmentation approaches could be helpful for VGG16 based FT2. VGG19, another architecture evaluated in this study — Section 3.2 —, led to remarkable sensitivity in [37]. However, the average accuracy was lower than the MCA in [37], suggesting that the network would not outperform a baseline method that labels every image as the majority class (non-malignant). A similar issue is noted in [26].

A point in common among most of the papers highlighted in this Section [2, 4, 8, 27, 45, 60, 75] is that they took into account a learning scheme to differentiate melanoma from non-melanoma lesion images. On the other hand, this work considered 53 melanoma and 5 carcinoma lesions as the 58 malignant images in the first dataset. Although keeping both SC types in the positive class represents an additional difficulty our classifiers had to deal with, it approaches better a scenario in which a CAD system classifies suspect lesions as malignant, regardless of the cancer type, or non-malignant.

Recall that we applied the CV strategy three times for each DL setting. A deeper analysis into the FT2 prediction errors in the first dataset showed that this setting mislabeled, respectively, 4, 2 and 6 malignant images in testing folds in the first, second and third CV executions. Although only one carcinoma was wrongly classified in the third run, this lesion type is less frequent in our dataset. Running FT2 more times in future work is an alternative to find if it recognizes carcinoma lesions better than melanoma ones. Another idea to study is using explanation methods, such as Gradient-weighted Class Activation Mapping [49, 74], to search for meaningful visual patterns in relevant skin cancer areas of incorrectly classified malignant instances.

This work dealt with the smallest target image dataset in Table 7. A low number of training images may cause deep and shallow learning algorithms to overfit, especially if the dataset is

not representative enough [48, 72]. As a result, a classifier might not generalize well to new images. To tackle this problem, we found that fine-tuning a pre-trained network built from ImageNet was the best alternative. In particular, other strategies recommended for convnets dealing with little training data [22] — feature extraction with the same network and data augmentation — did not perform well in our preliminary experimental evaluations in the first dataset. It should be emphasized that proposing and evaluating different fine-tuning settings to classify a relatively small group of target images represents a contribution from this work. As our flexible proposal succeeded well in Dermatology, it can also be assessed in other medical domains, such as Colonoscopy, when a low number of training images is available. This reasoning was briefly verified by us in a dataset of another field (COVID-CT), as addressed in Section 5.4.

The number of images (dataset instances) is only one of the properties that can influence classification complexity [41]. In fact, if one compares the eight settings proposed by us in the two datasets — Tables 3 and 4 —, one will find that all settings reached better accuracy and specificity in the second dataset (more images), but better sensitivity in the first dataset (fewer images). Thus, characterizing image datasets' complexity and associating the findings with classification performance is an additional direction to investigate.

This work focused on the family of architectures VGG, as it is relatively easy to understand, has been often employed in applications and has inspired extensions [22, 26]. Within this family, we selected VGG16 and VGG19 due to their public implementation (e.g., in the Keras/TensorFlow framework) and their use in related work. The latter architecture has three convolutional layers more than the former one. It makes sense to investigate a distinct number of layers in VGG, as the classification error can decrease in some cases if more weight layers are included [64].

As mentioned, besides the base architecture, the eight fine-tuning settings vary on the block index B from which freezing is applied and the learning rate used in the two training procedures — Table 1 and Fig. 1. Although it was not found a significant difference among the eight settings in both dermoscopic datasets, evaluating different values for these parameters is considered relevant, as discussed in what follows.

Section 3.2 defined two values for B: three and four. The idea behind these values was to freeze and unfreeze at least two different blocks from the convolutional base — Fig. 1. Thus, a reasonable number of weights from the lowest layers, learned from a large non-medical dataset, can be reused. At the same time, a few layers are available to learn specialized features from small dermoscopic datasets. Using a B value closer to one can promote overfitting [44], especially with little data to train the custom network at the top and the unfrozen blocks. It should be emphasized that an image classification study with three mammography datasets fine-tuned VGG16 and other architectures with the B values three, four and five [23]. In all the datasets, $B = 3$ and $B = 4$ reached the highest accuracy values for VGG16.

