Received 23 June 2022, accepted 8 July 2022, date of publication 19 July 2022, date of current version 26 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3192444

RESEARCH ARTICLE

Multiple Instance Learning Using 3D Features for Melanoma Detection

PEDRO M. M. PEREIRA[1,2], LUCAS A. THOMAZ[1,3], (Member, IEEE), LUIS M. N. TAVORA[1,3], PEDRO A. A. ASSUNCAO[1,3], (Senior Member, IEEE), RUI FONSECA-PINTO[1,3], RUI PEDRO PAIVA[2], AND SERGIO M. M. FARIA[1,3], (Senior Member, IEEE)

[1]Instituto de Telecomunicações, Morro do Lena—Alto do Vieiro, 2411-901 Leiria, Portugal
[2]Centre for Informatics and Systems, Department of Informatics Engineering, University of Coimbra, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
[3]School of Technology and Management, Polytechnic of Leiria, Morro do Lena—Alto do Vieiro, 2411-901 Leiria, Portugal

Corresponding author: Pedro M. M. Pereira (pedrommpereira@co.it.pt)

This work was supported in part by the Fundação para a Ciência e a Tecnologia (FCT), Portugal, through Ph.D. work under Grant SFRH/BD/128669/2017; in part by the Programa Operacional Regional do Centro, PlenoISLA Project POCI-01-0145-FEDER-028325; and in part by FCT/MCTES through National Funds through the European Union (EU) Funds under Project UIDB/50008/2020 and Project LA/P/0109/2020.

ABSTRACT This work presents a contribution to advance current solutions for the problem of melanoma detection based on deep learning (DL) approaches. This is an active research field, which aims to aid on the detection and classification of melanoma (the most lethal type of skin cancer) with non-invasive solutions. By exploiting both 2D and 3D characteristics of the skin lesion surface, the proposed approach advances beyond commonly used colour features of dermoscopic images. Two competing classification methods are exploited, namely Multiple Instance Learning (MIL) and DL, which are combined using an uncertainty-aware decision function. The DL method performs classification resorting to RGB data, while MIL performs 3D feature extraction, selects the most significant set, and performs classification at two different learning instances. The novel aspects of this work include DL uncertainty evaluation mechanisms along with MIL to train a robust ensemble classifier, and also the use of dense light-fields for skin lesion classification. Despite the large class imbalance (often present in medical image datasets), the ensemble model achieves cross-validated melanoma classification accuracy of 84.00% when trained against nevus lesions, and 90.82% accuracy when discriminating against all present lesion types. The results show that, in the absence of discriminative 2D characteristics, the 3D surface provides redeeming results, demonstrating that existing methods can benefit from the proposed method by looking beyond 2D image characteristics.

INDEX TERMS 3D features, classification, light-fields, medical image analysis, melanoma, skin lesion.

I. INTRODUCTION

Medical image processing, identification, and classification has been evolving for decades to assist dermatologists [1]. One of the most common research topics related to skin lesions involves the identification of melanoma skin cancer, a life-threatening dermatological disease. This type of skin lesion, which is increasingly common in the global population [2], develops from pigmented melanocytes and is hard to distinguish from other benign lesions – namely nevi. Therefore, the world population would benefit from automation's that lead to the immediate and automatic classification of cancerous lesions through non-invasive methodologies.

Seeking a solution, current systems tend to use the same type of information as dermatology experts – i.e. dermoscopic imagery (2D/colour). However, since a fully satisfactory solution is still to be found, other image modalities or fairly unexplored data dimensions ought to be considered. One of these modalities is 3D imaging (e.g., stereo), which has already proven to enhance skin lesion classification performances due to the additional dimension, i.e. depth information [3], [4].

The main contribution of this paper is to demonstrate that features of such third dimension are beneficial for the melanoma classification process. With this aim, an ensemble

The associate editor coordinating the review of this manuscript and approving it for publication was Chulhong Kim.

model is proposed. It enables melanoma classification by resorting to 3D surface data when the initial colour classification is uncertain, thus demonstrating that features of such third dimension are beneficial for the classification process. Both the colour of 2D images and corresponding depth information are used by resorting to a dataset of light-field skin lesions, which grants the inspection of the new dimension, providing more knowledge and decision boundaries to classification pipelines. Here, classification of colour or depth information is performed separately. For 2D information, a Transfer Learning (TL) approach [5], comprising a Deep Learning (DL) model, is used. While for the depth information, features extracted from the 3D surface feed a Multiple Instance Learning (MIL) approach when the DL model shows high uncertainty towards classification. Both local and global features are used to characterise the 3D depth surfaces. Feature selection also takes place and is performed by an automatic feature reduction algorithm that allows the model to cope with the dataset size. The usage of a small DL network to process the colour dimension is preferred to Machine-Learning feature-extraction approaches (as in [6]–[8]) since it can better couple with any shift in the dataset.

The remainder of the paper is organised as follows: Section II presents the background information and state-of-the-art that is relevant for this work. Section III describes the proposed model and the corresponding pipeline, including relevant details about data pre-processing, segmentation, model training, feature extraction and selection, and classification. Finally, Section IV presents and discusses the attained results while Section V highlights the conclusions and future work.

II. BACKGROUND

Machine Learning (ML), in particular for image recognition or classification, has become a major topic in a wide range of research fields because of the ability to learn abstract data models and intrinsic discriminative properties. The datasets used for training, validation and testing are crucial elements required for research on image classification with machine learning. Amongst the most important of them is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), currently considered one of the standard reference and used in recent years as a benchmark standard for large-scale object recognition, i.e. image classification, single-object location and object detection. ImageNet has been used by many authors to improve their image classification/recognition algorithms. Its use promoted an exponential growth of research results and significant improvements to the state-of-the-art techniques [9].

The remainder of this section is structured into subsections addressing the different concepts that support this work.

A. DL APPROACHES FOR SKIN LESION CLASSIFICATION

In recent years, the DL paradigm has attracted research in several domains of medical image analysis, demonstrating that noticeable improvements are achieved beyond conventional approaches [10]–[13]. In the field of skin lesion classification, Convolutional Neural Networks (CNN) have also produced promising results [13]–[17]. In [18] a CNN, pre-trained on the ILSVRC, is used as a feature extractor (rather than trained from scratch). This work demonstrated that the existing filters (used on the ILSVRC natural images) generalise well for a set of 10 classes using non-dermoscopic images. More recently, research with such pre-trained models reported the highest performance measurements ever published across multiple test datasets [5], [19]. The use of pre-trained models is typically accompanied by a TL approach [20], [21], which can be further aided by manually extracted features (e.g., as in [22]).

In [5], classification of segmented colour skin lesions is performed using TL with the pre-trained AlexNet CNN [23]. In this work, data augmentation is based on image rotation and the model classification layer is replaced with the appropriate Softmax layer for either: melanoma and nevus (binary); or melanoma, seborrheic keratosis, and nevus (ternary) classification. After fine-tuning the model weights on each dataset and performing augmentation in both train and test sets, the reported system accuracy performance was measured as 96.86%, 97.70%, and 95.91% for three different datasets. While performance without augmentation was recorded at 88.24%, 91.18%, and 87.31% for the same datasets.

B. UNCERTAINTY

Sometimes, DL classification results are enhanced with the model's inner statistics, namely the features' distribution that exists before a Softmax layer. If the model's values prior to this layer are not well separated, it might indicate that the model is uncertain of the which target label is the correct answer, or even that both are equally correct. For this reason, some researchers look for a better solution to replace the Softmax layer [24]. These model's values can be used to determine network class uncertainties or, for example, the CNN belief in the classification of the segmented pixels [25]–[27]. Such uncertainty values, which exist before the Softmax layer, have been used to improve CNN models [25], [27], [28]. As highlighted in [29], further inspection of uncertain decisions results in better performance. Additional research on uncertainty can be found in [27].

C. MULTIPLE INSTANCE LEARNING (MIL)

Assuming the calculation and usage of such uncertainties, other models can be used, or combined, to compensate, when a previous DL model is uncertain of its classification output. These other models need not be of matching technique and can even be a new combination of multiple models. When a new model depends on multiple outputs of a previous one, such composition is known as Multiple Instance Learning (MIL). This concept was introduced in 1991 [30], and was later used, in 1998 [31], to solve a machine vision scene classification problem.

In [30], an instance is defined as one or more fixed-size sub-images of a given image, and the bag of instances is the image itself. An image is labelled positive if it contains a target scene related instance or negative otherwise. For this to work, it is assumed that a relationship between the instances within a bag and the class label of the bag exists, allowing the classification itself to be performed in several ways. For example, given each instance classification, a bag of instances can be given the final label by a thresholding model, by a count-based assumption, by the presence of a single positive or negative class, or by more complex models – like for example a multidimensional-polynomial-border created by a Support Vector Machine (SVM) model [32].

