Robust feature spaces from pre-trained deep network layers for skin lesion classification

Fernando Pereira dos Santos, Moacir A. Ponti
Institute of Mathematical and Computer Sciences (ICMC)
University of São Paulo (USP), São Carlos, SP, Brazil
Email: {fernando_persan,ponti}@usp.br

***Abstract*—The incidence of skin cancer in the world population is a public health concern, and the first diagnosis takes into account the appearance of lesions on skin. In this context, automated methods to aid the screening for malign lesions can be an important tool. However, the efficiency of developed methods depends directly on the quality of the generated feature space which may vary when considering different image datasets and sources. We present a detailed study of feature spaces obtained from deep convolutional networks (CNNs), using the benchmark PH2 dataset, considering three CNN architectures, as well as investigating different layers, impact of dimensionality reduction, use of colour quantisation and noise addition. Our results show that, features have discriminative capability comparable to competing methods with balanced accuracy 94%, and 95% with noise injection. Additionally, we present a study of fine-tuning and generalisation across image quantisation and noise levels, contributing to the discussion of learning features from deep networks and offering a guideline for future works.**

I. Introduction

Skin cancer is an abnormal and uncontrolled growth in cells with potential to be invasive, spreading to other tissues or organs. Early diagnosis is crucial to favour probabilities of cure: if the disease is detected before malignant cells grow or spread to other parts of the body, the treatment is often more efficient [1]. Because the appearance of skin is used as screening for an initial diagnosis in such scenarios, digital image processing methods have been continuously developed for skin cancer diagnosis, via analysis of images and classification of skin lesions [2], [3]. However, there are some important challenges in addressing this problem because images are acquired under different conditions of illumination, may have blur due to the focal field, and also present strong texture due to the appearance of skin and other confusing components [4].

The standard pipeline in this application comprises: pre-processing, segmentation, feature extraction, and classification [5]. In order to perform feature extraction, ABCD Rule is often adopted [4], [6], [7] to characterise asymmetry, edge regularity, colour homogeneity, and texture uniformity. Some studies consider only part of ABCD Rule with handcrafted methods to represent feature spaces [2], [3]. However, by treating each feature separately, in this approach, it is necessary to apply different algorithms aiming to form an ideal vector of features [8]. Accordingly, the efficiency of one classifier depends directly on the appropriate descriptors choice.

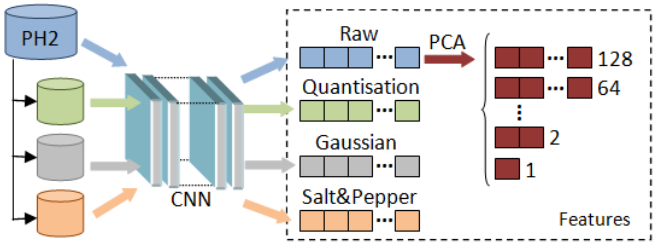

Fig. 1. Feature Extraction. From raw images (PH2), new sets were generated by colour quantisation and noise injection (Gaussian and impulsive). Feature maps, extracted using a pre-trained CNN, were used to evaluate skin lesion classification and generalisation capacity between raw and different levels of distortions. PCA was applied to raw set to measure space reduction efficiency.

Recent studies have shown the efficiency of deep convolutional neural networks (CNNs) as extractors of low-level (shapes and edges) and high-level features (textures and semantics) due to the high abstraction capacity codified in many layers [9]. In some scenarios, layers are able to filter relevant features so it is even possible to neglect filtering or segmentation steps [10].

A basic requirement to guarantee that a CNN can learn concepts of one domain is the amount of labeled data [11], which can be very expensive [12], [13]. Accordingly, one domain rarely has representation enough to learn its own concepts, and the higher number of parameters in hidden layers, the greater amount of samples is required [14]. In skin cancer images scenario, the aforementioned issues are even more evident, since images are obtained in hospital environments without any acquisition control (noise, different illumination, and presence of clutter [4]). Therefore, the domain alone has discrepancies that strengthen the challenge of CNN training. In addition, the number of collected and annotated images is limited making it almost impossible to train a CNN from scratch [15]. In this context, the use of pre-trained networks with different domains becomes a viable option in skin cancer analysis.

CNN methods were found to be efficient also for skin cancer classification [16]. In [17], authors collected a large image dataset, with many categories, and they were capable of building a model that was comparable to dermatologists when classifying lesions, but this dataset is not fully available. The generality of feature spaces in this context is put into question,

since small variations in the test set can lead to a significant decrease in testing error [18], and some models may memorise the data distribution, resulting in over-training [19].

Since the classification depends directly on the built-in feature space, the goal of this paper is to analyse feature spaces generated by a pre-trained CNN and impacts caused by distortions applied to images, to discuss the generality of such features and the potential for transfer learning. We found that, as expected, it is possible to use a CNN model to learn features from the raw dataset. Features from PH2 benchmark dataset [20] were extracted in three different CNNs pre-trained with ImageNet challenge dataset [21]: MobileNet [22]; VGG-19 [23]; and ResNet50 [24]. In the best raw feature space extracted was applied dimensionality reduction with PCA. In parallel, colour quantisation and noise addition (Gaussian and impulsive) were assigned, separately, to the raw set, generating new feature spaces. This structure is illustrated in Fig. 1.