The learning rate is one of the network hyperparameters that one could tune to find an optimal configuration [22]. This work evaluated two variations of this hyperparameter in a scenario with eight fine-tuning settings — Table 1: (1) a constant LR=1×10^{-5} for the 4^{th} and 6^{th} training procedures reported in Section 3.2 and (2) a higher LR=2×10^{-5} in the 4^{th} procedure and a smaller LR=1×10^{-5} in the 6^{th} one. Using a relatively small LR in the 6^{th} procedure makes sense, as we wanted to restrict the magnitude of the changes in the representations associated with the unfrozen blocks, *i.e.*, the layers under fine-tuning. On the other hand, applying a too large LR value in this step could hinder these representations and the reuse of a previously learned model to deal with little training data.

5.4 Proposal application in a second medical domain

Table 8 shows the FT2 results, averaged across three HV runs in the COVID-CT dataset. According to four evaluation measures, the setting performance can be considered satisfactory for two reasons, at least. First, FT2 average accuracy was higher than MCA (0.53), even with some CT slices with artifacts [63]. In particular, annotations, marks, letters or noise can be noted in these images, many of which are associated with COVID-19 clinical findings (positive cases). This issue makes sense, as the images were directly collected from papers. Second, holdout has a few limitations already discussed in Section 5.2 and does not take into account that each patient in the COVID-CT case can have more than one slice. Strategies such as CV and Leave one patient out [53] can be more appropriate.

Using holdout to evaluate our proposal in this additional study allowed us to consider literature results [28] as experimental references in the same dataset. Note that two deep networks methods in [28] reached higher Acc than the VGG-based one: DenseNet169 (0.85) and Self-Trans (0.86). This achievement can be associated with structural differences between these two methods and ours, such as using several layers densely connected to others in the former [31] or the self-supervised TL approach in the latter [66]. Moreover, different from the approaches in [28], FT2 did not have COVID-CT enlarged by data augmentation, which indicates that this method works well with only a few hundred medical images for training. This achievement was also noted in Dermoscopy examinations in the previous sections. In any case, if one analyzes only the best FT2 run in the CT images, an accuracy closer to the literature ones was achieved: 0.8276.

The findings obtained in a dataset larger than the two dermoscopic ones suggest that the best setting of this work, identified in a group of 104 skin images, was flexible and performed adequately in another medical domain with 746 computerized tomographic slices.

6 Conclusion

This work successfully proposed and evaluated eight fine-tuning settings in two small datasets with 104 and 200 dermoscopic images. Two settings led to the best results: FT1 and FT2. Both unfreeze from layer Block $B = 3$ to Block 5 in their VGG16 architecture during the fine-tuning process. However, only FT1 uses a constant learning rate during training.

Using the FT2 setting in the smallest dataset was competitive compared to recent deep and shallow learning papers. Moreover, FT1 and FT2 led to performance comparable with four deep learning methods in the other dataset. This work also illustrated that DL can be effective — learn properly and generalize — if one has only a few dozen malignant and non-malignant lesion images to study and classify in Dermatology. An additional study with CT scans related to COVID-19 suggested that FT2 is flexible enough to perform satisfactorily in another domain with more training images.

This study supports research on health informatics by addressing and assessing an intelligent method that can learn from small labeled datasets from distinct domains. These datasets illustrate many real-world scenarios in which it is hard to find a reasonable amount of labeled images for infrequent diseases or conditions [73], often leading to imbalanced class distribution. The best deep learning settings investigated in this work performed competitively, offering an alternative to classify medical image examinations with a relatively low number of abnormal and normal samples.