In [33], the same concept is exploited. The authors proposed a weakly supervised DL framework with uncertainty estimation in order to address a disease classification problem. Firstly, a CNN instance-level classifier is iteratively refined by using the proposed uncertainty-driven deep MIL scheme. Secondly, a Recurrent Neural Network takes each of the previous instances features (from a same bag/image) as input and generates the final prediction, considering each local instance and their global aggregated representation.

D. SEGMENTATION

Most methods dealing with skin lesion classification require some form of prior lesion segmentation or region identification [5]. Several previous works present some form of skin lesion segmentation to prepare the data for classification, such as [10], [12]–[15], [21], [22], [34], [35]. This preprocessing step is typically needed since skin information (or image acquisition artefacts) can produce outlier features or expand the dimension of the hyperspace in which the parameter search is performed by DL algorithms (as, for example, with CNN) – both undesirable outcomes. A relevant example of such method is described in [36], where the image is segmented into super-pixels using local features and then iteratively merged into regions to form two classes of regions (lesion and non-lesion), while considering a spatial continuity constraint on the super-pixels colour.

E. DATASET

The majority of publicly available datasets for skin lesion classification only include conventional 2D images [37]. In these datasets, the images can be separated into two sources: dermoscopic, as those in the PH^2 [38], Dermofit [39], and Atlas [40] datasets; and macro, as in MED-NODE [41]. Typically, these datasets are comprised of small resolution images that, given the nature of this type of medical data, have high class imbalanced and a relatively small amount of samples.

Due to the limitations imposed by the planar nature of 2D images, few works have attempted to use different image modalities. The main motivation has been to find out whether additional information, beyond RGB, can be helpful for skin lesion classification. An example of such alternative modalities is stereoscopic technology, as in [3], [4], which has already shown to be more efficient than single-view images to identify skin lesions, using the obtained disparity information. Even so, literature on 3D surface of melanoma or related skin lesions is still very scarce, as well as datasets including this type of information. However, existing research indicates that using richer information, including depth information (3D) of the lesion surface, contributes to improve the classification accuracy of skin lesions. For instance, the study presented in [42] attempted to use artificially generated 3D information to enhance an existing 2D dataset, with moderate success.

As a contribution for research using 3D skin lesion data, a dataset named Skin Lesion Light-fields (SKINL2) was acquired and made public to enable further advances in skin lesion classification using 3D surface information [43]. This dataset was created with light-field images, using a handheld light-field camera. Light-fields have the advantage of enabling the extraction of several multi-view photographs in one single shot, and the reconstruction of a detailed skin surface map. At the time of writing, to the best of the authors knowledge, there are no works published by other authors resorting to this recent dataset for the purpose of 3D skin lesion classification. There are, however, some previous exploratory works by the authors dealing with 3D features extracted from this dataset [15], [34], [35].

F. HAND-CRAFTED FEATURES

As mentioned before, most works in the literature rely on 2D datasets, that either extract hand-crafted features for melanoma classification or, more recently, use DL or TL to automate the process. Some of these hand-crafted features include: colour, distribution, shape, texture, and border irregularity [1], [37], [44]. After the feature extraction step more automated machine learning methods such as K-Nearest Neighbours, Artificial Neural Networks (ANN), Logistic Regression, Decision Trees, and SVMs are used to perform classification – typically with no more than moderate success [1], [37]. Hence the literature transition in recent years to more rewarding DL methods, which relieve the research on new features. Examples of related work using 2D hand-crafted features and known classifiers can be found in [1], [21].

So far, there are no 3D features specifically studied for melanoma classification. Thus a primary approach towards defining a relevant set of such features is to look at other research fields, where 3D features have been used. Depending on the target recognition task, several 3D features have been developed and generalised across multiple 3D datasets and tasks. This type of generalisation is performed to propose a set of features that capture a broad spectrum of 3D characteristics – typically applied to key regions. In general, an algorithm responsible for extracting the designed features is called feature extractor and the key regions where these feature extractors are applied are determined by a keypoint detector. In the scope of this work, the Normal Aligned Radial Features (NARF) [45] is used as both a keypoint detector and feature extractor. Other relevant feature extractors are the following:

- Radius-based Surface Descriptor (RSD) [46],
- Global RSD (GRSD) [47],
- Globally Aligned Spatial Distribution (GASD) [48],
- Rotation Invariant Feature Transform (RIFT) [49],
- Point Feature Histogram (PFH) [50],
- Fast PFH (FPFH) [51],
- Signature of Histograms of OrienTations (SHOT) [52],
- Ensemble of Shape Functions (ESF) [53],
- 3D Shape Context (3DSC) [54],
- Unique Shape Context (USC) [55].

G. FEATURE SELECTION

In many cases, the initial number of features can be overwhelming for the classification algorithm, particularly when the number of data samples is not enough to enable a correct understanding of all feature space combinations. Thus, feature reduction is necessary to select the most meaningful ones, which can be done by using several methods such as, for instance, using a diagonal adaptation of Neighborhood Component Analysis (NCA) [56]. NCA is a non-parametric algorithm that enables feature selection with the goal of maximising the prediction accuracy of regression and classification algorithms. The algorithm performs better when estimating feature importance for distance-based supervised models that use pairwise distances between observations to predict the response. NCA can be understood as a pre-processing step before the classification step, as in [57], allowing the removal of similar or noisy features from the feature space. But it can also be used between models [58], namely when initial DL models produce too many latent features in comparison with the amount of available data samples [59].

III. PROPOSED APPROACH

As pointed out before, in addition to conventional colour (RGB, left column in Fig. 1), the proposed approach also explores depth information (Z, middle column in Fig. 1) to improve beyond current classification results. To this end, a new pipeline was devised (as summarised in Section III-A), to operate over a dataset with lesion segmentation masks (generated as described in Section III-B). This pipeline utilises both a DL process, as a baseline 2D classification model (Section III-C), as well as a two-step model scheme that resorts to hand-crafted features from the 3D surface (Section III-D). This is an ensemble classification approach, where the objective is to collectively obtain better predictive performances than those from any of the individual learning algorithms on its own. In order to increase the reliance on the attained model (and the produced results), a cross-validation scheme is used to show that the outcome is consistent and thus supportive of the findings.

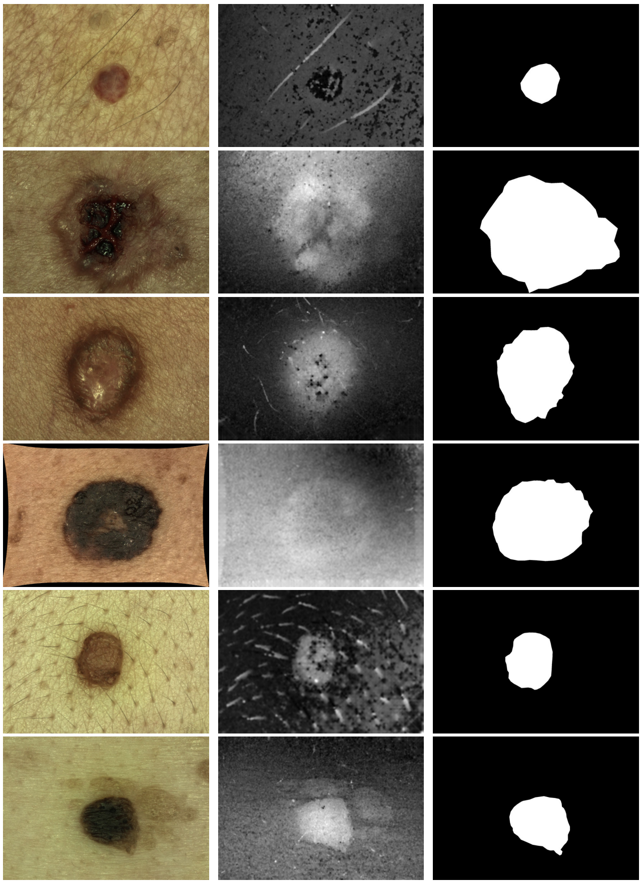

FIGURE 1. Sample SKINL2 dataset images. The left column displays RGB images, the middle column shows depth (Z) values in grayscale, and the right column contains generated segmentation mask images. From top to bottom, samples show: angioma, carcinoma, dermatofibroma, melanoma, nevus, and seborrheic keratosis.

A. OVERVIEW

An overview of the pipeline is depicted in Fig. 2. Given a 4D dataset (with its lesion segmentation masks), at any 10-fold cross-validation (CV) partition k, a Train$_k$ and Test$_k$ datasets are received by the ensemble pipeline. As training precedes the test step, the Train$_k$-set is first used to train both a TL model and a MIL model prior to the use of the Test$_k$-set. Note that the same data is used by both the TL and MIL models.

TL is performed with a DL model to update its weights to the classification problem at hand. The other part of the ensemble classifier (MIL) comprises a two-step learning approach.

The Softmax layer present in the CNN model allows to predict the level of confidence the CNN has in its prediction, which is known as the model certainty. It can be asserted either naively or by imposing alternative computations. Therefore, if the CNN 2D classification model is certain of its prediction it is set as the ensemble prediction, otherwise, the MIL 3D-classification model is preferred.