Consequently, our contribution include: (i) the use of several CNN models and different layers for feature extraction and skin lesions image classification with and without fine-tuning process; (ii) a detailed study of the impact of dimensionality reduction in the final classification; (iii) in-depth analysis of colours space contraction and noisy effects in the feature space; and (iv) feature generalisation analysis between raw and distorted sets. To our knowledge, this study is the first that analyses feature space's robustness for skin lesions, in particular using PH2 dataset.

II. Related Work

The use of convolutional networks as feature extractor in medical field is seen in several studies, such as analysis of blood images for leukemia diagnosis [25], mammography images classification [26], and chest pathology [27]. Accordingly, this scenario may be expanded to other diagnoses, e.g skin lesion classification. Despite the use of the same pre-trained CNN, AlexNet [28], these studies [16], [29], [30] are distinguished by the layer choice used as descriptor and by different skin lesion images datasets, making a direct comparison among performances impossible. Majtner et. al [30] compared handcrafted methods in contrast to second last layer of AlexNet. Meanwhile, Pomponiu et. al [16] explored the last three layers as descriptors, separately, to measure the best accuracy among them. Similarly to [16], Mahbod et. al [29] used the same adjustment (layers) from AlexNet, including feature map fusion and latest VGG-16 [23] layers. However, only dense layers were exploited, considering plenty of semantic information contained in ImageNet dataset. Also, there are newer architectures showing better properties when compared to AlexNet, that can be explored.

Considering that there is not a large labeled dataset enough to fine-tune a state-of-the-art CNN and feature extraction from a CNN pre-trained in another domain is widely used, this paper explores different layers (intermediate-level and top-level) for skin lesion classification. As we described, our investigation is supported by many experiments, including generalisation, which remains well discriminative, practically maintaining same accuracy. We also show that fine-tuning with same skin lesion domain does not offer discriminative capability as ImageNet due to the limited amount of examples available. Consequently, our study explores a new horizon in this field, in order to analyse more deep layers systematically [31] in more complex networks. We consider that intermediate layers may offer greater discriminative spaces than the latter ones.

III. Feature Extraction

Feature extraction was performed using the following CNNs: MobileNet [22]; VGG-19 [23]; and ResNet50 [24]. These models were chosen because they have different structures and layer depths. While MobileNet is considered a lightweight model, ResNet50 is very demanding in processing. VGG-19 is intermediate in this factor. First, skin images were re-sized to 224×224 pixels (architectures restriction). For each CNN, pre-trained with ImageNet [21], the feature space was obtained using activation values from each last seven layers.

Among many layers contained in a CNN, three types stand out with great relevance: convolutional; pooling; and fully connected (FC). Convolutional layers provide a filters set, with fixed size, in which an activation function generates a space representation as input for the next layer. Due to the large increase in the number of parameters accumulated during successive convolutions, pooling layers operate dimensionality reduction. When there is a transition between convolutional and FC, after the pooling layer is placed a flatten (reshape) layer to realign the input for the next layer without content transformation. At the model top, FC layers aim to vectorize feature maps, converting the data to classes probabilities contained in the training dataset [19].

Every CNN has in its architecture filters capable of providing features of low-level and high-level [9]. As the image proceeds through first layers, the built-in feature map adds both shape, border, and color information. Because of hierarchical model, the aspect of these layers, regardless architecture, refers almost exclusively to Gabor filters or color blobs [31]. This important property allows networks pre-trained in datasets from different domains to be used as feature descriptors for a target domain. In situations where dissimilarity between the target domain and the training domain applied to the network is evident, the semantic information contained in last layers should be avoided or minimized. However, bottom-level and top-level layers threshold is still uncertain, with several heuristics prevailing to determine the optimal layer for each problem [31].

MobileNet uses the concept of depthwise separable convolutions. A standard convolutional layer joins inputs and filters into an output set in a single step. However, depthwise convolution keeps the data separated, one layer for filtering (depthwise convolution) and another for combining them (pointwise convolution). In this structure, pointwise convolution (1×1) performs linear combination among filters applied to input (single filter per channel). The factorisation allows a model size reduction and less computation cost [22].

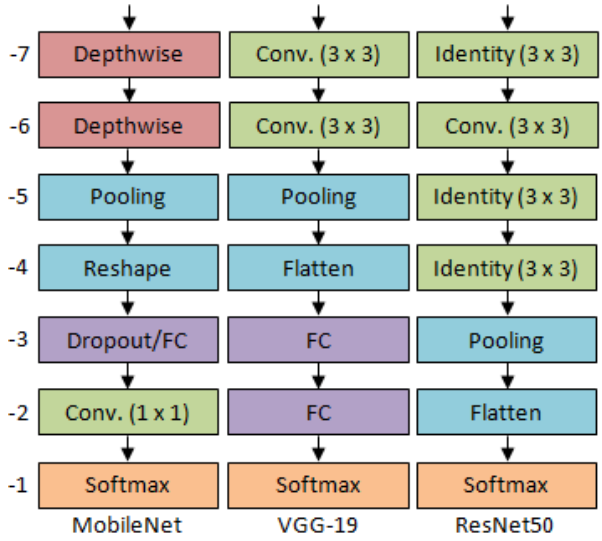

Fig. 2. Last seven layers from MobileNet, VGG-19, and ResNet50. These CNNs have different structures: MobileNet used dephtwise convolution and dropout layers; and ResNet50 does not contain FC layers. Meanwhile, they all have a prediction layer as the last one (softmax). Reshape and flatten layers have same feature space of its previous layer in a new shape. Dropout is a layer with the purpose of deactivates neurons with a defined probability. In ResNet50, identity layers do not have the addition of input with data transformation, occurring only in the convolutional layer. We refer to layers as -1 (last layer), then -2 (one before the last), and so on until -7.