Future work includes developing and evaluating transfer learning strategies in more medical domains in which small datasets can be found, such as Colonoscopy. Another research direction to investigate is to assess the eight proposed settings in larger public datasets with dermoscopic images, such as the International Skin Imaging Collaboration: Melanoma Project [3] one used in [61]. Adapting some FT variants for other neural network architectures is also relevant to study. Finally, combining oversampling [2, 5, 68] or alternative schemes with fine-tuning could be explored to extend our DL strategies and perform better in PH^2 and other small databases in which malignant lesions represent the minority class.

Acknowledgements The authors thank Gabriel Humpire for his help in the background acquisition.

Author Contributions N. Spolaôr: Investigation, Writing - Original Draft H. D. Lee: Methodology, Visualization, Writing - Review & Editing A. I. Mendes: Methodology, Supervision, Writing - Original Draft C. V. Nogueira: Investigation, Writing - Review & Editing A. R. S. Parmezan: Visualization, Software W. S. R. Takaki: Validation, Software C. S. R. Coy: Conceptualization, Supervision F. C. Wu: Conceptualization, Methodology R. Fonseca-Pinto: Conceptualization, Writing - Original Draft.

Funding We would like to acknowledge eurekaSD: Enhancing University Research and Education in Areas Useful for Sustainable Development - grants EK14AC0037 and EK15AC0264. We thank Araucária Foundation for the Support of the Scientific and Technological Development of Paraná through a Research and Technological Productivity Scholarship for H. D. Lee (grant 028/2019). We also thank the Brazilian National Council for Scientific and Technological Development (CNPq) through the grant number 142050/2019-9 for A. R. S. Parmezan. The Portuguese team was partially supported by Fundação para a Ciência e a Tecnologia (FCT). R. Fonseca-Pinto was financed by the projects UIDB/50008/2020, UIDP/50008/2020, UIDB/05704/2020 and UIDP/05704/2020 and C. V. Nogueira was financed by the projects UIDB/00013/2020 and UIDP/00013/2020. The funding agencies did not have any further involvement in this paper.

Data Availability The three datasets are public and were already used in previous work from different authors, as described in Section 3. The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request. Currently, the code is unavailable, although some libraries employed in this work, such as Tensorflow, are free to use.

Declarations

Ethics approval Not applicable.

Consent Not applicable.

Conflict of interest The authors declare that they do not have any other conflict of interest.

References

1. Abadi M, Agarwal A, Barham P et al (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/,software available from tensorflow.org
2. Abayomi-Alli OO, Damasevicius R, Misra S et al (2021) Malignant skin melanoma detection using image augmentation by oversamplingin nonlinear lower-dimensional embedding manifold. Turk J Electr Eng Comput Sci 29(8).https://doi.org/10.3906/elk-2101-133
3. Abuzaghleh O, Faezipour M, Barkana BD (2015) A comparison of feature sets for an automated skin lesion analysis system for melanoma early detection and prevention. In: Long Island Systems, Applications and Technology, pp 1–6. https://doi.org/10.1109/LISAT.2015.7160183
4. Ain QU, Al-Sahaf H, Xue B et al (2022) Genetic programming for automatic skin cancer image classification. Expert Syst Appl 197(116):680. https://doi.org/10.1016/j.eswa.2022.116680

