DAILY QA
Finished before your first cup of coffee

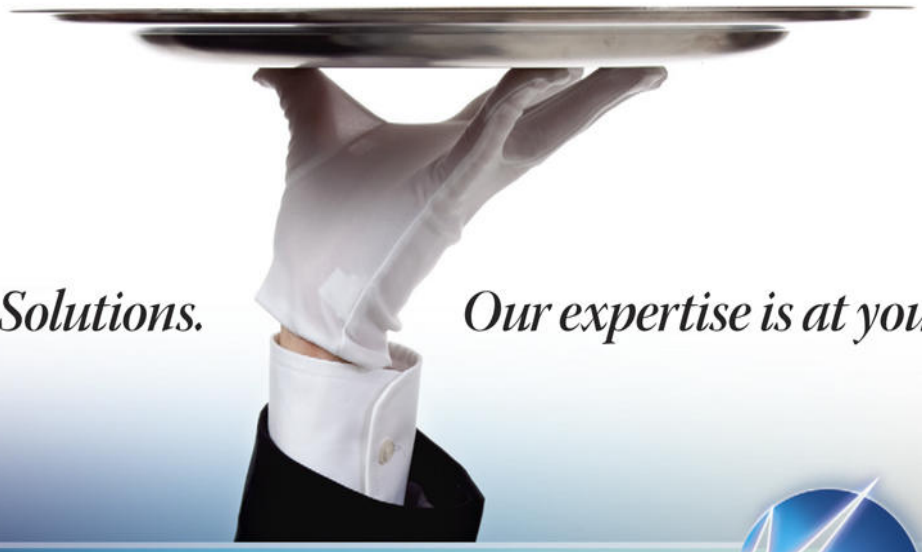
MONTHLY QA
Never re-learn workflow again

ANNUAL QA
Confidence with fewer gray hairs

STEREOTACTIC
When millimeters matter most

ABSOLUTE DOSIMETRY
"Pinpoint" is our margin of error

PATIENT DOSIMETRY
Pre-treatment is just the beginning



QA Solutions. *Our expertise is at your service.*

STANDARD IMAGING®

MORE THAN JUST QA PRODUCTS, FIND QA SOLUTIONS AT

www.standardimaging.com

1411-00, 2/18

Fully Automated Identification of Skin Morphology in Raster-Scan Optoacoustic Mesoscopy using Artificial Intelligence

Serafeim Moustakidis [a)]
AIDEAS OÜ, Narva mnt 5, Tallinn, Harju maakond, 10117, Estonia

Murad Omar, Juan Aguirre, Pouyan Mohajerani, and Vasilis Ntziachristos
Technische Universität München and Institute of Biological and Medical Imaging, Helmholtz Zentrum München, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

Purpose: Identification of morphological characteristics of skin lesions is of vital importance in diagnosing diseases with dermatological manifestations. This task is often performed manually or in an automated way based on intensity level. Recently, ultra-broadband raster-scan optoacoustic mesoscopy (UWB-RSOM) was developed to offer unique cross-sectional optical imaging of the skin. A machine learning (ML) approach is proposed here to enable, for the first time, automated identification of skin layers in UWB-RSOM data.
Methods and materials: The proposed method, termed SkinSeg, was applied to coronal UWB-RSOM images obtained from 12 human participants. SkinSeg is a multi-step methodology that integrates data processing and transformation, feature extraction, feature selection and classification. Various image features and learning models were tested for their suitability at discriminating skin layers including traditional machine learning along with more advanced deep learning algorithms. An SVM-based post-processing approach was finally applied to further improve the classification outputs.
Results: Random forest proved to be the most effective technique, achieving mean classification accuracy of 86.89% evaluated based on a repeated leave-one-out strategy. Insights about the features extracted and their effect on classification accuracy are provided. The highest accuracy was achieved using a small group of 4 features and remained at the same level or was even slightly decreased when more features were included. Convolutional neural networks provided also promising results at a level of approximately 85%. The application of the proposed post-processing technique was proved to be effective in terms of both testing accuracy and 3D visualization of classification maps.
Conclusions: SkinSeg demonstrated unique potential in identifying skin layers. The proposed method may facilitate clinical evaluation, monitoring and diagnosis of diseases linked to skin inflammation,

1. INTRODUCTION

Raster scan optoacoustic mesoscopy (RSOM) is evolving as a powerful alternative for non-invasive, high-resolution three-dimensional imaging of skin features based on optical absorption contrast. The technique can resolve epidermal and dermal features, including microvasculature, at resolution-to-depth ratios that go beyond optical coherence tomography (OCT)[1]. For example, OCT in the visible range[2,3] can image vascular networks non-invasively to depths of only ~400 μm[4-6]. As an alternative, high-frequency ultrasound can resolve microvasculature to depths of several millimeters. However, visualizing vessels with diameters smaller than 100 μm using this method requires microbubbles as contrast agents[7], which makes it challenging to apply in humans.