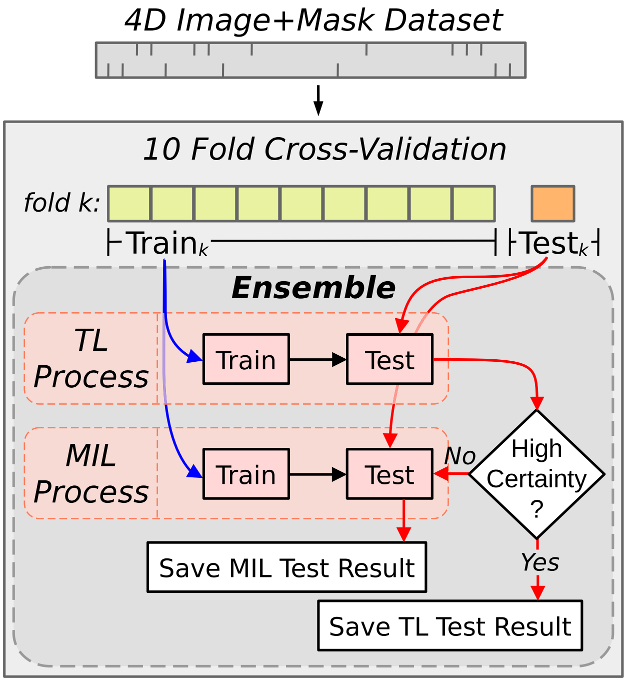

FIGURE 2. Proposed pipeline: a given dataset is partitioned into 10 folds for cross-validation; at any stage, both TL and MIL are trained on 9 training folds and later tested on 1 test fold; afterwards, if the TL process certainty is high, the TL test image classification result is recorded, otherwise the MIL classification result is recorded. Blue arrows indicate the pipeline training sequence. Red arrows indicate the pipeline testing sequence. Black arrows indicate previous dependencies and/or abstract progressions through the pipeline.

B. SEGMENTATION

The employed segmentation method is based on a modified version of the Lazy Snapping algorithm [60], which resorts to an internal method to group similar pixels. However, in this work, such method is replaced by a more recent approach described in [61], which has been observed to achieve good performance in coloured images of skin lesions in [36].

Given an RGB coloured image (Fig. 3, top-left), pixels are first grouped into super-pixels (Fig. 3, top-right) using the Simple Linear Iterative Clustering (SLIC) method [61]. This method serves as a pre-processing step for the Lazy Snapping algorithm, as it compacts the problem dimension to less samples (super-pixels). In this work, the SLIC compactness is set to 10 and its clustering phase is performed for 10 iterations. Then, the Lazy Snapping algorithm constructs a graph of the image super-pixels, where each super-pixel is a node connected by weighted edges. The higher the probability that pixels are related, the higher the weighted edge. The algorithm cuts along weak edges, achieving the object segmentation by maximising the colour similarity within the object. To generate the necessary binary segmentation mask, that separates foreground from background, the graph-cut is guided with user provided information (Fig. 3, bottom-left) about pixels belonging to the lesion (foreground, green points in the figure) and pixels belonging to the non-lesion skin (background, red points in the figure). Given the user input, the Lazy Snapping algorithm then outputs the segmentation mask (Fig. 3, bottom-right).

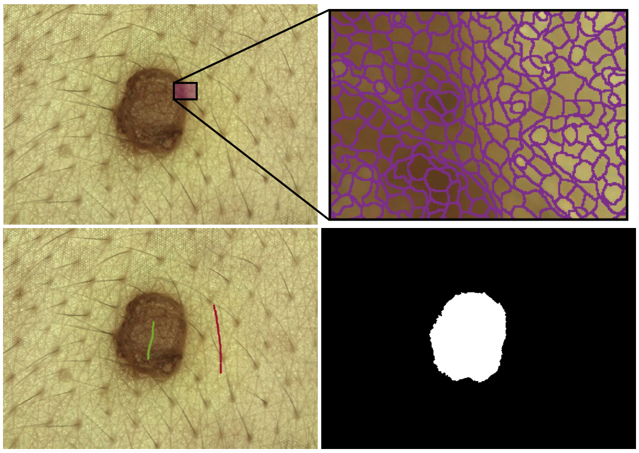

FIGURE 3. Segmentation method: given a dataset coloured central-view image (top-left); the image pixels are grouped through superpixel over-segmentation (top-right); then, some pixels regarding the lesion (green) and skin region (red) are marked to guide the segmentation process (bottom-left); lastly, a segmentation mask is generated (bottom-right).

C. TL PROCESS

Starting with the employed TL process, only the 2D RGB images are processed along with their segmentation masks. This process mainly comprises a model named AlexNet, which is obtained with the ILSVRC pre-trained weights. The DL training process is performed for 32 epochs of batch size 10, with a learning rate of 0.001. The colour images first undergo a segmentation process (described in Section III-B), so that non-lesion skin can be coloured black – effectively removing colour information and forcing the CNN to focus on the RGB characteristics of the target lesion area. This is performed in alternative to a crop and resize process because applying different crop shapes and different resize ratios would make the problem more difficult for the network. Data augmentation (by online rotation) is also performed, as described in [5]. The retrained layers are also the same as in [5].

Additionally, in the present work an enhancement is performed to allow the model to be aware of its classification uncertainty. Instead of naively using the internal Softmax probabilities for the ensemble model uncertainty, the model is reinforced with the capability to generate its internal classification certainty during training. Thus, the loss function is changed from the default Softmax cross entropy to the sum of two components [25], as expressed by

$$\text{loss} = \frac{\|v_1\|_2^2 + \|v_2\|_2^2}{2} \times 0.005 + \text{UIF}(e^o, a), \tag{1}$$

where $\|X\|_2^2$ represents the L2-norm, defined as $\frac{1}{2}\sum x_i^2$, where x_i are the elements of the vector X, v_1 and v_2 are the outputs of the first and last classification layers, o are the values at the end of the network, and a the target classification one-hot label probabilities. The mean-square-error uncertainty-infused function (UIF), is expressed as

$$\mathrm{UIF}(b,a) = (b - \frac{a}{s})^2 + \frac{a(s-a)}{s^2(s+1)} + \mathrm{KL}(\mathrm{P}(b,a)\|Q), \quad (2)$$

where s is the sum of all one-hot exponential values and KL is the Kullback-Leibler divergence term, defined as $\mathrm{KL}(P'\|Q)$, where P' is the result obtained from applying Eq. (3) and Q is the one-hot distribution.

$$\mathrm{P}(b,a) = (a-1) \times (1-b) + 1 \quad (3)$$

The KL divergence is used in this context to regularise the predictive distribution by penalising predictions that diverge from the desired uncertainty (which is known as Learned Loss Attenuation [25]).

At the end of the Alexnet uncertainty-infused-model training stage, a classification uncertainty for each class can be obtained by dividing the number of possible output classes by the natural (Euler) exponential of the values outputted by the network. For the proposed ensemble, a classification certainty above 50% is considered high (refer to Fig. 2, ''High Certainty''). This means that the final output of the ensemble model will be: the TL process output if the classification certainty is above 50%; or reevaluated with the MIL process if below 50% (or equal) certainty.

Both the classification labels and uncertainties are output to the ensemble definition described in Section III-A.

D. MIL PROCESS

The MIL process performs skin lesion classification using only 3D surface information. This means that, from the available SKINL2 data exemplified in Fig. 1, the RGB data shown in the left column is discarded and not used in this process. A detailed pipeline of the this process is depicted in Fig. 4. Note that the correct dependency-flow starts with the training stage (blue-arrows), which might initiate black-arrow flows. Any procedure is only executed if all input training flows (arrows) are present or if it has already been executed for training.

The process comprises four main blocks, each being executed only after the previous one's completion. Blocks named 1 and 2 comprise the dataset pre-processing stage with feature extraction and selection, while blocks 3 and 4 comprise the actual MIL aspect of the process. Detailed information about each block is provided in the following four subsections.