[3] https://www.isic-archive.com/

5. Alazzam MB, Alassery F, Almulihi A (2021) Diagnosis of melanoma using deep learning. Mathematical Problems in Engineering 2021
6. Ashraf R, Afzal S, Rehman AU et al (2020) Region-of-interest based transfer learning assisted framework for skin cancer detection. IEEE Access 8:147858–147871
7. Asiri N, Hussain M, Adel FA et al (2019) Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. Artif Intell Med 99(101):701. https://doi.org/10.1016/j.artmed.2019.07.009
8. Bansal P, Garg R, Soni P (2022) Detection of melanoma in dermoscopic images by integrating features extracted using handcrafted and deep learning models. Comput Ind Eng 168(108):060. https://doi.org/10.1016/j.cie.2022.108060
9. Barata C, Ruela M, Francisco M et al (2014) Two systems for the detection of melanomas in dermoscopy images using texture and color features. IEEE Syst J 8(3):965–979. https://doi.org/10.1109/JSYST.2013.2271540
10. Barata C, Celebi ME, Marques JS (2015) Improving dermoscopy image classification using color constancy. IEEE J of Biomed and Health Inform 19(3):1146–1152. https://doi.org/10.1109/JBHI.2014.2336473
11. Barata C, Celebi ME, Marques JS (2021) Explainable skin lesion diagnosis using taxonomies. Pattern Recognit 110(107):413. https://doi.org/10.1016/j.patcog.2020.107413
12. Barata C, Marques JS, Celebi ME (2014a) Improving dermoscopy image analysis using color constancy. In: IEEE International conference on image processing, pp 3527–3531. https://doi.org/10.1109/ICIP.2014.7025716
13. Barata C, Marques JS, Mendonça T (2013c) Bag-of-features classification model for the diagnose of melanoma in dermoscopy images using color and texture descriptors. In: International Conference Image Analysis and Recognition, pp 547–555
14. Barata C, Marques JS, Rozeira J (2013a) The role of keypoint sampling on the classification of melanomas in dermoscopy images using bag-of-features. In: Sanches JM, Micó L, Cardoso JS (eds) Pattern Recognition and Image Analysis: 6th Iberian Conference, Funchal, Madeira, Portugal, June 5–7, 2013. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 715–723. https://doi.org/10.1007/978-3-642-38628-2_85
15. Barata C, Marques JS, Rozeira J (2013b) Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model. In: Bebis G, Boyle R, Parvin B et al (eds) Advances in Visual Computing: 9th International Symposium, ISVC 2013, Rethymnon, Crete, Greece, July 29–31, 2013. Proceedings, Part I. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 40–49. https://doi.org/10.1007/978-3-642-41914-0_5
16. Boer A, Nischal K (2007) A growing online resource for learning dermatology and dermatopathology. Indian J Dermatol Venereol Leprol 73(2):138–140. https://doi.org/10.4103/0378-6323.31909
17. Carvalho VAM, Spolaôr N, Cherman EA et al (2014) A framework for multi-label exploratory data analysis: Ml-eda. In: Latin american computing conference, pp 1–12
18. Celebi ME, Codella N, Halpern A (2019) Dermoscopy image analysis: Overview and future directions. IEEE Journal of Biomedical and Health Informatics 23(2):474–478
19. Chen Z, Fu Y, Chen K et al (2019) Image block augmentation for one-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 3379–3386
20. Cheng G, Lang C, Han J (2023) Holistic prototype activation for few-shot segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(4):4650–4666. https://doi.org/10.1109/TPAMI.2022.3193587
21. Chen H, Wang Y, Shi Y et al (2018) Deep transfer learning for person re-identification. In: IEEE fourth international conference on multimedia big data, pp 1–5. https://doi.org/10.1109/BigMM.2018.8499067
22. Chollet F, Allaire JJ (2018) Deep learning in R, 1st edn. Manning publications
23. Chougrad H, Zouaki H, Alheyane O (2018) Deep convolutional neural networks for breast cancer screening. Comput Methods Prog Biomed 157:19–30. https://doi.org/10.1016/j.cmpb.2018.01.011
24. Feng X, Jiang Y, Yang X et al (2019) Computer vision algorithms and hardware implementations: A survey. Integr 69:309–320. https://doi.org/10.1016/j.vlsi.2019.07.005
25. Fujita H (2020) Ai-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiol Phys Technol 13:6–19. https://doi.org/10.1007/s12194-019-00552-4
26. Grochowski M, Kwasigroch A, Mikolajczyk A (2019) Selected technical issues of deep neural networks for image classification purposes. Bull of the Pol Acad Sci Technical Sci 67(2):363–376. https://doi.org/10.24425/bpas.2019.128485
27. Gulati S, Bhogal RK (2019) Detection of malignant melanoma using deep learning. In: Singh M, Gupta P, Tyagi V, et al (eds) Advances in Computing and Data Sciences, pp 312–325