RSOM offers advantages of non-invasiveness and penetration depth over all these methods. For best performance, RSOM should be performed using ultra-wideband (UWB) detection, spanning a range of 200 MHz[8]. UWB-RSOM can generate high-resolution images of neovascularization in a growing tumor [9], visualize neovascularization in neoplastic gastrointestinal tissues [10], and observe clinically relevant features of the skin microvascular structures[11-12]. For this imaging technique to be clinically relevant, it is necessary to accurately define the regions and subregions of tissue in the field of view[10,13]. For example, skin images should be annotated to indicate the boundaries of epidermis, dermis, and, within the dermis, the areas that have a dense microvascular structure (herein referred to as the vascular plexus), since identifying particular features in each of these subregions may facilitate disease diagnosis and assessment of its severity[8].

So far, skin layers in UWB-RSOM images have been manually segmented by visual inspection of vasculature morphology or automatically based on signal intensity levels. Such procedures are slow or inaccurate and unsuitable for processing larger numbers of patients or for making clinical decisions during the patient's visit. Manual segmentation is also subjective and compromises the reproducibility and robustness of UWB-RSOM as a clinical tool.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.13725

Machine learning (ML) has ushered in new possibilities for automated image processing and diagnostic imaging[14-16]. Intelligent algorithms are expected to speed up clinical workflows and improve diagnostic accuracy and sensitivity. They can identify risk factors in medical images and provide a basis for quantitative, less subjective decision-making. ML has yet to be applied to the segmentation of UWB-RSOM images.

Here we examined the feasibility of using ML to identify skin layers automatically in cross-sectional UWB-RSOM images. We developed a multi-step method, termed SkinSeg, that integrates deterministic feature engineering and ML algorithms. Feature engineering is a fundamental term widely used in ML and refers to the process of using domain knowledge of the data to create features. The feature engineering part of SkinSeg involves data processing and transformation, feature extraction and feature selection. Various image features and ML models were tested for their suitability at discriminating skin layers. Deep learning was also investigated in two different ways: (i) well known pre-trained models were employed for extraction of deep features from the collected UWB-RSOM images and (2) CNN models were also trained to classify skin layers directly using the collected images as inputs skipping the feature engineering part of the methodology . The best performing classifier was selected, and a post-processing approach was finally applied to further improve the classification outputs.

The main contributions of this paper can be summarized as follows. The paper contains original content in the first ever application of learning empowered solutions on a very promising imaging technology (UWB-RSOM) in the task of skin morphology identification, facilitating extraction of new quantitative means for measuring skin and disease landmarks. Traditional machine learning approaches along with more advanced deep learning ones were employed to implement the classification task, whereas a novel post-classification technique was proposed being tailor made on the particular skin layer identification problem.

The rest of the paper is organized as follows. In Section II, we present the SkinSeg methodology. Section III describes the classification results achieved when various classification models, feature selection and post-processing were used. Section IV discusses the results, the significance of the extracted features and their effect on the classification results.

2. MATERIALS AND METHODS

2.A. Imaging System

We used an RSOM system developed in-house[8] (Figure 1) with a custom-made, spherically focused ultrasound detector (50 MHz central frequency, 10-120 MHz bandwidth, 3 mm focal distance, and f-number ~1)[1,8]. The collected signals were amplified by a low-noise amplifier (63 dB, AU-1291; Miteq, Hauppauge, NY, USA), then sampled by a high-speed data acquisition card (1 GS/s; CS121G2, Dynamic Signals, Lockport, IL, USA). The region of interest was illuminated by a fast nanosecond laser at 532 nm (0.9 ns pulse width, 2 kHz maximum pulse repetition frequency, and 1 mJ max pulse energy; HB532, Bright Solutions, Cura Carpignano, Italy), which was coupled to the scanning head through a customized fiber bundle. To scan the region of interest, the detector and fiber bundle were attached to a motorized *x*-*y* stage. The motorized stages scanned a region of 4×2 mm^2 in step sizes of 7.5 µm along the *x*-axis and 15 µm along the *y*-axis. These step sizes were chosen to satisfy the Nyquist criterion, maximize the signal to noise ratio, and minimize the scanning time. The scanning head was attached to an articulated arm to facilitate accurate positioning over the region of interest.