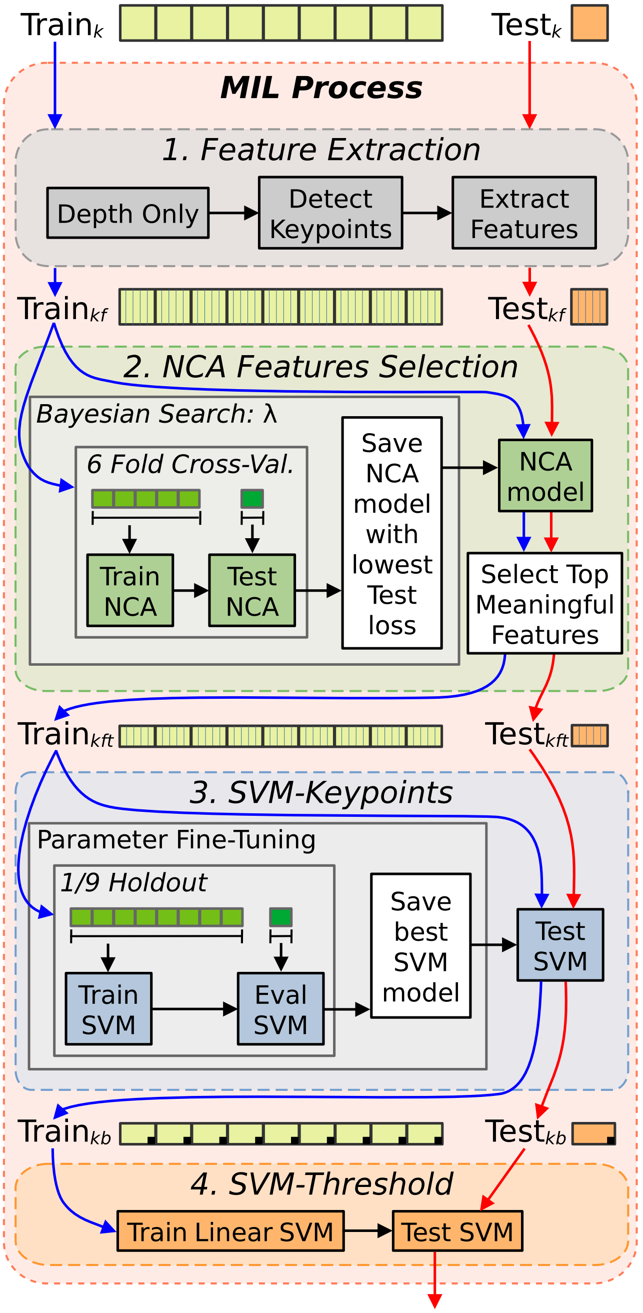

FIGURE 4. MIL process pipeline comprising four blocks: *1)* given an image dataset, only depth information is kept and features from detected image keypoints are extracted; *2)* given a keypoints training-set, an NCA model is created with the lowest possible loss. This is obtained by performing a Bayesian search over a 6-fold cross-validation of the train data to find the NCA optimal λ value. Given the NCA model, the algorithm advances to next block with the top meaningful features; *3)* given the selected features, a fine-tuned SVM model for keypoint classification is created. This tuning occurs through a parameter search using 1-out-of-9 folds for evaluation of said SVM model, and the SVM keypoint classification labels bagged by image advance to the next block; finally, *4)* given a dataset with bags of labels, a linear SVM is trained to provide the grouped image final classification label. Blue arrows indicate the pipeline training sequence. Red arrows indicate the pipeline testing sequence. Black arrows indicate previous dependencies and/or abstract progressions through the pipeline.

1) FEATURE EXTRACTION

Given either a training or a test-set of 4-channel images (RGB+Z), pixel values in the RGB channels of all input images are replaced with zeros. This operation is performed to guarantee that no colour related feature is generated, meaning that further processing only uses depth. Having only the 3D surface, the NARF keypoint detector elects several key locations in each image. Using the lesion masks (as described in Section III-B), after a dilation process to extend each mask by 25 pixels, keypoints not belonging to the new lesion region are discarded. Fig. 5 provides

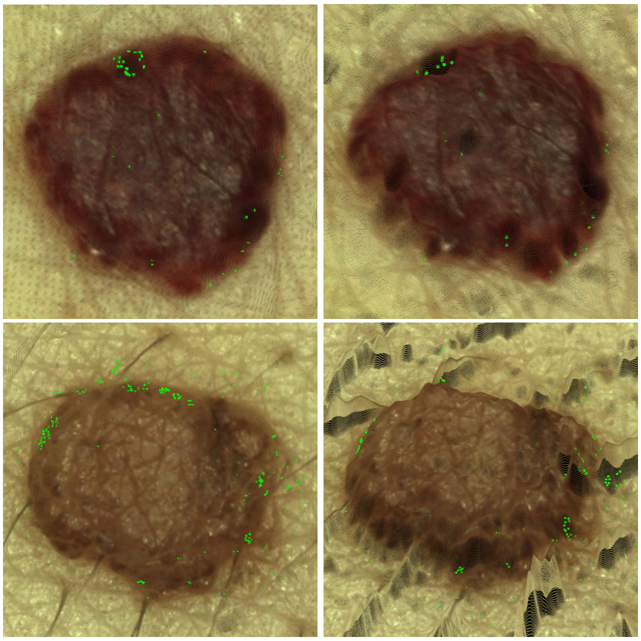

FIGURE 5. **Visualization of keypoint locations (in green) in excerpts of two image samples from the SKINL2 dataset. Left column images provide a top visualization in 3D space, while images on the right column provide a different point-of-view to highlight depth perception.**

examples of keypoint distribution on a lesion's surface. Feature extractors are applied to each of the remaining keypoint locations. In essence, this block generates a new dataset$_{kf}$ from the input set of images, where each image is now represented by multiple instances (keypoints) of multiple features.

The NARF keypoint detector was selected because it seems specially suited for skin lesion images, since it selects the surface locations where abrupt changes occur and takes object borders into account, such as skin-to-lesion borders, which have already been recognised as relevant information [62]. The keypoint detector has two major characteristics. First, keypoints are extracted in areas where the direct underlying surface is smooth and the neighbourhood contains major surface changes. The resulting keypoints are located in the vicinity of significant geometric structures and not directly on them. Second, NARF takes object borders into account. Such objects are detected when non-continuous transitions from the foreground to the background arise. Thus, the silhouette of an object has a strong influence on the resulting keypoints.

As for the feature extractor methods, 11 are utilised, generating a total of 5726 features per keypoins. The distribution of keypoints per method is shown in Table 1. These extractors were selected based on the relevance of their characteristics for the type of input signal in use (i.e., 3D information). RIFT (32 features) was selected because it provides invariance to illumination, viewpoint, scale, and rotation. Like RIFT, NARF (42 features), and PFH/FPFH (125/33 features) also possess some of these characteristics, PFH/FPFH, in particular, provides robustness against outliers and noise. Other features extractors as SHOT (361 features), SC3D (1989 features), and USC (1969 features) also provide robustness against noise. Additionally, both SHOT and USC are reported to provide uniqueness amongst detection, as well as unambiguous representations. Finally, ESF (640 features), RSD (2 features), GASD (512 features), and GRSD (21 features) were selected for being descriptive, simple, and intuitive shape descriptors. ESF has proven to be efficient and expressive, while GRSD adds expressiveness to the simple RSD by partitioning the image point cloud into several voxel-surfaces of understandable shapes.

2) NCA FEATURE SELECTION

Given a (training) feature dataset, feature reduction is performed resorting to an NCA model. This is done because some of the extracted features might not contribute for the adequate label separation during later classification process.

Since NCA is a data-driven algorithm, it is possible that, without due care, the generated feature's meaningfulness-weight is overfitted to the training data. To overcome this problem, NCA includes a regularisation parameter λ that helps to prevent overfitting. Since this parameter has to be predefined, the method performs a Bayesian search for the λ value that originates the lowest average test loss of a six-fold CV partitioning scheme of the given (training) features dataset. This inner CV is implemented to further prevent data overfit.

Having found the NCA model with the optimal λ, the (training) features dataset can now be reduced to the most meaningful features. Meaningfulness-weights obtained from the training data can be applied to later testing-sets. In this model, only features with a normalised absolute meaningfulness greater than 0.02 are selected – meaning that features with meaningfulness-weights below 2% are discarded. In essence, this block generates a new dataset$_{kft}$ from the feature dataset, where only features relevant to classification are maintained.

The implemented Bayesian search is performed by constraining λ values to the range $[0.00001, 0.1]$, using four initial seeds randomly chosen from the λ search range. This search is executed for 50 steps, comprising 24 evaluations each. To promote a balance between the search exploitation and exploration [63], the Bayesian propensity to explore is 0.5. In addition, to avoid over-exploiting, the acquisition function in [63] is modified as suggested in [64].

As for the NCA model parameters, the inner network is optimised using Stochastic Gradient Descent and an initial learning rate is determined by selecting 200 random dataset samples and training a temporary model on increasing learning rates for 15 epochs. The learning rate providing the lowest loss is selected as the initial learning rate (on average, the initial learning rate is 51.2000). With the initial learning rate defined, the network is trained using all training data (five-folds) over 10 epochs with a mini-batch size that enables

TABLE 1. Summary of used features extractors and the number of features their provide.

Reference	[45]	[46]	[47]	[48]	[46]	[50]	[51]	[52]	[53]	[54]	[55]
Extractor Name	NARF	RSD	GRSD	GASD	RIFT	PFH	FPFH	SHOT	ESF	SC3D	USC
Total Features	42	2	21	512	32	125	33	361	640	1989	1969

at least 40 iterations per epoch. At each epoch, the learning rate is decreased when a convergence tolerance of step size 0.000005 is met.