28. He X, Yang X, Zhang S et al (2020) Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. https://doi.org/10.1101/2020.04.13.20063941
29. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
30. Hinton G (2020) Neural networks for machine learning.http://tiny.cc/zjlruz
31. Huang G, Liu Z, Van Der Maaten L et al (2017) Densely connected convolutional networks. In: IEEE Conference on computer vision and pattern recognition, pp 2261–2269. https://doi.org/10.1109/CVPR.2017.243
32. JS M, P M, Aravindan C et al (2023) Classification of skin cancer from dermoscopic images using deep neural network architectures. Multimedia Tools and Applications 82(10):15763–15778. https://doi.org/10.1007/s11042-022-13847-3
33. Kaur R, Albano PP, Cole JG et al (2015) Real-time supervised detection of pink areas in dermoscopic images of melanoma: importance of color shades, texture and location. Skin Res Technol 21(4):466–473. https://doi.org/10.1111/srt.12216
34. Khan MA, Sharif M, Akram T et al (2021) Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. Diagnostics 11(5). https://doi.org/10.3390/diagnostics11050811
35. Kokabi M, Donnelly M, Xu G (2020) Benchmarking small-dataset structure-activity-relationship models for prediction of wnt signaling inhibition. IEEE Access 8:228831–228840. https://doi.org/10.1109/ACCESS.2020.3046190
36. Kumar V, Abbas A, Aster J (2014) Robbins & Cotran Pathologic Basis of Disease, 9th edn. Elsevier
37. Kwasigroch A, Mikolajczyk A, Grochowski M (2017) Deep neural networks approach to skin lesions classification – a comparative analysis. In: International conference on methods and models in automation and robotics, pp 1069–1074. https://doi.org/10.1109/MMAR.2017.8046978
38. Lee HD, Mendes AI, Spolaôr N et al (2018) Dermoscopic assisted diagnosis in melanoma: Reviewing results, optimizing methodologies and quantifying empirical guidelines. Knowl-Based Syst 158:9–24. https://doi.org/10.1016/j.knosys.2018.05.016
39. Liu XJ, Kl Li, Hy Luan et al (2022) Few-shot learning for skin lesion image classification. Multimedia Tools and Applications 81(4):4979–4990
40. Li W, Xu J, Huo J et al (2019) Distribution consistency based covariance metric networks for few-shot learning. In: Proceedings of the AAAI conference on artificial intelligence, pp 8642–8649
41. Lorena AC, Garcia LPF, Lehmann J et al (2019) How complex is your classification problem? a survey on measuring classification complexity. ACM Comput Surv 52(5):1–34. https://doi.org/10.1145/3347711
42. Machado M, Pereira J, Fonseca-Pinto R (2015) Classification of reticular pattern and streaks in dermoscopic images based on texture analysis. J Med Imaging 2(4):044503–044503. https://doi.org/10.1117/1.JMI.2.4.044503
43. Mahajan K, Sharma M, Vig L (2020) Meta-dermdiagnosis: Few-shot skin disease identification using meta-learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 730–731
44. Mahdianpari M, Salehi B, Rezaee M et al (2018) Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. Remote Sens 10(7):1119. https://doi.org/10.3390/rs10071119
45. Maia LB, Lima A, Pinheiro Pereira RM, et al (2018) Evaluation of melanoma diagnosis using deep features. In: International conference on systems, signals and image processing, pp 1–4. https://doi.org/10.1109/IWSSIP.2018.8439373
46. Mendonça TF, Ferreira PM, Marçal ARS et al (2016) PH2: A public database for the analysis of dermoscopic images. In: Celebi ME, Mendonça TF, Marques JS (eds) Dermoscopy Image Analysis. CRC Press, Boca Ratón, pp 419–440. https://doi.org/10.1201/b19107
47. Menegola A, Fornaciali M, Pires R et al (2017) Knowledge transfer for melanoma screening with deep learning. In: International Symposium on Biomedical Imaging, pp 297–300. https://doi.org/10.1109/ISBI.2017.7950523
48. Mitchell TM (1997) Machine learning. McGraw-Hill
49. Nunnari F, Kadir MA, Sonntag D (2021) On the overlap between grad-cam saliencymaps and explainable visual features in skin cancer images. In: Holzinger A, Kieseberg P, Tjoa AM et al (eds) Machine learning and knowledge extraction. Springer International Publishing, Cham, pp 241–253
50. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191
51. Pereira J, Fonseca-Pinto R (2015) Segmentation strategies in dermoscopy to follow-up melanoma: combined segmentation scheme. Online J Sci Technol 5(3):56–61