VGG-19 was developed with 19 layers in which most of filters are 3×3 size. The almost exclusive use of this size is based on the concept that two consecutives 3×3 filters have an effective receptive field equivalent to one 5×5 filter, and three 3×3 filters can be use as one 7×7 filter. Moreover, features from this hierarquical model are more discriminative and the amount of parameters is smaller [23].

ResNet applies residual blocks to allow training networks with greater number of layers, e.g. 152 layers. Residual blocks aim to preserve features from input vector before its transformation, adding both values as output of delimited block. Another interesting property in this architecture is the absence of FC layers. ResNet used in this paper applies three weight layers for each residual block, resulting in 50 layers [24].

In common, the last layer of each CNN corresponds to probabilities of each class (1000 categories from ImageNet), as shown in Fig. 2. As we can see, the final structure of each CNN differs in layers composition and, consequently, in the quantity of attributes. Therefore, for MobileNet, layers output from 1000 to 50176 features, and for VGG-19 and ResNet50 from 1000 to 100352 features.

IV. Experiments

PH2 [20] is a benchmark dataset for skin lesion classification, composed of 200 dermoscopic images, divided into two main categories: malignant (40 melanomas) and non-malignant (80 common nevi and 80 dysplastic nevi). These lesions have a variety of chromatic and texture appearance, and an original resolution of 768×574 RGB pixels, as shown in Fig. 3. By the appearance of skin lesions, categories differ in shapes, edges, and colors (low-level features). In principle, the more irregular these properties are, the greater is the malignancy likelihood in the lesion.

(a) (b) (c)

Fig. 3. Skin Lesions: (a) Common Nevus; (b) Displasic Nevus; and (c) Melanoma. The preliminary diagnosis of skin cancer includes visual analysis of low-level features, being that common nevi have more regular structures than melanomas. Additionally, in these samples are also evident presence of confusing objects in the image composition, such as the black circle on margins and some bubbles superimposed on the lesion (c), increasing the challenge of finding an adequate feature space.

A. Classifiers

For our analysis, three classifiers were used: Linear SVM (LSVM); Random Forest (RF); and AdaBoost (ADA). The choice of LSVM is to verify how linearly separable the generated space is. RF is a more complex classifier that constructs an ensemble of decision trees, producing many decision boundaries. ADA algorithm is also an ensemble method based on linear classifiers. Our experiment used 100 iterations for ADA, and 100 trees for RF. Intuitively, a more adequate feature space performs well in LSVM then in ensemble methods, since LSVM is the algorithm with a more restricted bias, and it has stronger learning guarantees [32].

B. Feature Spaces

After performing a standard scale normalisation on feature vectors obtained from last layers of MobileNet, VGG-19, and ResNet50, we computed the balanced accuracy using a 20-folds cross validation (each fold is class-balanced) for LSVM, RF, and ADA. For each layer, we also verify the real amount of features used (Variance). Despite the high dimensionality provided from layers, those are sparse, with many attributes having no variance, i.e. with equal value on all examples [33]. These specific attributes do not contribute to the classification, only increase the computational cost. Due to this reason, a cleanup is performed on the data to eliminate these attributes.

Results in Table I show details of the experiments, including the number of features actually used. It can be seen that the best performance was obtained in earlier layers: LSVM achieved 94% in MobileNet (-3); RF and ADA achieved 88.5% and 93%, respectively, in MobileNet (-5) and ResNet50 (-5). As expected, layers containing predictions (-1) achieved significantly poorer results: LSVM 80.5% in ResNet50; RF 84%, and ADA 83%, both using VGG-19. Due to the randomly cross validation folders creation, layers with the same space representation (MobileNet layers -4 and -5, VGG-19 layers -4 and -5, and ResNet layers -2 and -3) do not have the exactly balanced accuracy, meanwhile the variation is minimal.

In addition to best accuracy, the feature space provided by MobileNet is more compact, containing attributes more

TABLE I
CNNs: 20-FOLDS CROSS VALIDATION BY BALANCED ACCURACY (%)