Since there is only one regularisation parameter (λ) for all weights, and the weight magnitudes must be comparable, i.e. within the same range, any dataset data entering the model is normalised with zero mean and unit standard deviation.

3) KEYPOINT-LEVEL CLASSIFICATION

Given a (training) dataset of meaningful-features, which comprises multiple instance data (keypoints) for each image, label classification of each image keypoint takes place resorting to a SVM. As SVMs have several hyper-parameters, parameter fine-tuning is necessary at this stage. Due to the complexity of the pipeline, the theoretical determination of the optimal SVM implementation is not feasible. Assuming an initial data partitioning into 10 folds CV (as detailed in Section III-A), the training data comprises nine folds. Therefore, the last fold is hold-out from the SVM classification training process, so that it can be later used for the SVM selection during the parameterisation fine-tuning. This single fold is called evaluation fold. Having found the SVM model with the best performance in the evaluation fold, the same model can be applied to later testing-sets. In essence, this block generates a new dataset$_{kb}$ comprising a bag of classified keypoints, that is, a label classification for each keypoint of each image. During pipeline training stage, classification label results from both training and evaluation folds advance to the next block as one training-set – i.e., maintaining the original dataset data sample counts. Evaluation results will not be perfect, but this is helpful during the next pipeline block training stage as it provides behavioural insight of how the model operates on unseen data.

As for the SVM parameter fine-tuning, instead of using a full Bayesian search, several predefined parameters were evaluated for simplicity. The SVM kernel function can be either linear, polynomial, or Gaussian. In the case of polynomial, it can be either of order 2 or 3. In the case of Gaussian, it can be either of kernel scale 0.9, 3.6, or 14. Data normalisation always takes place and the box constraint is set to 1. This enables the evaluation of six different SVM models in total. The quadratic kernel SVM is typically the top performing.

The SVM solver is the Iterative Single Data Algorithm (ISDA) [32], given that this is a binary classification problem. In addition, the SVM also comprises a custom cost matrix, which is set to [0 1; 2 0] in order to enforce a double penalty when miss-classifying the melanoma class. In this matrix, each element consists of the cost of guessing that a sample belongs to class X (lines) when it belongs to class Y (columns), leaving all elements of its main diagonal equal to zero. This matrix was empirically defined considering that misclassification in the melanoma class have a more severe outcome.

4) IMAGE-LEVEL CLASSIFICATION

Finally, given bags of labels, a last SVM model provides the image-level label classification. Since the objective is to reduce a variable-sized list of keypoint-level labels to a single image-level label, the data is summarised to enable thresholding. That is, given an arbitrary number of data samples belonging to an image, the data is transformed into two sums: the number of melanoma labels and the number of non-melanoma labels. Then, these sums are normalised to the [0, 1] range, while making their sum 1 – producing a probability distribution over predicted output classes, as occurs in a Softmax layer. Furthermore, these probabilities are given as features to a SVM model of linear function and ISDA solver, with box constraint set to 1, and without implicit data standardisation. In a training pipeline, this effectively produces a threshold along the probability distribution that attempts to separate the target class labels. A SVM is used rather than a common thresholding technique, due to its capability for better forming the threshold boundary and also because it would enable future work beyond binary classification. As in the keypoint-level classification, the SVM cost matrix is adjusted in order to enforce a double penalty when miss-classifying the melanoma class ([0 1; 2 0]).

In a testing pipeline, the linear SVM model image-level labels are sent to the ensemble, as described in Section III-A.

IV. RESULTS AND DISCUSSION

The performance of the proposed method is evaluated and discussed in this section, encompassing two classification experiments, both executed applying 10-fold CV, as previously mentioned. The selection of the number of folds was based on a balance between the significance of the results and the diversity and representativeness of the training data. A larger number of folds would be possible but would not add significant value, while less folds would make inner training sets less representative of the underlying data. The first experiment, named "M vs Nevus", consists in melanoma classification against nevus samples, i.e., a more difficult task, and the second experiment, named "M vs All", covers classification of melanoma versus all other skin lesion types (including nevus).

TABLE 2. Features outputted in feature selection block.

Feature Extractor Name	Features Inside Extractor	Unique Features Selected†‡		Total # Features Selected†‡		# Times Extractor is Used†‡	
ESF	640	28	26	126	92	10	10
FPFH	33	0	0	0	0	0	0
GASD	512	35	32	200	149	10	10
GRSD	21	2	2	14	13	9	7
NARF	42	0	2	0	2	0	1
PFH	125	0	0	0	0	0	0
RIFT	32	2	3	7	17	5	9
RSD	2	0	0	0	0	0	0
SC3D	1989	0	0	0	0	0	0
SHOT	361	0	0	0	0	0	0
USC	1969	0	0	0	0	0	0

† Values resulting from the aggregation of 10-fold cross-validation.
‡ Values for *M vs Nevus* and *M vs All* results, respectively.

A. DATASET

The proposed pipeline was applied to the publicly available SKINL2 dataset [43]. The dataset comprises light-field imagery of skin lesions acquired using a Raytrix R42 camera, captured at a hospital facility (Centro Hospitalar de Leiria, Portugal) from patients previously screened by a physician during dermatology clinical appointments. The procedure and purpose of the study was explained to all volunteers, who also signed an informed consent form. Procedures related to the image acquisition, storage, and publication were evaluated and approved by a health ethics committee. The information available from acquisition process are the RGB channels with resolution of 3858 × 2682 and relative depth of each pixel.

Particularly in this work, the second [65] and third versions of this dataset were used. Both versions are present due to their increase in lens magnification of approximately 30% (which means more detail) in comparison to its first version. At the time of publication of this paper, the third version is still in development and is used in this work as an extension to the second version. The combined dataset includes 14 melanomas, 36 nevi, and 48 other lesion types (16 angiomas, six basal cell carcinomas, one dermatofibroma, 24 seborrheic keratoses, and one verruca). Among these lesions, 70 belong to the second dataset and 28 from the third. All images undergo the pipeline described in Section III.

Therefore, experiment *M vs Nevus* comprises 14 melanoma samples against 36 nevus samples, while experiment *M vs All* comprises 14 melanoma samples against all other 84 non-melanoma samples.

B. FEATURE SELECTION

In the pipeline described in Section III, the MIL process is responsible for performing the classification when the TL process does not have enough certainty. The feature selection performed within fold samples in this step is a key component of the former process. Depending on the fold, different dataset samples arrive at *NCA Feature Selection* block (III-D2), which in turn will induce different features to be marked as meaningfully in different folds for the classification objective.

Table 2, comprising five major columns, provides some statistics regarding feature selection. For each feature extractor in the first column, the number of inner features comprising said extractor is shown in the second column. Subsequent columns are sub-divided to provide information for either the *M vs Nevus* or the *M vs All* experiment, respectively. Across the 10-fold execution, the number of unique features that are selected at least once are defined in the third column, while the total amount of features (regardless of repetition) selected across folds is presented in the fourth column. Finally, the fifth column indicates how many times a feature extractor is used (that is, if any of its features were used in any given fold).

Table 2 shows that most literature features considered potentially relevant for melanoma surface discrimination are not selected. This can be considered a normal behaviour since features with higher discriminative power overshadow the lesser ones, making the NCA model algorithm reduce their meaningfulness to marginal values. This occurs as they do not present added information to the higher representative features.

Table 2 also shows that only the ESF, GASD, GRSD, and RIFT feature extractors are selected across the two experiments, with NARF being used in only one fold of the second experiment (*M vs All*). Concerning the feature extractors, it can be seen that, if the uniquely selected features were always the same across folds (third column), then the total amount of features selected (fourth column) would be 10 times that value – which is never the case. However, this does not mean that no feature is meaningful enough to be selected across folds.

For the five selected feature extractors, Fig. 6 presents the number of times each feature extractor is used across folds (bar plot representing the same information as in Table 2), as well as the number of times each feature extractor's feature is used (scatter plot). From this figure, it is possible to observe that: in the *M vs Nevus* experiment, four ESF and six GASD features are always selected (i.e. having usage count equal to 10) independently of the fold data, while in the *M vs All* experiment, only two ESF and two GASD features are always selected. This suggests that discrimination between melanoma and nevus is possible in more ways than in melanoma versus every other class (as evidenced by the scatter plot's data-points spread). Also, in the second experiment, the NCA model algorithm excluded some features while adding others, namely including two features from NARF in one fold (as previous mentioned). All in all, from one experiment to the other, a total of 50 features change from either being or not being used in the experiment pipeline, while 40 remain in usage at least once. On average across folds, the feature selection block chooses 33.8±4.8488 features in the *M vs Nevus* experiment, and 28.1±4.5080 features in the *M vs All* experiment.