52. Pereira J, Mendes A, Nogueira C et al (2013) An adaptive approach for skin lesion segmentation in dermoscopy images using a multiscale local normalization. In: Bourguignon JP, Jeltsch R, Pinto AA et al (eds) Dynamics, Games and Science: International Conference and Advanced School Planet Earth. Springer International Publishing Switzerland, pp 537–545. https://doi.org/10.1007/978-3-319-16118-1_29
53. Petersen D, Naveed P, Ragheb A et al (2017) Raman fiber-optical method for colon cancer detection: Cross-validation and outlier identification approach. Spectrochim Acta A Mol Biomol Spectrosc 181:270–275. https://doi.org/10.1016/j.saa.2017.03.054
54. Porta CAML (2011) Skin Cancers - Risk Factors. Prevention and Therapy, InTech, Rijeka
55. Prabhu V, Kannan A, Ravuri M et al (2019) Few-shot learning for dermatological disease diagnosis. In: Machine Learning for Healthcare Conference, PMLR, pp 532–552
56. Rafay A, Hussain W (2023) EfficientSkinDis: An EfficientNet-based classification model for a large manually curated dataset of 31 skin diseases. Biomedical Signal Processing and Control 85(104):869. https://doi.org/10.1016/j.bspc.2023.104869
57. Rastgoo M, Garcia R, Morel O et al (2015) Automatic differentiation of melanoma from dysplastic nevi. Comput Méd Imaging and Graph 43:44–52. https://doi.org/10.1016/j.compmedimag.2015.02.011
58. Sáez A, Sánchez-Monedero J, Gutiérrez PA et al (2016) Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images. IEEE Trans Méd Imaging 35(4):1036–1045. https://doi.org/10.1109/TMI.2015.2506270
59. Sánchez-Monedero J, Sáez A, Pérez-Ortiz M et al (2016) Classification of melanoma presence and thickness based on computational image analysis. In: Martínez-Álvarez F, Troncoso A, Quintián H et al (eds) Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016, Proceedings. Springer International Publishing, Cham, pp 427–438. https://doi.org/10.1007/978-3-319-32034-2_36
60. Seeja RD, Suresh A (2019) Melanoma segmentation and classification using deep learning. Int J Innov Technol Explor Eng 8(12):2667–2672. https://doi.org/10.35940/ijitee.L2516.1081219
61. Serte S, Demirel H (2019) Gabor wavelet-based deep learning for skin lesion classification. Comput Biol Medicine 113(103):423. https://doi.org/10.1016/j.compbiomed.2019.103423
62. Shuai W, Li J (2022) Few-shot learning with collateral location coding and single-key global spatial attention for medical image classification. Electronics 11(9):1510
63. Silva P, Luz E, Silva G et al (2020) Covid-19 detection in ct images with deep learning: A voting-based scheme and cross-datasets analysis. Informatics in Medicine Unlocked 20(100):427. https://doi.org/10.1016/j.imu.2020.100427
64. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, presented in the International Conference on Learning Representations
65. Singh R, Bharti V, Purohit V et al (2021) Metamed: Few-shot medical image classification using gradient-based meta-learning. Pattern Recogn 120(108):111
66. Sudowe P, Leibe B (2016) Patchit: Self-supervised network weight initialization for fine-grained recognition. In: British machine vision conference, pp 75.1–75.12. https://doi.org/10.5244/C.30.75
67. Sugata TLI, Yang CK (2017) Leaf app: Leaf recognition with deep convolutional neural networks. IOP Conf Ser: Mater Sci and Eng 273:012,004. https://doi.org/10.1088/1757-899x/245/1/012004, Licensed under CC BY 3.0. Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI
68. Tahir M, Naeem A, Malik H et al (2023) DSCC_Net: Multi-classification deep learning models for diagnosing of skin cancer using dermoscopic images. Cancers 15(7). https://doi.org/10.3390/cancers15072179
69. Thao LT, Quang NH (2017) Automatic skin lesion analysis towards melanoma detection. In: Asia Pacific Symposium on Intelligent and Evolutionary Systems, pp 106–111. https://doi.org/10.1109/IESYS.2017.8233570
70. Wang W, Yang Y, Wang X et al (2019) Development of convolutional neural network and its application in image classification: a survey. Opt Eng 58(4):1–19. https://doi.org/10.1117/1.OE.58.4.040901
71. Wang Z, Duan LY, Lin J et al (2015) Hamming compatible quantization for hashing. In: International Joint Conference on Artificial Intelligence, pp 2298–2304
72. Witten IH, Frank E, Hall MA et al (2016) Data Mining: Practical Machine Learning Tools and Techniques, 4th edn. Morgan Kaufmann, Burlington
73. Wu P (2022) A survey of few-shot learning research based on deep neural network. Frontiers Comput Intell Syst 2(1):110–115. https://doi.org/10.54097/fcis.v2i1.3177
74. Xin C, Liu Z, Zhao K et al (2022) An improved transformer network for skin cancer classification. Comput Biol Med 149(105):939. https://doi.org/10.1016/j.compbiomed.2022.105939