2.B. Data collection, reconstruction and labelling

A total of 12 participants (9 males and 3 females, mean age, 58 yr; range, 27-88 yr) were imaged using UWB-RSOM at two locations in the lower anterior shin area. Signals were corrected for motion[17] and then filtered into two sub-bands (14-40 and 40-120 MHz), which were reconstructed separately using a delay-and-sum algorithm with a dynamic aperture[1]. Frequency banding prior to reconstruction can improve the signal-to-noise ratio of smaller structures that would otherwise be masked by larger structures[1]. Reconstruction was performed using voxels of $10 \times 10 \times 3$ µm^3. The reconstructed RSOM dataset for each measurement was a 3-D matrix of size $333 \times 171 \times N_z$ pixels, where N_z varied between 300 and 600. The skin surface was identified as described then flattened to facilitate visualization of the different layers[1-2]. Imaging depth varied with the perfusion state of the skin and dermis thickness at the imaging site.

The different coronal layers (*z*-slices in the *x*-*y* plane) of the RSOM dataset were assigned to one of 4 classes: C1, dead zone, where no biological tissue was present; C2, epidermis; C3, vascular plexus; and C4, deeper structures. The depth ranges for each class were defined manually based on visual observation. Specifically, the second author visually examined the 3D volumes and assigned segment boundaries, according to his extensive experience with RSOM data and the typical image textures of different skin segment. This class assignment was performed separately for low- and high-frequency data.

Each *x-y* slice of the 3D image volume was split into sub-images following a sliding-window approach. A window of 100 × 100 pixels was scanned laterally over each *x-y* plane (row of 3D image) with a step of 50 pixels, producing 10 partially overlapping sub-images of 100 × 100 pixels per row (Figure 2). The window size was a compromise to generate a sufficiently large number of samples while remaining large enough to provide sufficient textural information. In this way, we acquired 39,100 sub-images, each of which was assigned to classes C1-C4 (Table I).

Figure 3 shows indicative images reconstructed with only high- or low-frequency data, along with images reconstructed with all data, with each image assigned to one of the 4 classes. Figure 3 leads to several observations: (i) deep structures appear darker than structures from epidermis (C2) or vascular plexus (C3); (ii) a specific textural pattern of the epidermal layer is visible in panels a), d) and i) that make the class C2 more easily identifiable; and (iii) a thin layer of the dermal vasculature is visible in C3 images of vascular plexus. Images reconstructed with low-frequency data appear less informative than images reconstructed with high-frequency data. Best are images reconstructed using data from both frequency bands which reveal fine spatial details together with lower-resolution skin structures. Therefore, images reconstructed with both low- and high-frequency data were used to train the ML algorithm to identify and extract features.

2.C. ML methodology

The dataset of 39,100 sub-images was split into training and testing sets using a repeated leave-one patient-out mechanism. Two feature extraction approaches were followed: (a) Traditional feature engineering: In the training phase, 29 features were generated from each sub-image of the training dataset. Feature selection was applied on the obtained feature set to (i) identify the most informative features, (ii) reduce the complexity of the classification models and (iii) improve the performance of the classifiers. (b) Deep feature extraction: using pretrained CNN models. Various classification models were applied to both the entire feature set and the subset of selected features. The best model was selected based on overall classification accuracy and the confusion matrix. Convolutional neural networks were also investigated as an alternative deep learning approach. Subsequently, this best fitted model was used to predict the responses for the sub-images in the testing dataset. The classifier model was deployed using the selected features assigning each testing sub-image to one of the four classes (biological layers).

2.C.1 Feature extraction and selection

Traditional feature engineering: Three first-order statistical parameters (maximum, minimum and standard deviation) were calculated to estimate properties of individual pixel values from the low- and high-frequency bands. Four gray-level co-occurrence matrix (*GLCM*) features (contrast, correlation, energy and homogeneity)[18-19] were also computed from both frequency bands to capture textural information and characterize the spatial variations and relationships between voxels within an image. Using 2D fast wavelet transform (FWT)[20], a two-level wavelet decomposition was performed on the image samples. At the first level analysis, four new sub-images were formed containing coefficients for approximation (LL1), horizontal detail (LH1), vertical detail (HL1), and diagonal detail (HH1). At the second level, the approximation sub-image LL1 was further decomposed to coefficients for LL2, LH2, HL2 and HH2. Finally, total energy measures were calculated from the associated wavelet coefficients of LL2, LH1-2, HL1-2 and HH1-2, generating seven features per frequency band. This energy distribution from FWT provides a detailed description of the frequency content of an image[21-22]. Table II summarizes the 29 features extracted: 14 came from each frequency band, in addition to depth (row index).