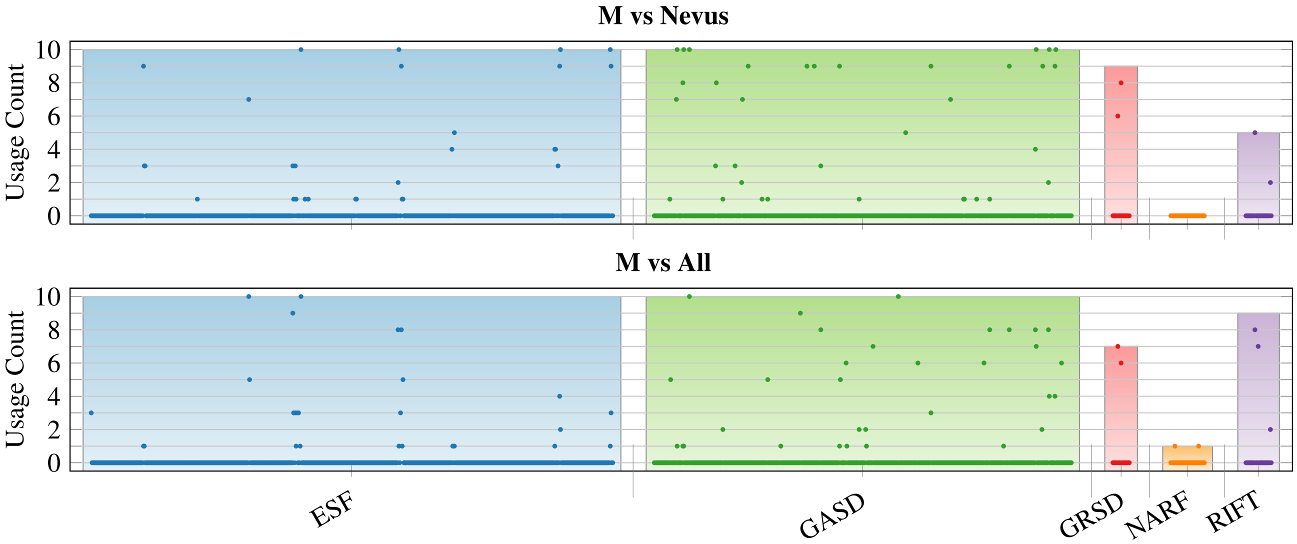

FIGURE 6. Number of times that features (or feature extractors) are selected during the 10-fold cross-validation process. Scatter plot values indicate how many times a given feature (from a feature extractor) is marked meaningful for classification during the feature selection block. Bar plot bars indicate how many times a feature extractor is meaningful for classification (.i.e having any of its features selected for usage in a fold) during the feature selection block. Only features extractors which had any meaningful features for classification are displayed.

TABLE 3. Experimental results.

Experiment	Model	ACC	SEN	SPE
M vs Nevus	TL-naive	68.00	21.43	86.11
M vs Nevus	TL (our)	66.00	35.71	77.78
M vs Nevus	MIL	72.00	78.57	69.44
M vs Nevus	Proposed Ensemble	**84.00**	**71.43**	**88.89**
M vs All	TL-naive	73.47	14.29	83.33
M vs All	TL (our)	71.43	21.43	79.76
M vs All	MIL	51.02	71.43	47.62
M vs All	Proposed Ensemble	**90.82**	**78.57**	**92.86**

naive: as in [5], without explicit uncertainty.

C. RESULTS

In this subsection, the results are presented in terms of percentage of classification accuracy (*ACC*), specificity (*SPE*), and sensitivity (*SEN*), where *SEN* represents the successful melanoma identification rate and *SPE* the successful identification of the other class.

Table 3 shows the achieved results. As detailed in Section III, the proposed ensemble model is comprised of two processes: TL and MIL – respectively, a 2D-only and a 3D-only classifier. The TL process includes alterations to enable the classification uncertainty to be determined in a non-naive manner. Therefore, the TL model without the mentioned uncertainty calculations is referred to as "TL-naive", which corresponds to the effective implementation of the state-of-the-art method described in [5], mentioned in Section III-C – providing the literature baseline classification result in both experiments. This method was selected from the literature for its superior results across several datasets. While it provides poor results in this dataset, it is important to remember that no adaptations were made from the original paper implementation description (meaning that this is an expected behaviour given this work's data constraints). The *ACC* performance of the TL-naive is of 68.00% and 73.47% for the *M vs Nevus* and *M vs All* experiments, respectively. While the accuracy increases in the experiment with more data (which has 48 additional samples in comparison with *MvsN*), it is important to point-out that the *SEN* metric decreases by 7.14 percentage points (*pp*), even though the number of melanoma samples is the same (14) in both experiments. This decrease represents one melanoma misclassification. The *SPE* metric is not comparable between both experiments, since the amount of samples differs between experiments. Across folds, TL-naive identifies 31 out of 36 nevus in the first experiment, and 70 out of 84 non-melanoma lesions in the second experiment.

In the *M vs Nevus* experiment, TL-naive incorrectly classifies 16 samples in the testing stage. Performing naive uncertainty calculations with the TL-naive model (using the internal Softmax probabilities) enables the identification of nine potential misclassifications. Among these, only six are actual misclassifications, while three were originally correct. If the uncertainty-awareness is performed during training, the (TL) model incorrectly classifies 17 samples in the testing stage, but enables the correct identification of 11 (instead of six) of these misclassifications (while also incorrectly identifying one sample that was actually correct). This improvement to the TL uncertainty identification enables MIL (the 3D-only classifier), which has a 72.00% *ACC* and 78.57% *SEN*, to potentially correct or disregard said misclassifications performed while observing 2D-only information (as described in the proposed ensemble pipeline, Section III-A). From the TL-uncertain-classifications (which uses only colour information), MIL corrects 10 out of 11 misclassifications (of which, five are melanomas) and only wrongly changes one sample that was originally correctly identified, although with low certainty – improving from the TL initial performance from 66.00% (2D-only) to 84.00% *ACC* (2D and 3D), as shown in Table 3 for the "Proposed Ensemble".

In *M vs All*, the detailed observations are similar to the previous experiment. The TL-naive incorrectly classifies 26 samples in the testing stage from which the naive uncertainty calculations enable the identification of 12 potential misclassifications – 10 comprising actual misclassifications and two originally correct. If trained with the uncertainty-awareness, the (TL) model incorrectly classifies 28 samples – but potentially enables the correct identification of 21 (instead of 10) misclassifications (while also incorrectly identifying three samples which were actually correct). As with the previously detailed-experiment results, this improvement to the TL uncertainty identification enables MIL, which has a 51.00% *ACC* and 71.43% *SEN*, to potentially correct or disregard the misclassifications. From the TL-uncertain-classifications, MIL corrects 20 out of the 21 misclassifications (of which, nine are melanomas) and incorrectly classifies one of the three uncertain (but correctly classified) samples – improving from the TL initial performance from 71.43% to 90.82% *ACC*, as shown in Table 3 for the "Proposed Ensemble".

In this section, all comparisons with the baseline classification results obtained with TL-naive have shown that the proposed ensemble method provides superior performance results. This can be interpreted as an indirect comparison with the works considered in [5] and other works that resorted to the same dataset and metrics as [5]. In essence, since TL-naive [5] reports results superior to 10 other works, it serves as indication that the proposed ensemble method could prove superior to these previous mentioned methods. This can be further extended to other literature works (as [13], [21], [22], [37], [62]), which experiment on the same datasets as [5] using the same metrics.

As a reminder to the reader, the purpose of this manuscript is not to provide a new model with the highest literature results, but to evidence that, when RGB image classification is uncertain, a second-level classification using depth information might increase the overall performance of the skin lesion classification system. For this reason, previous author's works are not directly compare here, which, in any case, provide inferior accuracy performances for the *M vs All* experiment [15], [34], [35].

V. CONCLUSION AND FUTURE WORK

The pursuit of a solution to automatically identify melanoma has been under research for many years. Automated melanoma detection is crucial to help dermatologists improve their diagnostic accuracy. Still, even with Deep Learning methods, current systems are yet to achieve satisfactory sensitivity performances. Instead of continuously attempting to improve algorithms with available colour (2D) datasets, which are commonly used by dermatology experts, new dimensions and modalities should be explored as, for example, surface (3D) information; which can potentially provide new melanoma discrimination capabilities. In order to advance beyond current state-of-the-art results, more reliable solutions might depend on the joint exploitation of both 2D and 3D information. Taking advantage of the recently introduced technology of light-field cameras, the main contribution of this work is to be the first to exploit both colour and depth information for classification of skin lesions using a recent dataset of multi-dimensional imaging, which was specifically acquired for this goal and has shown the ability to provide rich information for image classification. Accordingly, this work groups different literature domains, even if originally developed for different purposes, aiming to build a model that takes advantage of the recent literature improvements in both 2D and 3D modalities. As a result, this work is the first to incorporate Deep Learning uncertainty evaluation mechanisms with Multiple Instance Learning for the training of a robust synergistic ensemble classifier with the intent of performing skin lesion classification using light-field imagery.