CNN	Layer	Features	Variance	LSVM	RF	ADA
MobileNet	-1	1000	1000	85.0 ± 12.04	87.0 ± 7.14	91.0 ± 9.95
	-2	1000	1000	92.0 ± 8.72	87.0 ± 7.81	85.5 ± 12.44
	-3	1024	1024	**94.0 ± 6.63**	84.5 ± 7.4	87.5 ± 8.87
	-4	1024	1024	93.5 ± 7.26	86.0 ± 8.0	85.0 ± 9.22
	-5	1024	1024	93.0 ± 8.43	88.5 ± 8.53	84.5 ± 7.4
	-6	50176	(90.2%) 45263	90.5 ± 8.65	83.5 ± 8.53	89.0 ± 8.89
	-7	50176	50176	91.5 ± 7.26	86.5 ± 7.26	87.5 ± 8.87
VGG-19	-1	1000	1000	81.0 ± 12.61	84.0 ± 5.83	83.0 ± 9.0
	-2	4096	(93.7%) 3837	88.5 ± 6.54	88.0 ± 8.72	89.0 ± 8.31
	-3	4096	(93.7%) 3839	88.5 ± 8.53	86.0 ± 7.35	86.0 ± 7.35
	-4	25088	(86.8%) 21774	89.0 ± 6.24	84.0 ± 5.83	87.5 ± 8.87
	-5	25088	(86.8%) 21774	88.5 ± 7.26	82.0 ± 5.1	87.5 ± 9.94
	-6	100352	(75.2%) 75440	91.5 ± 7.92	86.5 ± 7.26	87.5 ± 8.87
	-7	100352	(92.8%) 93085	91.5 ± 6.54	88.0 ± 7.48	91 ± 5.39
ResNet50	-1	1000	1000	80.5 ± 11.17	88.5 ± 9.1	88.5 ± 7.92
	-2	2048	2048	90.0 ± 7.75	88.5 ± 9.63	88.0 ± 9.8
	-3	2048	2048	90.5 ± 7.4	85.5 ± 7.4	88.0 ± 10.77
	-4	100352	(96.3%) 96684	91.5 ± 7.92	85.0 ± 6.71	88.5 ± 4.77
	-5	100352	100352	91.5 ± 7.26	85.0 ± 7.42	93.0 ± 8.43
	-6	100352	100352	90.5 ± 7.4	85.0 ± 8.06	90.0 ± 7.75
	-7	100352	100352	90.5 ± 9.73	84.0 ± 5.83	88.5 ± 8.53

discriminative than VGG-19 and ResNet50 for skin images classification. This discriminative capacity is evidenced not only by the amount of attributes, but mainly by the variance contained in these attributes. VGG-19 generates more attributes without variance (all layers), and its performance is also surpassed by ResNet50. However, the best result achieved by ResNet50 (ADA in layer -5) demonstrates the complexity of the generated space: 100352 features and 100 iterations.

Hence, the lighter CNN, MobileNet, provides the best results in terms of accuracy versus model complexity and feature space dimensionality. This indication corroborates the idea that smaller datasets that are application-specific do not need networks with high capacity. Results with feature spaces obtained from different layers, shown in Fig. 4, imply that semantic features, extracted from last layers are less relevant for skin lesions classification, while earlier layers may be more adequate. MobileNet presents a steady curve, falling only in layer -6. In contrast, VGG-19 has two growth peaks (layers -1 to -2 and -5 to -6), and ResNet50 maintains itself balanced from layers -2 to -7. MobileNet also provides, overall, a more linearly separable feature space, while VGG-19 and ResNet50 features require more hyperplanes to discriminate classes, which was confirmed by better performances using RF and ADA (see Table I). Therefore, all further analysis are carried out on MobileNet best layer (-3).

Based on these results, we can analyze layers in three groups: (i) prediction layers; (ii) dense layers between softmax and pooling/flatten; and (iii) layers prior to pooling/flatten. As expected, softmax layers provide the worst results because they represent probabilities of training dataset classes in other domain, as shown in Fig. 5. Therefore, the dissimilarity between domains makes those not adequate. Other layers have different behaviour for different CNN architectures. Because it is more compact, MobileNet has less complexity and the

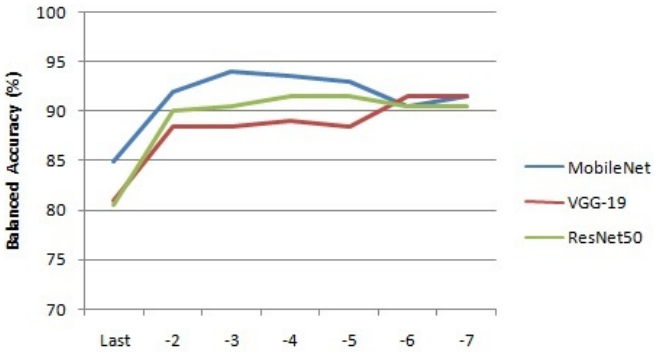

Fig. 4. LSVM balanced accuracy in last seven layers. MobileNet has better performance (94% in layer -3) and stays on top (layers -2 to -5). However, ResNet50 is more consistent (90% to 91, 5% from layers -2 to -7). VGG-19 has a growing accuracy from softmax layer into hidden layers.

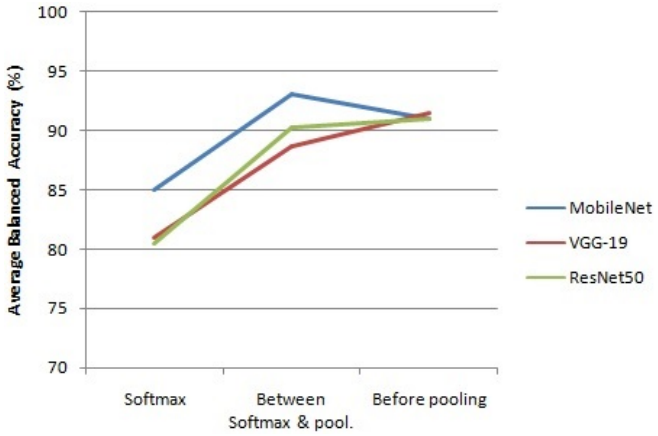

Fig. 5. Results in group of layers. In the second group, MobileNet and VGG-19 have four layers and ResNet50 only two. Hence, ResNet50 is more concentrated in convolutional layers (before pooling) with four layers.

second group is more discriminative (93.12%). All three CNNs are equivalent considering layers prior to pooling (91% to 91.5%). Due to the large number of layers in Resnet50, the second and third groups are similar (90.25% and 91%). For these reasons, the ideal layer to be used as feature extractor must be investigated depending on the capacity of the network.