The efficient feature selection method SVM-FuzCoC[23-24] was applied to these features to select an informative, non-redundant feature subset. This method can provide a reasonable trade-off between accuracy and computational complexity[23].

Feature extraction using pre-trained deep learning models: The pre-trained models ResNet50[25] and AlexNet[26] were also employed for extracting deep features directly from the sub-images. The 100×100 pixels sub-images were resized to match the input size requirements of the deep learning models (224×224 for ResNet50 and 227×227 for AlexNet). AlexNet is a convolutional neural network that is trained on more than a million images from the ImageNet database. It consists of a total of 23 layers, whereas 3 of them are fully connected layers represented as FC6, FC7, and FC8 consisting of 4096, 4096 and 1000 features, respectively. These fully connected layers learn higher level image features and are better suited for image recognition tasks. FC6 and FC8 were used in SkinSeg for feature extraction. Due to the large number of extracted features, principal component analysis (PCA) was applied to reduce feature dimensionality. The first ten principal components were finally selected. ResNet50 is a 50-layer trained on the entire ImageNet 2012 classification dataset. The last fully connected layer of the network (FC1000) was used to extract 1000 deep features from each subimage. Similarly, PCA was also employed to reduce feature dimensionality to 10 features per sub-image. Overall, the following three configurations were employed: (i) Pre-trained AlexNet (using fully connected layer fc6)

followed by PCA. (ii) Pre-trained AlexNet (using fully connected layer fc8) followed by PCA. (iii) Pre-trained ResNet (using fully connected layer fc1000) followed by PCA.

2.C.2 Classification

Several algorithms were evaluated for classification of sub-images to specific biological layers. Given the relatively small dimensionality of the task (29 features), we tested linear discriminant analysis (LDA)[27] to provide a baseline for comparisons with more advanced models. We also evaluated decision trees[28-29] driven by Gini's diversity index, KNN (k=1) and weighted KNN[30] (k=10), as well as linear and non-linear support vector machines (SVM) algorithms[31-32], which can deal with the curse of dimensionality that typically appears in high-dimensional spaces. The ensemble techniques AdaBoost[33], Random Forest[34] and RUS-boost[35] were also evaluated using LDA models as weak learners. A neural network[36] with 3 hidden layers was evaluated, and the number of nodes per layer was varied. For this neural network, rectified linear units (ReLU) were chosen as the optimum activation functions, and the adaptive moment estimation (Adam) optimizer was selected for its fast convergence rate. Adaptive learning rate[37] was adopted here with a frequency step of 10 epochs. Batch size was 512.

The entire dataset was split into two subsets: a training/validation dataset (comprising data generated by 10 subjects) and a testing dataset (data from the remaining 2 subjects):

- Validation (29100 samples): A repeated leave-one-out cross validation (LOOCV) schedule was implemented on the first dataset to validate the classifiers employed and select the optimal hyperparameters. Specifically, the entire data obtained from a participant was held out for evaluation, whereas the remaining datasets from the rest 9 participants were used for training. This process was repeated 10 times (once for each one of the 10 participants of the training/validation set) leading to 10 non-overlapping holdout testing sets, and the resulting accuracies were averaged. The LOOCV mechanism was iterated through a wide range of hyperparameter combinations and the set of parameters for which LOOCV reports the highest accuracy was finally selected per model. The selected hyperparameters of all ML models are given in Table IV.
- Testing (10000 samples): The finetuned classification models (using the selected hyperparameters) were finally validated on the testing dataset.

Deep learning was also employed as an alternative approach to the problem of skin layers recognition. Convolutional Neural Networks[36] were trained directly on the extracted sub-images skipping the feature engineering part of SkinSeg. Different networks configurations were tested with respect to their classification capability. Due to CNN increased complexity, LOOCV was just performed once splitting the training/validation dataset into two sets: the training set (subimages generated from 9 participants) where the CNN models were applied and the validation set (subimages generated from 1 participant). In total, 65 network configurations were investigated and the one that maximized the validation accuracy was selected. Table III presents the characteristics and hyperparameters of the best CNN architecture that was finally selected. The selected CNN model was tested on the same testing dataset as described above.

The ML algorithms were implemented on a CPU using MATLAB 2017b (Version 9.3). Deep neural networks were developed using the MATLAB-based deep learning framework LightNet[38].