Targeting the melanoma class with this model, despite the large class imbalance (often present in medical image datasets) and limited data samples, the ensemble model achieves a cross-validation accuracy of 84.00%, with 71.43% sensitivity and 88.89% specificity. These results account for the classification against nevus lesions and show an accuracy increase of 16.00*pp* (supported by a sensitivity increase of 50.00*pp*) from the baseline method (applied to the SKINL2 dataset). In a more challenging setting, discrimination of melanomas against all other available skin lesions was achieved with 90.82% accuracy, 78.57% sensitivity, and 92.86% specificity, with a similar accuracy increase of 17.35*pp* from the baseline, also supported by a sensitivity increase of 64.28*pp*. The performed experimental assessment allows to extrapolate that melanoma skin lesion classification can be improved by including unexploited 3D information, such as surface depth. This claim is supported by the different constraints employed in the experiment that aimed at increasing the confidence in the attained results although operating on a small dataset.

Expanding on the presented concepts, further research can be done in the field of skin lesion image classification to either improve existing methods that lack in performance or refine existing top performers, as shown in this research. Thus, future works should try to enlarging existing datasets and acquisition modalities to enable the emergence of features specifically tailored for skin lesion detection and classification. In the presence of untrustworthy 2D features, the achieved results indicate that the 3D surface provides redeeming results, showing that improvement of existing methods is still possible when looking beyond 2D image characteristics.

REFERENCES

[1] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artif. Intell. Med.*, vol. 56, no. 2, pp. 69–90, Oct. 2012.

[2] C. Karimkhani, A. C. Green, T. Nijsten, M. A. Weinstock, R. P. Dellavalle, M. Naghavi, and C. Fitzmaurice, "The global burden of melanoma: Results from the global burden of disease study 2015," *Brit. J. Dermatol.*, vol. 177, no. 1, pp. 134–140, Jul. 2017.

[3] S. McDonagh, R. Fisher, and J. Rees, "Using 3D information for classification of non-melanoma skin lesions," in *Medical Image Understanding and Analysis*. Dundee, U.K., Jul. 2008, pp. 164–168.

[4] L. N. Smith, M. L. Smith, A. R. Farooq, J. Sun, Y. Ding, and R. Warr, "Machine vision 3D skin texture analysis for detection of melanoma," *Sensor Rev.*, vol. 31, no. 2, pp. 111–119, Mar. 2011.

[5] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with Alex-net," *PLoS ONE*, vol. 14, no. 5, pp. 1–17, May 2019.

[6] K. H. Cheong, K. J. W. Tang, X. Zhao, J. E. W. Koh, O. Faust, R. Gururajan, E. J. Ciaccio, V. Rajinikanth, and U. R. Acharya, "An automated skin melanoma detection system with melanoma-index based on entropy features," *Biocybernetics Biomed. Eng.*, vol. 41, no. 3, pp. 997–1012, Jul. 2021.

[7] A. Javaid, M. Sadiq, and F. Akram, "Skin cancer classification using image processing and machine learning," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Islamabad, Pakistan, Jan. 2021, pp. 439–444.

[8] S. Oukil, R. Kasmi, K. Mokrani, and B. García-Zapirain, "Automatic segmentation and melanoma detection based on color and texture features in dermoscopic images," *Skin Res. Technol.*, vol. 28, no. 2, pp. 203–211, Mar. 2022.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[10] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, Dec. 2016.

[11] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[12] Z. Li, X. Zhang, H. Müller, and S. Zhang, "Large-scale retrieval for medical image analytics: A comprehensive review," *Med. Image Anal.*, vol. 43, pp. 66–84, Jan. 2018.

[13] P. Tang, Q. Liang, X. Yan, S. Xiang, and D. Zhang, "GP-CNN-DTEL: Global—Part CNN model with data-transformed ensemble learning for skin lesion classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2870–2882, Oct. 2020.

[14] I. Gonzalez-Diaz, "DermaKNet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 547–559, Feb. 2018.

[15] P. Pereira, L. Thomaz, L. Tavora, P. Assuncao, R. Fonseca-Pinto, R. Paiva, and S. Faria, "Melanoma classification using light-fields with Morlet scattering transform and CNN: Surface depth as a valuable tool to increase detection rate," *Med. Image Anal.*, vol. 75, pp. 1–16, Jan. 2022.

[16] X. Lu and Y. A. F. A. Zadeh, "Deep learning-based classification for melanoma detection using XceptionNet," *J. Healthcare Eng.*, vol. 2022, pp. 1–10, Mar. 2022.

[17] R. Kaur, H. GholamHosseini, R. Sinha, and M. Lindén, "Melanoma classification using a novel deep convolutional neural network with dermoscopic images," *Sensors*, vol. 22, no. 3, pp. 1–15, 2022.

[18] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Prague, Czech Republic, Jun. 2016, pp. 1397–1400.

[19] A. Menegola, M. Fornaciali, R. Pires, F. Bittencourt, S. Avila, and E. Valle, "Knowledge transfer for melanoma screening with deep learning," in *IEEE Int. Symp. Biomed. Imag.*, Melbourne, VIC, Australia, Apr. 2017, pp. 297–300.

[20] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[21] C. Barata, M. E. Celebi, and J. S. Marques, "A survey of feature extraction in dermoscopy image analysis of skin cancer," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1096–1109, Jun. 2018.

[22] J. R. Hagerty, R. J. Stanley, H. A. Almubarak, N. Lama, R. Kasmi, P. Guo, R. J. Drugge, H. S. Rabinovitz, M. Oliviero, and W. V. Stoecker, "Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images," *IEEE J. Biomed. Health*, vol. 23, no. 4, pp. 1385–1391, Jan. 2019.

[23] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014, *arXiv:1404.5997*.

[24] S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao, "Striking the right balance with uncertainty," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 103–112.

[25] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2018, pp. 3179–3189.

[26] T. DeVries and G. W. Taylor, "Leveraging uncertainty estimates for predicting segmentation quality," 2018, *arXiv:1807.00502*.

[27] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," 2020, *arXiv:2011.06225*.

[28] C. Cho, W. Choi, and T. Kim, "Leveraging uncertainties in softmax decision-making models for low-power IoT devices," *Sensors*, vol. 20, no. 16, pp. 1–32, Aug. 2020.

[29] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, Dec. 2017.

[30] J. D. Keeler, D. E. Rumelhart, and W. K. Leow, "Integrated segmentation and recognition of hand-printed numerals," in *Proc. Int. Conf. Neural Inf. Process Syst.*, Denver, CO, USA, Nov. 1990, pp. 557–563.

[31] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in *Proc. Int. Conf. Mach. Learn.*, Madison, WI, USA, vol. 98, Jul. 1998, pp. 341–349.

[32] V. Kecman, T. Huang, and M. Vogt, "Iterative single data algorithm for training kernel machines from huge data sets: Theory and performance," in *Support Vector Machines: Theory and Applications*. Springer, May 2005, pp. 255–274.

[33] X. Wang, F. Tang, H. Chen, L. Luo, Z. Tang, A.-R. Ran, C. Y. Cheung, and P.-A. Heng, "UD-MIL: Uncertainty-driven deep multiple instance learning for OCT image classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 12, pp. 3431–3442, Dec. 2020.

[34] P. M. M. Pereira, L. A. Thomaz, L. M. N. Tavora, P. A. A. Assuncao, R. Fonseca-Pinto, R. P. Paiva, and S. M. M. Faria, "Skin lesion classification using bag-of-3D-features," in *Proc. Telecoms Conf. (ConfTELE)*, Leiria, Portugal, Feb. 2021, pp. 1–6.

[35] P. M. M. Pereira, L. A. Thomaz, L. M. N. Tavora, P. A. A. Assuncao, R. Fonseca-Pinto, R. P. Paiva, and S. M. M. Faria, "Skin lesion classification using features of 3D border lines," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2726–2731.

[36] F. Navarro, M. Escudero-Vinolo, and J. Bescos, "Accurate segmentation and registration of skin lesion images to evaluate lesion change," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 501–508, Apr. 2018.

[37] S. Pathan, K. G. Prabhu, and P. C. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review," *Biomed. Signal Process. Control*, vol. 39, pp. 237–262, Jan. 2018.

[38] 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 *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Osaka, Japan, Jul. 2013, pp. 5437–5440.

[39] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*. Dordrecht, The Netherlands: Springer, vol. 6. 2013, pp. 63–86.

[40] G. Argenziano, H. Soyer, V. De Giorgi, D. Piccolo, P. Carli, and M. Delfino, *Interactive Atlas of Dermoscopy (Book and CD-ROM)*. Milan, Italy: EDRA Medical Publishing & New Media, 2000.

[41] I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov, "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6578–6585, Nov. 2015.

[42] T. Y. Satheesha, D. Satyanarayana, M. N. G. Prasad, and K. D. Dhruve, "Melanoma is skin deep: A 3D reconstruction technique for computerized dermoscopic skin lesion classification," *IEEE J. Translational Eng. Health Med.*, vol. 5, pp. 1–17, Jan. 2017.