75. Yang S, Oh B, Hahm S et al (2017) Ridge and furrow pattern classification for acral lentiginous melanoma using dermoscopic images. Biomed Signal Process and Control 32:90–96. https://doi.org/10.1016/j.bspc.2016.09.019
76. Yu L, Chen H, Dou Q et al (2016) Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 36(4):994–1004
77. Zhang X, Zou J, He K et al (2015) Accelerating very deep convolutional networks for classification and detection. IEEE Transactions On Pattern Analysis And Machine Intelligence 38(10):1943–1955
78. Zhao J, Zhang Y, He X et al (2020) COVID-CT-Dataset: a CT scan dataset about COVID-19. arXiv:2003.13865

Authors and Affiliations

Newton Spolaôr[1] · Huei Diana Lee[1] · Ana Isabel Mendes[2] · Conceição Veloso Nogueira[2,3] · Antonio Rafael Sabino Parmezan[1,4] · Weber Shoity Resende Takaki[1] · Claudio Saddy Rodrigues Coy[5] · Feng Chung Wu[1,5] · Rui Fonseca-Pinto[2,6,7]

Huei Diana Lee
huei.lee@unioeste.br

1 Laboratory of Bioinformatics, Western Paraná State University (UNIOESTE), Presidente Tancredo Neves Avenue 6731, Foz do Iguaçu 85867-900, Paraná, Brazil

2 Polytechnic Institute of Leiria, General Norton de Matos Street, 4133, Leiria 2411-901, Portugal

3 Center of Mathematics, University of Minho, Braga, Portugal

4 Laboratory of Computational Intelligence, Institute of Mathematics and Computer Science, University of São Paulo (USP), São Carlos, São Paulo, Brazil

5 Service of Coloproctology, Faculty of Medical Sciences, University of Campinas, Campinas, São Paulo, Brazil

6 CiTechCare - Center for Innovative Care and Health Technology, Polytechnic Institute of Leiria, Leiria, Portugal

7 IT - Instituto de Telecomunicações - Leiria, Leiria, Portugal