2.C.3 Post-processing

Given that the class ordering is a-priori known and that the four skin layers are non-overlapping, an SVM-based post-processing algorithm was applied to further improve the classification outputs of the best model. The proposed algorithm is presented below:

Step 0: *The outputs (labels) of the best classifier on the testing data are considered as inputs of the post-processing algorithm.*

Step 1*: Find the most frequent label per row of the 3D image and generate a label vector* $L=\{l_i\}$*, where* l_i *the most frequent label at row i,* $i=1,\ldots,N_z$ *and* N_z*: number of rows in the 3D image.*

Step 2*: For* $c = 1\ldots (N_c - 1)$*, where* N_c *is the number of classes*

- ***2.1*** *Gather labels belonging to classes c and c+1 and construct sets* $S_c=\{l_{i,c}\}$ *and* $S_{c+1}=\{l_{i,c+1}\}$*, where* $l_{i,c}$ *and* $l_{i,c+1}$ *are the labels from L belonging to classes c and c+1, respectively.*
- ***2.2*** *Train a linear SVM to separate sets* S_c *and* S_{c+1}
- ***2.3*** *Update those labels in L that belong to sets* S_c *and* S_{c+1} *according to the outputs of the SVM classifier,*

End of loop

Step 3*: Set* l_i *all the labels at the row i of the 3D image. Repeat for all rows of the 3D image.*

The proposed classification methodology produces one decision (label) per sample (sub image at x-y). Applying classification on all the samples generates a number of labels spatially distributed in the z axis (depth). However, we already know that skin layers are non-overlapping and the classes are ordered as follows: 1-2-3-4. The decisions (labels) obtained from the classifier typically overlap leading to mixed skin layer. Applying any linear classifier on the spatially distributed decisions (1-dimensional space) would lead to non-overlapping

classes. Non-linear classifiers were avoided since they would lead to more complex and overlapping class separations. Among the existing available linear classifiers, linear SVM was preferred given that: (i) it is a powerful classifier, one of the most effective, (ii) it provides a generalised class separation maximizing the margin between the adjacent classes and thus leading to the optimal linear separating hyperplane and (iii) it is fast when implemented in low dimensional spaces (1D here). Figure 4 shows a graphical example of the application of the proposed post-processing in a three-class classification problem. The algorithm was experimentally validated in Section III.

2.C.4 Visualization of classification outputs

Three-dimensional classification maps were generated from the classification decisions of the best model after the application of post-processing. A pixel-based approach was used in which a neighborhood (100 × 100 pixels) was considered around each testing pixel across the *x*-*y* plane (Figure 2), then the classification algorithm was applied to the extracted sub-image and a decision was made. The generated decision was finally updated (if needed) by the post-processing unit. The decision was expressed as an integer scalar that declares the biological structure/layer (class) assigned. Next, the value of the testing pixel was replaced with the class number, and a color was assigned to it. This process was repeated for all the pixels in the 3D image, and a 3D classification map was generated in which skin layers were depicted in different colors.

To simplify the computations, the testing data were sub-sampled by selecting 10 pixels per row of the 3D image, resulting in the extraction of 10 overlapping areas of 100 × 100 pixels per row. The proposed ML methodology was applied in the extracted sub-images generating 10 decisions per row in a layout of 5 × 2. The decisions were reshaped to the initial dimension (333 × 171) using bilinear interpolation based on the weighted average of pixels in the nearest 2-by-2 neighborhood. The final 3D classification map was obtained by repeating the process for each one of the N_z rows in the image cube.

3. RESULTS

3.A. Classification performance

Table IV compares the average performance of the classification models employed to recognize skin layers. The full set of 29 features was used to train and test the ML classification models. As far as the deep leaning approaches, 4096 deep features were extracted from the 'fc6' layer of AlexNet, where 1000 features were extracted from the 'fc8' layer of AlexNet and the 'fc1000' layer of ResNet50. The dimensionality was reduced by applying PCA. Random forest applied on the 29 manually extracted features proved to be the most effective technique, achieving classification performance of 86.13%. Linear discriminant models were used as the base classifier during ensemble learning. The optimal number of weak learners was determined to be 35 since it maximized the LOOCV validation accuracy. Statistical significance analysis was also performed by applying t-tests at two confidence levels (1% and 5%) on the accuracies obtained on the 12 LOOCV data folds. The results of RF without post-processing were significantly different (at both confidence levels) compared with all the remaining models. However, no significant differences at the confidence level of 5% were obtained on the results of RF, CNN and RUS-Boost followed by post-processing (third column of Table IV), with RF being only marginally better than CNN and RUS-Boost (at the confidence level of 1%).