C. Dimensionality Reduction

The output of CNN layers are often high-dimensional vectors. Depending on the number of available examples the dimensionality can be an issue in this scenario. We use PCA algorithm in the feature space provided by MobileNet layer -3, originally with 1024 features, to select from 128 to 1 principal components, halving the size each step. As seen in Fig. 6, a space with dimensionality between 64 to 16 features (LSVM $\approx$ 92%) is sufficient to discriminate the data without serious performance loss. Other classifier results follow similar behaviour: there is a smaller feature space with similar accuracy: RF has higher performance gain (between 32 and 4 dimensions); and ADA is more stable from 128 to 32 features. However, LSVM tends to show better performance overall, implying that reducing dimensionality does not affect the linear separability of the space, and that the manifold of the data lies in a subspace with less dimensionality.

Fig. 6. Dimensionality reduction by PCA, MobileNet (layer -3). LSVM is only surpassed by RF and ADA with extreme space representativity contraction. Curves of each classifier have a lack of smoothing due to the small size of the dataset. These dimensions were selected to show a gradual feature space reduction. In addition, PCA imposes as a limitation the minimum between samples (200 skin lesions) and amount of features (1024 for MobileNet layer -3).

TABLE II
PCA VARIANCE FROM MOBILENET (LAYER -3)

Number of Features	Variance
1024	1.0000
128	0.9585
64	0.8533
32	0.7314
16	0.6080
8	0.4870
4	0.3502
2	0.2181
1	0.1254

To complement the dimensionality reduction discussion, we show the percentage of variance maintenance in Table II, showing that a 60.80% variance threshold allow class separability (LSVM $\approx$ 92%, 16 features). As expected, PCA variance decreases gradually as the space contraction increases.

D. Colour Quantisation and Image Noise

We investigate the impact of colour quantisation in the quality of feature spaces, as it can play a significant role in feature extraction [34]. New versions of the raw dataset were created by computing 64, 32, and 16 colours per channel. In general, the feature space becomes less linearly separable, yielding lower performance, as presented in Table III.

TABLE III
QUANTISED AND NOISY SPACE: 20-FOLDS CROSS VALIDATION BY BALANCED ACCURACY (%)

Set	LSVM	RF	ADA
Quant 64	94.5 $\pm$ 4.97	84.5 $\pm$ 4.97	88.0 $\pm$ 6.78
Quant 32	92.5 $\pm$ 8.29	86.5 $\pm$ 7.92	89.5 $\pm$ 7.4
Quant 16	90.0 $\pm$ 9.49	87.0 $\pm$ 8.43	89.5 $\pm$ 7.4
G 0.008	93.0 $\pm$ 7.81	85.0 $\pm$ 7.42	89.0 $\pm$ 6.24
G 0.016	93.0 $\pm$ 6.4	87.0 $\pm$ 7.14	86.0 $\pm$ 9.17
G 0.032	94.5 $\pm$ 7.4	87.5 $\pm$ 8.29	87.0 $\pm$ 10.1
SP 0.005	95.0 $\pm$ 6.71	88.0 $\pm$ 8.72	91.5 $\pm$ 7.26
SP 0.01	91.5 $\pm$ 9.1	88.0 $\pm$ 6.78	89.0 $\pm$ 8.31
SP 0.02	90.5 $\pm$ 8.65	87.5 $\pm$ 6.98	88.5 $\pm$ 9.1

Similarly to the quantisation experiment, we artificially generated Gaussian and impulsive (salt & pepper) noisy images, with results also presented in Table III. For each type of noise, three sets of images were generated: Gaussian noise variances are 0.008, 0.016, and 0.032; salt & pepper noise probabilities are 0.005, 0.01, and 0.02. With the progressive increase of Gaussian noise, the accuracy of classifiers remains relatively constant, especially with LSVM. On the other hand, the impulsive noise shows initially a positive impact (with SP 0.005 by LSVM), but then results degrade.

E. Cross-training/test for generalisation analysis

To study feature spaces in more depth, we employed different dataset versions (regarding quantisation and noise levels) for training and testing. This experiment has the intention of measuring how well the feature space generalises for images with unseen quantisation or noise levels. We performed Hold-out 50/50 in which folder contained exactly 80 non-malignant and 20 malignant lesions. The final balanced accuracy is the average of each tested set. As in previous experiments, LSVM obtained the best result in general. Contrary to LSVM performance, RF and ADA are less robust to colour quantisation influence and noise injection, with accuracies dramatically impoverished. It is interesting to observe that the experiment showed some average when quantised or noisy versions of images are used and, in particular improve the robustness of the representation when testing on noisy and quantised images. Regarding the quantisation, the generalisation reduces according the bigger distances among colours spaces, as expected. The same behaviour occurs with noisy application. All results are shown in Table IV.