[43] S. Faria, J. Filipe, P. Pereira, L. Tavora, P. Assuncao, M. Santos, R. Fonseca-Pinto, F. Santiago, V. Dominguez, and M. Henrique, "Light field image dataset of skin lesions," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Berlin, Germany, Jul. 2019, pp. 3905–3908.

[44] S. S. Mahmouei, M. Aldeen, W. V. Stoecker, and R. Garnavi, "Biologically inspired QuadTree color detection in dermoscopy images of melanoma," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 570–577, Mar. 2019.

[45] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Point feature extraction on 3D range scans taking into account object boundaries," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Shanghai, China, May 2011, pp. 2601–2608.

[46] Z. Marton, D. Pangercic, N. Blodow, J. Kleinehellefort, and M. Beetz, "General 3D modelling of novel objects from a single view," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Taipei, Taiwan, Dec. 2010, pp. 3700–3705.

[47] A. Kanezaki, Z. Marton, D. Pangercic, T. Harada, Y. Kuniyoshi, and M. Beetz, "Voxelized shape and color histograms for RGB-D," in *Proc. IROS Workshop Act. Semantic Perception*, San Francisco, CA, USA, Sep. 2011, pp. 1–6.

[48] J. P. S. do Monte Lima and V. Teichrieb, "An efficient global point cloud descriptor for object recognition and pose estimation," in *Proc. Conf. Graph., Patterns Images*, Sao Paulo, Brazil, Oct. 2016, pp. 56–63.

[49] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1265–1278, Jun. 2005.

[50] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst*, Nice, France, Sep. 2008, pp. 3384–3391.

[51] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. 26th IEEE Int. Conf. Robot. Autom.*, Kobe, Japan, May 2009, pp. 3212–3217.

[52] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. 11th Eur. Conf. Comput. Vis.*, Heraklion, Crete, Sep. 2010, pp. 356–369.

[53] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3D object classification," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Karon Beach, Phuket, Dec. 2011, pp. 2987–2992.

[54] A. Frome, D. Huber, R. Kolluri, T. Bülow, and J. Malik, "Recognizing objects in range data using regional point descriptors," in *Proc. Eur. Conf. Comput. Vis.*, Prague, Czech Republic, May 2004, pp. 224–237.

[55] F. Tombari, S. Salti, and L. Di Stefano, "Unique shape context for 3D data description," in *Proc. ACM Workshop 3D Object Retr. (3DOR)*, Florence, Italy, 2010, pp. 57–62.

[56] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, pp. 161–168, Jan. 2012.

[57] A. A. Jiménez, F. P. G. Márquez, V. B. Moraleda, and C. Q. G. Muñoz, "Linear and nonlinear features and machine learning for wind turbine blade ice detection and diagnosis," *Renew. Energy*, vol. 132, pp. 1034–1048, Mar. 2019.

[58] T. Akram, M. A. Khan, M. Sharif, and M. Yasmin, "Skin lesion segmentation and recognition using multichannel saliency estimation and M-SVM on selected serially fused features," *J. Ambient Intell. Hum. Comput.*, pp. 1–20, Sep. 2018.

[59] F. K. G. A., T. Akram, B. Laurent, S. R. Naqvi, M. M. Alex, and N. Muhammad, "A deep heterogeneous feature fusion approach for automatic land-use classification," *Inf. Sci.*, vol. 467, pp. 199–218, Oct. 2018.

[60] Y. Li, J. Sun, C. Tang, and H. Shum, "Lazy snapping," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 303–308, Aug. 2004.

[61] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[62] P. M. M. Pereira, R. Fonseca-Pinto, R. P. Paiva, P. A. A. Assuncao, L. M. N. Tavora, L. A. Thomaz, and S. M. M. Faria, "Skin lesion classification enhancement using border-line features—The melanoma vs nevus problem," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101765.

[63] M. A. Gelbart, J. Snoek, and R. P. Adams, "Bayesian optimization with unknown constraints," 2014, *arXiv:1403.5607*.

[64] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2879–2904, Oct. 2011.

[65] S. Faria, M. Santos, P. Assuncao, L. Tavora, L. Thomaz, P. Pereira, R. Fonseca-Pinto, F. Santiago, V. Dominguez, and M. Henrique, "Dermatological imaging using a focused plenoptic camera: The SKINL2 light field dataset," in *Proc. Conf. Telecommun.*, Lisbon, Portugal, Jun. 2019, pp. 1–4.

PEDRO M. M. PEREIRA was born in Lisbon, Portugal, in 1992. He received the B.Sc. and M.Sc. degrees in informatics engineering from the Polytechnic of Leiria, in 2014 and 2016, respectively, and the Ph.D. degree from the University of Coimbra, in 2022. He has been a Researcher with the Instituto de Telecomunicações, Portugal, since 2014. His research interests include high performance computing, image classification, artificial intelligence, and pattern recognition.

LUCAS A. THOMAZ (Member, IEEE) was born in Niterói, Brazil. He received the B.Sc. degree *(cum laude)* in electronic and computer engineering from the Universidade Federal do Rio de Janeiro (UFRJ), Brazil, in 2013, and the M.Sc. and D.Sc. degrees in electrical engineering from COPPE/UFRJ, in 2015 and 2018, respectively.

From 2017 to 2018, he was a Visiting Researcher Scholar with North Carolina State University. Since 2019, he has been a Researcher at the Instituto de Telecomunicações and an Associate Professor with the School of Technology and Management, Polytechnic of Leiria, Portugal. His research interests include computer vision, digital signal processing, video and image processing, image and video codification, and medical image analysis and coding. He is also involved in the JPEG standardization activities, serves as a member of the Student Services Committee. He is the Chair of the Engagement and Career Training Subcommittee of IEEE Signal Processing Society.

LUIS M. N. TAVORA was born in Lisbon, in 1970. He received the bachelor's degree in engineering physics from the University of Coimbra, Portugal, in 1993, and the Ph.D. degree in physics from the University of Surrey, U.K., in 1998. He is currently a Professor with the Department of Electrical Engineering, Polytechnic of Leiria, and a Researcher at the Instituto de Telecomunicações, Portugal. He has a solid research experience on X-ray-based imaging systems for medical and security applications and more recently on image processing and coding. His current research interests include machine learning, image processing and coding, and medical imaging.

PEDRO A. A. ASSUNCAO (Senior Member, IEEE) was born in Coimbra, Portugal, in 1965. He received the bachelor's and master's degrees in electrical engineering from the University of Coimbra, in 1988 and 1993, respectively, and the Ph.D. degree in electronic systems engineering from the University of Essex, U.K., in 1998.

He is currently a Full Professor in electronics and multimedia communications at the Polytechnic of Leiria and a Senior Researcher at the Instituto de Telecomunicações, Portugal. He has authored/coauthored over 200 publications in international conferences, journals, book chapters, two books, and four U.S. patents. His current research interests include learning-based coding and processing of image/video, light fields, machine-learning methods for multimedia systems, green computing for video coding, and lossless coding for multimodal medical imaging. He received the title of Agregado from IST, University of Lisbon, Portugal, in 2020.

RUI FONSECA-PINTO was born in Resende, Portugal, in 1977. He received the Ph.D. degree in biomedical engineering and biophysics from the University of Lisbon and the Medical Doctor degree.

He is currently a Professor with the Polytechnic of Leiria, Portugal. He has authored more than 100 scientific publications, resisted six brands and one international patent, and is a member of one spin off. His research interests include the application of decision support techniques in medicine, in particular, in the evaluation of the autonomic nervous systems and medical image processing.

RUI PEDRO PAIVA was born in Luanda, Angola, in 1973. He received the bachelor's, master's, and Ph.D. degrees from the University of Coimbra, in 2007, 1999, and 1996, respectively. He is currently a Professor at the Department of Informatics Engineering, University of Coimbra. His research interests include the areas of health informatics and music information retrieval. The common research hat is the study of feature engineering, machine learning, and signal processing to the analysis of musical and bio signals. He is also a member of the CMS Group, CISUC.

SERGIO M. M. FARIA (Senior Member, IEEE) was born in Horta, Portugal, in 1965. He received the Ph.D. degree in electronics and telecommunications from the University of Essex, U.K., in 1996. He has been a Full Professor with the Department of Electrical Engineering, Polytechnic of Leiria, since 2016. He was the Former Vice-President of the Polytechnic of Leiria, from 1997 to 1999. He was the Head of the IT Leiria Site, in 2014, and the Former Head of the Multimedia Signal Processing Group, Leiria. His research interests include 2D/3D image and video processing and coding and medical imaging. He is the co-inventor of two patents, co-founded a Tech startup, coauthor of one book, 14 book chapters, and more than 200 publications in international journals and conferences. He participates in the MPEG and JPEG activities and is currently an Associate Editor of IEEE Transactions on Image Processing and an Associate Editor of *Signal Processing: Image Communication* (Elsevier). He received the title of Agregado from IST, University of Lisbon, in 2014.