Table V shows the confusion matrix as obtained by the best model of random forest. Despite the class imbalance problem, where classes 1, 2, 3 and 4 correspond to 10.15%, 18.23%, 52.68% and 18.92% of the entire dataset, adequately high per-class accuracies were observed from three out of the four classes (higher than 88% for classes 1, 3 and 4). Class 2 was recognized with the moderate accuracy of 66.28% whereas misclassifications occurred only between physically adjacent classes.

3.B. Influence of feature selection on classification

Next, we applied feature selection and repeated the classification using random forest. Figure 5 depicts the average classification performance with respect to the number of features selected for the best model with and without post-processing. The first four most important features (Table VI) alone gave an accuracy of more than 88% after post-processing, which remained at the same level or was even slightly decreased when more features were included. The methodology with post-processing outperformed the methodology without post-processing for all the different subsets investigated (Figure 5). The confusion matrixes for the best model trained on the four features without and with post-processing are given in Table VII and VIII, respectively.

Depth, as expressed in terms of the row index, was the feature most critical for achieving coarse separation of the layers, with an average accuracy of 70.2%. That depth would be the best feature is not surprising, given that the four layers are stacked one above the other in a fixed sequence. However, depth cannot guarantee fine layer separation, which can vary from one individual to the next. Thus, GLCM features in high- and low-frequency images were the second and fourth most effective features. The third most important feature was minimum value.

3.C. Visualization of classification outputs

Figure 6b and 6c show the 3D classification maps obtained from one subject's data using the SkinSeg method trained in the first four most important features without and with post-processing, respectively. Dead zone was depicted in purple; epidermis, blue; vascular plexus, green; and deeper structures, yellow. Figure 6a depicts a cross-sectional UWB-RSOM image rendered by taking the maximum intensity projection (MIPs) of the reconstructed images along the *y* direction.

4. DISCUSSION

UWB-RSOM shows unique potential for *in vivo* diagnostic imaging in dermatology. Here we demonstrate the potential of SkinSeg to reliably identify skin layers automatically from UWB-RSOM images. Our method may promote future development and clinical implementation of UWB-RSOM. The morphological characteristics of skin lesions are key elements for diagnosing skin diseases as well as systemic diseases that involve skin changes.

Various classification models were tested, and the best model was random forest, which achieved 86.13% classification accuracy. The confusion matrix revealed that misclassifications occurred only between adjacent classes (layers). Further analysis showed that a small group of 4 features can even achieve a better accuracy of over 88% (Table VIII, Figure 5). The 3D map (Figure 6c) and the cross-sectional UWB-RSOM image (Figure 6a) from the same subject showed good correlation between predicted and actual classes. The application of the proposed post-processing techniques was proved to be effective in terms of both testing accuracy (Tables VII and VIII) and 3D visualisation of classification maps (Figures 6b and 6c). CNNs were proved to be the second most promising classifier leading to a testing performance of 82.5% and 84.95% without and with post-classification, respectively. The relatively low performance of CNNs could be attributed to the number of subjects considered in our cohort (12 in total). More data is needed for a deep 19-layers network to generalize well compared with random forest that is an ensemble of weak learners having more relaxed requirements in terms of data size and variability. The accuracy and generalization of the proposed CNN is expected to increase with the inclusion of data generated from a larger subject cohort.

SkinSeg offers a faster alternative to current manual practices to assess skin morphology. Coronal images (taken in the *x*-*y* plane) proved informative, providing the necessary information content for implementing the classification task. The size of the coronal images (100×100 pixels) was large enough to contain enough textural information to reveal biological structures, but not large enough to make the method computationally burdensome. It may also be possible to use cross-sectional images (in the *z*-*x* plane) using segmentation analysis.

Two feature families (1st order and GLCM) significantly affected classification performance: features from those categories appeared among the first four most significant features. Two of the top four features came from the high-frequency image, which suggests that it is more informative for classification than the low-frequency image. Visual inspection of the sub-images (Figure 3) confirms the superiority of the high-frequency image in capturing class information. Nevertheless, GLCM correlation from low-frequency images also fell within the top five features for classification. Feature selection is a key part of the ML method here because it reduces the computational complexity of the classifier and accelerates feature extraction. Deep features extracted from pre-trained CNN models led to moderate testing accuracies and were not finally selected.

Figure 5 shows that classification accuracy increased only for the first four selected features, then remained almost stable until the thirteenth one and finally decreased with increasing features because of overfitting. We believe the marginal improvement in accuracy with increasing features reflects correlations among the 29 features.

One limitation of SkinSeg is that it requires flattening the 3D volume, such that the *z*-axis is approximately normal to the skin. This was performed automatically during the pre-processing stage. This flattening may be difficult if measurements are taken when the recording head is positioned quite obliquely to the skin surface.