TABLE IV
FEATURE SPACE GENERALISATION - BALANCED ACCURACY (%)
HOLD-OUT 50/50

Training	Testing	LSVM	RF	ADA
Raw	Quant 64	91.5	84.0	86.5
	Quant 32	90.0	86.0	86.0
	Quant 16	86.5	82.0	87.5
Quant 64	Raw	91.0	82.5	90.0
	Quant 32	90.5	85.0	89.0
	Quant 16	87.0	84.5	88.5
Quant 32	Raw	90.0	82.0	86.5
	Quant 64	91.0	84.5	89.0
	Quant 16	87.0	83.5	86.0
Quant 16	Raw	91.0	81.5	81.5
	Quant 64	91.5	84.5	82.5
	Quant 32	91.5	85.5	84.0
Raw	G 0.008	90.5	83.0	85.0
	G 0.016	89.0	80.5	87.0
	G 0.032	88.5	83.5	86.0
G 0.008	Raw	90.0	83.0	86.0
	G 0.016	89.0	83.5	87.5
	G 0.032	88.0	85.0	84.0
G 0.016	Raw	89.5	84.5	84.5
	G 0.008	88.5	85.5	85.5
	G 0.032	88.0	84.0	86.0
G 0.032	Raw	90.0	82.5	81.0
	G 0.008	90.0	85.5	80.5
	G 0.016	89.0	85.5	84.0
Raw	SP 0.005	89.5	85.0	88.0
	SP 0.01	88.5	84.0	85.5
	SP 0.02	88.5	81.0	84.0
SP 0.005	Raw	92.5	81.5	84.0
	SP 0.01	89.0	83.5	85.5
	SP 0.02	91.0	82.5	83.0
SP 0.01	Raw	88.0	84.0	85.0
	SP 0.005	88.0	83.0	85.0
	SP 0.02	89.5	82.0	85.0
SP 0.02	Raw	89.0	84.5	84.0
	SP 0.005	89.5	84.0	85.5
	SP 0.01	89.0	81.5	82.0

The average balanced accuracy obtained by each training set in the colour quantisation implies in the training of CNN with mild colour reduction parameter (16 bins per channel) which provides the best generalisation of the feature space (average of 91.33% with LSVM). In this context, the high complexity in colour space implies a greater difficulty of generalisation for images with less amount of intensities available (average of 89.3% using Raw as training and LSVM as classifier). With respect to noise types, additive noise causes a positive impact: the lowest average reached among Gaussian sets was 88.66% (for G 0.016); similarly for salt & pepper, with a highest average of 90.83% (for SP 0.005). These results indicate that noise can cause positive perturbations on the data so that the CNN model produces a more robust space, facilitating the classification algorithm to find a linearly separable classifier even for unseen noise/quantisation levels.

Furthermore, as highlighted in Fig. 7, the performance reduces according to the greater dissimilarity between original and distorted sets. Considering only the Raw set as training, colour quantisation proves deep degradation of feature space (Raw with 94% and Quant 16 with 86.5%). Despite the recurrent deterioration, noise addition behaves with less variation between transitions (Gaussian by 2% and SP by only 1%).

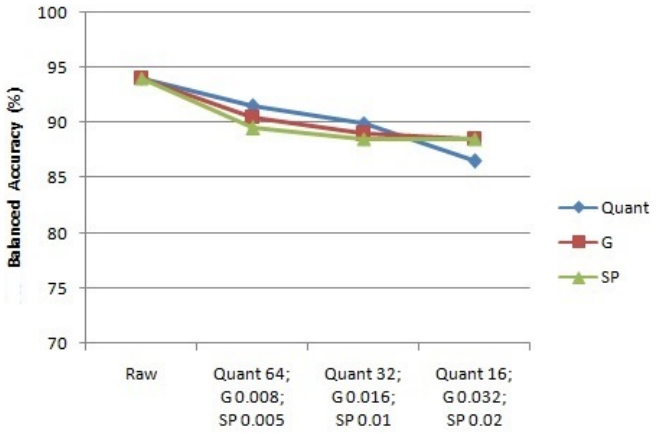

Fig. 7. LSVM balanced accuracy by training with Raw. As expected, as the colour space contraction or the noise concentration increases, the similarity between sets decreases and, consequently, the performance. The horizontal axis of the graph denotes test sets.

F. Fine-tuning

Aiming to confirm the high performance from MobileNet (layer -3) pre-trained with ImageNet, we performed fine-tuning with all CNNs used in previous experiments. Fine-tuning consists of using initial parameters obtained with a large dataset (e.g ImageNet), copying the first n layers, and re-train some of the last layers using available annotated images. To fine-tune the CNNs, we employed the HAM10000 dataset [35], [36] with 10000 skin lesions images categorized into 7 distinct classes (training set).

In the fine-tuning process, one may freeze pre-trained layers, allow them to adapt, or even reinitialise them with random values [37]. For VGG-19 and ResNet50, all layers before the pooling were maintained. The last layers (4 layers in VGG-19 and 2 layers in ResNet50) are randomly initialised. From that, the last seven layers were allowed to adapt, freezing previous ones. For MobileNet we employed two approaches: the first one resets all dense layers (5 layers); and the second one resets only the last two layers which are related to the classifier.

We used a batch size of 32 images (stipulated optimal value [38]), and the Adaptive Moment Estimation (Adam) with a binary cross-entropy loss function [37], [39]. In addition, this experiment was carried out with 10, 25, 50, 100, and 500 epochs. In Table V, we present the best result achieved in each adaptation. MobileNet (last 2) had the best overall result (91.5% with LSVM). Other adaptations showed similar results, and slightly below the ones without fine-tuning.