Overall, the application of intelligent approaches is expected to revolutionise image-based diagnostics enabling new possibilities for enhanced monitoring and treatment. This paper makes a significant first step in applying ML empowered solutions on UWB-RSOM data in the task of skin morphology identification, facilitating extraction of new quantitative means for measuring skin and disease landmarks. Specifically, SkinSeg is expected to improve the accuracy of other ML-based methods which examine and analyze RSOM images for diagnostic purposes. Without a segmented description of the imaged volume, any diagnostic algorithm is bound to treat (and learn from) the entire volume as one piece. This condition could lead to problems such as overfitting. In return, the segmentation afforded by SkinSeg effectively reduces the input dimension and, hence, can potentially improve prediction accuracy.

The proposed methodology is envisioned to be employed in the processing pipeline of image and sensor systems based on the RSOM principles. As an example, we are using the proposed method to improve the results of a diabetic detection and grading system, which is currently under development. The results are

promising, and the accuracy of our method can be increased by using data from larger studies. Nevertheless, we should note that, since the method is going to be used as a component in a larger processing pipeline, its accuracy is to be understood and evaluated in conjunction with other components and the overall system in particular. Future steps include research into alternative image classification techniques such as autoencoders along with deep learning based segmentation techniques. The full feature space, as has been generated by a variety of pre-trained deep learning models, should be also exploited as an alternative approach. The performance of SkinSeg should be assessed in studies with more participants and in which both coronal images (*x*-*y* plane) and cross-sectional images (*x*-*z* plane) are used. Ultimately, ML-based UWB-RSOM should be applied to the clinic to examine the feasibility of automated disease diagnosis and assessment.

ACKNOWLEDGMENT

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement 694968 (PREMSOT) and 687866 (INNODERM).

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

[a] Author to whom correspondence should be addressed. Electronic mail: info@aideas.eu; Telephone: +30-6972863096

REFERENCES

[1] Omar M, Soliman D, Gateau J, Ntziachristos V. Ultrawideband reflection-mode optoacoustic mesoscopy. Opt Lett. 2014;39(13):3911. doi:10.1364/ol.39.003911
[2] Zabihian B, Weingast J, Liu M et al. In vivo dual-modality photoacoustic and optical coherence tomography imaging of human dermatological pathologies. Biomed Opt Express. 2015;6(9):3163. doi:10.1364/boe.6.003163
[3] Blatter C, Weingast J, Alex A et al. In situ structural and microangiographic assessment of human skin lesions with high-speed OCT. Biomed Opt Express. 2012;3(10):2636. doi:10.1364/boe.3.002636
[4] Qin J, Jiang J, An L, Gareau D, Wang R. In vivo volumetric imaging of microcirculation within human skin under psoriatic conditions using optical microangiography. Lasers Surg Med. 2011;43(2):122-129. doi:10.1002/lsm.20977
[5] Lacarrubba F, Pellacani G, Gurgone S, Verzì A, Micali G. Advances in non-invasive techniques as aids to the diagnosis and monitoring of therapeutic response in plaque psoriasis: a review. Int J Dermatol. 2015;54(6):626-634. doi:10.1111/ijd.12870
[6] Archid R. Confocal laser-scanning microscopy of capillaries in normal and psoriatic skin. J Biomed Opt. 2012;17(10):101511. doi:10.1117/1.jbo.17.10.101511
[7] Errico C, Pierre J, Pezet S et al. Ultrafast ultrasound localization microscopy for deep super-resolution vascular imaging. Nature. 2015;527(7579):499-502. doi:10.1038/nature16066
[8] Aguirre J, Schwarz M, Garzorz N et al. Precision assessment of label-free psoriasis biomarkers with ultra-broadband optoacoustic mesoscopy. Nat Biomed Eng. 2017;1(5):0068. doi:10.1038/s41551-017-0068
[9] Omar M, Schwarz M, Soliman D, Symvoulidis P, Ntziachristos V. Pushing the Optical Imaging Limits of Cancer with Multi-Frequency-Band Raster-Scan Optoacoustic Mesoscopy (RSOM). Neoplasia. 2015;17(2):208-214. doi:10.1016/j.neo.2014.12.010
[10] Knieling F, Gonzales Menezes J, Claussen J et al. Raster-Scanning Optoacoustic Mesoscopy for Gastrointestinal Imaging at High Resolution. Gastroenterology. 2018;154(4):807-809.e3. doi:10.1053/j.gastro.2017.11.285
[11] Aguirre J, Hindelang B, Berezhnoi A et al. Assessing nailfold microvascular structure with ultra-wideband raster-scan optoacoustic mesoscopy. Photoacoustics. 2018;10:31-37. doi:10.1016/j.pacs.2018.02.002
[12] Berezhnoi A, Schwarz M, Buehler A, Ovsepian S, Aguirre J, Ntziachristos V. Assessing hyperthermia-induced vasodilation in human skin in vivo using optoacoustic mesoscopy. J Biophotonics. 2018;11(11):e201700359. doi:10.1002/jbio.201700359
[13] Schwarz M, Omar M, Buehler A, Aguirre J, Ntziachristos V. Implications of Ultrasound Frequency in Optoacoustic Mesoscopy of the Skin. IEEE Trans Med Imaging. 2015;34(2):672-677. doi:10.1109/tmi.2014.2365239
[14] Lee J, Jun S, Cho Y et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol. 2017;18(4):570. doi:10.3348/kjr.2017.18.4.570
[15] Wu G, Shen D, Sabuncu M. Machine Learning In Medical Imaging. 1st ed. Elsevier, Academic Press; 2016:512.
[16] Razzak M, Naz S, Zaib A. Deep Learning for Medical Image Processing: Overview, Challenges and the Future. Lecture Notes in Computational Vision and Biomechanics. 2017:323-350. doi:10.1007/978-3-319-65981-7_12
[17] Schwarz M, Garzorz-Stark N, Eyerich K, Aguirre J, Ntziachristos V. Motion correction in optoacoustic mesoscopy. Sci Rep. 2017;7(1). doi:10.1038/s41598-017-11277-y
[18] Haralick R, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst Man Cybern. 1973;SMC-3(6):610-621. doi:10.1109/tsmc.1973.4309314
[19] Liao Y, Tsui P, Li C et al. Classification of scattering media within benign and malignant breast tumors based on ultrasound texture-feature-based and Nakagami-parameter images. Med Phys. 2011;38(4):2198-2207. doi:10.1118/1.3566064
[20] Strela V, Heller P, Strang G, Topiwala P, Heil C. The application of multiwavelet filterbanks to image processing. IEEE Transactions on Image Processing. 1999;8(4):548-563. doi:10.1109/83.753742
[21] Alobaidli S, McQuaid S, South C, Prakash V, Evans P, Nisbet A. The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning. Br J Radiol. 2014;87(1042):20140369. doi:10.1259/bjr.20140369
[22] Singh S, Urooj S. Wavelets: biomedical applications. Int J Biomed Eng Technol. 2015;19(1):1. doi:10.1504/ijbet.2015.071405