Note that, due to the greater complexity of architectures, both VGG-19 and ResNet50 need more epochs to obtain their best results. Note that MobileNet and ResNet50 quickly converged to perfect training accuracy, For the VGG-19 network, the loss saturated at 1.477, which kept unchanged even after 500 epochs. This indicates the images used for fine-tuning did not offer a relevant gradient in terms of classification loss that allowed improvement. This result can be interpreted

TABLE V
BEST RESULTS FROM FINE-TUNING

CNN	Training Loss (%)	Training Accuracy (%)	Epochs	Layer	Features	Classifier	Test Accuracy (%)
MobileNet (Last 2)	0.7	100.0	10	-3	1024	LSVM	91.5 ± 9.1
MobileNet (Full Top)	1.6	99.0	10	-5	1024	ADA	90.5 ± 7.4
VGG-19	147.7	90.7	50	-7	100352	ADA	91.0 ± 7.68
ResNet50	0.0	100.0	50	-5	100352	LSVM	90.0 ± 7.75

TABLE VI
FEATURE SPACE GENERALISATION - BALANCED ACCURACY (%)
HOLD-OUT 50/50 AFTER FINE-TUNING

Training	Testing	LSVM	RF	ADA
Raw	Quant 64	83.0	81.5	83.0
	Quant 32	84.0	80.5	81.5
	Quant 16	84.5	81.0	83.5
Quant 64	Raw	90.0	87.0	88.5
	Quant 32	84.5	82.5	84.0
	Quant 16	83.0	79.5	82.5
Quant 32	Raw	90.5	86.0	85.5
	Quant 64	85.0	83.0	84.0
	Quant 16	84.0	83.0	83.5
Quant 16	Raw	89.5	82.5	85.0
	Quant 64	85.5	83.5	86.0
	Quant 32	86.5	84.0	86.5
Raw	G 0.008	83.0	83.0	84.0
	G 0.016	82.0	82.5	81.5
	G 0.032	82.5	81.0	84.0
G 0.008	Raw	85.5	82.0	84.0
	G 0.016	86.5	85.0	84.0
	G 0.032	85.0	80.5	82.0
G 0.016	Raw	85.5	83.0	82.5
	G 0.008	85.5	81.0	81.5
	G 0.032	84.0	80.0	81.5
G 0.032	Raw	87.5	80.5	83.0
	G 0.008	86.0	83.0	85.5
	G 0.016	85.0	80.0	81.5
Raw	SP 0.005	82.0	79.5	82.5
	SP 0.01	80.5	80.5	80.0
	SP 0.02	79.5	81.0	79.5
SP 0.005	Raw	85.0	81.5	83.5
	SP 0.01	83.5	80.0	80.5
	SP 0.02	84.0	82.0	83.0
SP 0.01	Raw	83.0	80.0	83.0
	SP 0.005	82.0	82.0	82.0
	SP 0.02	82.5	77.5	79.5
SP 0.02	Raw	79.5	80.0	81.5
	SP 0.005	80.0	81.5	80.0
	SP 0.01	79.5	80.5	80.0

intuitively according to the architectures: MobileNet is smaller allowing convergence, while ResNet50 converges due to the skipping layers. On the other hand VGG-19 is to deep for the problem, and, without using skipping layers, could not converge properly. We also performed generalisation experiments with feature extraction from fine-tuning MobileNet (last 2), as shown in Table VI. However, feature generalisation from HAM10000 is lower in comparison to ImageNet training parameters, being worse, mainly, with noise injection.

G. Competing Methods

We compared our results with competing state-of-the-art methods in PH2 dataset (see Table VII). Overall, CNN feature extraction followed by a LSVM classification produces results above all competing methods. MobileNet achieved even better results, reaching 94% balanced accuracy with raw images, while competing methods comprise preprocessing steps to achieve at most 90.31%. As shown in our feature generalisation experiments without fine-tuning, even with noisy images, results are robust (see SP 0.005 results).

TABLE VII
COMPETING METHODS RESULTS (%), OURS IN BOLD

Method	Accuracy	Balanced Accuracy
Barata et. al (2015) [40]	—	84.3
Bi et. al (2016) [41]	92.0	90.31
Salido and Ruiz Jr. (2018) [42]	93.0	—
VGG-19 (LSVM)	90.5	91.5
Fine-tuning (LSVM)	84.0	91.5
Hold-out [Raw, SP 0.005] (LSVM)	89.5	92.5
ResNet50 (ADA)	91.5	93.0
MobileNet (LSVM)	95.0	94.0
SP 0.005 (LSVM)	94.0	95.0

V. CONCLUSION

We report an in-depth analysis of feature spaces from raw PH2 dataset, extracted using last seven layers of state-of-the-art CNNs. The best space was generated by a dense layer of MobileNet (94% balanced accuracy), bettering competing methods. Only 16 features were sufficient to represent the raw space with good accuracy (92%), which is relevant for a low running time in practical scenarios. Furthermore, we performed generalisation studies, indicating that the space remains adequate and more robust with artificial perturbation.

In this sense, our study offers important guidelines for future studies with the use of pre-trained CNNs for feature extraction. Researchers can leverage pre-trained CNNs with ImageNet and other very large datasets to obtain feature spaces even for different image domains, such as skin lesion images, exploring more mid-level layers. Noise injection of SP type may help producing more robust feature spaces, while fine-tuning with images from different datasets and same domain should be investigated with care. The depth of the CNN must be considered when performing fine-tuning: architectures with less capacity or employing skipping layers seem to converge better.

VI. ACKNOWLEDGMENTS

The authors would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), FAPESP (grant #2016/16111-4), and CNPq (grant #307973/2017-4) for financial support. This work is also partially supported by the CEPID-CeMEAI (FAPESP grant #2013/07375-0).

REFERENCES

[1] M. Fornaciali, S. Avila, M. Carvalho, and E. Valle, "Statistical learning approach for robust melanoma screening," in *Graphics, Patterns and Images (SIBGRAPI), 2014 27th SIBGRAPI Conference on*. IEEE, 2014, pp. 319–326.

[2] E. Bernart, J. Scharcanski, and S. Bampi, "Segmentation and classification of melanocytic skin lesions using local and contextual features," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2633–2637.

[3] M. H. Jafari, S. Samavi, S. M. R. Soroushmehr, H. Mohaghegh, N. Karimi, and K. Najarian, "Set of descriptors for skin cancer diagnosis using non-dermoscopic color images," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2638–2642.

[4] R. B. Oliveira, N. Marranghello, A. S. Pereira, and J. M. R. Tavares, "A computational approach for detecting pigmented skin lesions in macroscopic images," *Expert Systems with Applications*, vol. 61, pp. 53–63, 2016.

[5] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble model," *IEEE transactions on medical imaging*, vol. 36, no. 3, pp. 849–858, 2017.

[6] M. Mete and N. M. Sirakov, "Optimal set of features for accurate skin cancer diagnosis," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2256–2260.

[7] A. Parolin, E. Herzer, and C. R. Jung, "Semi-automated diagnosis of melanoma through the analysis of dermatological images," in *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*. IEEE, 2010, pp. 71–78.

[8] A. Pennisi, D. D. Bloisi, D. Nardi, A. R. Giampetruzzi, C. Mondino, and A. Facchiano, "Skin lesion image segmentation using delaunay triangulation for melanoma detection," *Computerized Medical Imaging and Graphics*, vol. 52, pp. 89–103, 2016.

[9] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[10] D. Mishkin, N. Sergievskiy, and J. Matas, "Systematic evaluation of cnn advances on the imagenet," *arXiv preprint arXiv:1606.02228*, 2016.

[11] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya, "Understanding the mechanisms of deep transfer learning for medical images," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 188–196.

[12] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1667–1680, 2012.

[13] X. Li, M. Fang, and J.-J. Zhang, "Projected transfer sparse coding for cross domain image representation," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 265–272, 2015.

[14] R. F. de Mello, M. D. Ferreira, and M. A. Ponti, "Providing theoretical learning guarantees to deep learning networks," *arXiv preprint arXiv:1711.10292*, 2017.

[15] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4068–4076.

[16] V. Pomponiu, H. Nejati, and N.-M. Cheung, "Deepmole: Deep neural networks for skin mole lesion classification," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2623–2627.

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

[18] T. Nazare, G. P. da Costa, W. Contato, and M. A. Ponti, "Deep convolutional neural networks and noisy images," in *Iberoamerican Conference on Pattern Recognition (CIARP 2017)*, vol. LNCS 10657, 2017.

[19] M. Ponti, L. S. Ribeiro, T. S. Nazare, T. Bui, and J. Collomosse, "Everything you wanted to know about deep learning for computer vision but were afraid to ask," in *30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T 2017)*, 2017, pp. 17–41.

[20] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 5437–5440.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Internat. Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] L. H. S. Vogado, R. D. M. S. Veras, A. R. Andrade, F. H. D. De Araujo, R. R. V. e Silva, and K. R. T. Aires, "Diagnosing leukemia in blood smear images using an ensemble of classifiers and pre-trained convolutional neural networks," in *Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on*. IEEE, 2017, pp. 367–373.

[26] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 652–660.

[27] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 294–297.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[29] A. Mahbod, R. Ecker, and I. Ellinger, "Skin lesion classification using hybrid deep neural networks," *arXiv preprint arXiv:1702.08434*, 2017.

[30] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, "Combining deep learning and hand-crafted features for skin lesion classification," in *Image Processing Theory Tools and Applications (IPTA), 2016 6th International Conference on*. IEEE, 2016, pp. 1–6.

[31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[32] V. Vapnik, *Statistical learning theory*. Wiley, 1998.

[33] H. Bagherinezhad, M. Rastegari, and A. Farhadi, "Lcnn: Lookup-based convolutional neural network," in *Proc. IEEE CVPR*, 2017.

[34] M. Ponti, T. S. Nazaré, and G. S. Thumé, "Image quantization as a dimensionality reduction procedure in color and texture feature extraction," *Neurocomputing*, vol. 173, pp. 385–396, 2016.

[35] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *arXiv preprint arXiv:1803.10417*, 2018.

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

[37] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244*, 2016.

[38] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.

[39] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Trans. Med. Imaging*, vol. 36, no. 9, pp. 1876–1886, 2017.

[40] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE journal of biomedical and health informatics*, vol. 19, no. 3, pp. 1146–1152, 2015.

[41] L. Bi, J. Kim, E. Ahn, D. Feng, and M. Fulham, "Automatic melanoma detection via multi-scale lesion-biased representation and joint reverse classification," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 1055–1058.

[42] J. A. A. Salido and C. Ruiz Jr, "Using deep learning for melanoma detection in dermoscopy images," *International Journal of Machine Learning and Computing*, vol. 8, no. 1, pp. 61–68, 2018.