[23] Moustakidis S, Theocharis J. SVM-FuzCoC: A novel SVM-based feature selection method using a fuzzy complementary criterion. Pattern Recognit. 2010;43(11):3712-3729. doi:10.1016/j.patcog.2010.05.007
[24] Moustakidis S, Theocharis J, Giakas G. Feature selection based on a fuzzy complementary criterion: application to gait recognition using ground reaction forces. Comput Methods Biomech Biomed Engin. 2012;15(6):627-644. doi:10.1080/10255842.2011.554408
[25] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778. doi: 10.1109/CVPR.2016.90
[26] Krizhevsky A, Sutskever I, Hinton G. E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems. 2012.
[27] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed. Wiley Inter science, 2000.
[28] Belson W. Matching and Prediction on the Principle of Biological Classification. Applied Statistics. 1959;8(2):65. doi:10.2307/2985543
[29] Witten, I., Frank, E. & Hall, M. Data mining. (Morgan Kaufmann, 2011).
[30] Atkeson C, Moore A, Schaal S. Artificial Intelligence Review. 1997;11(1/5):11-73. doi:10.1023/a:1006559212014
[31] Cortes C, Vapnik V. Mach Learn. 1995;20(3):273-297. doi:10.1023/a:1022627411411
[32] Scholkopf B, Support Vector Learning, R. Oldenbourg Verlag, Munich, 1997
[33] Freund Y, Schapire R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci. 1997;55(1):119-139. doi:10.1006/jcss.1997.1504
[34] Breiman L. Mach Learn. 2001;45(1):5-32. doi:10.1023/a:1010933404324
[35] Seiffert C, Khoshgoftaar T, Van Hulse J, Napolitano A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 2010;40(1):185-197. doi:10.1109/tsmca.2009.2029559
[36] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444. doi:10.1038/nature14539
[37] Ruder S. (2016). An overview of gradient descent optimization algorithms. CoRR, abs/1609.04747.
[38] Ye C, Zhao C, Yang Y, Fermüller C, Aloimonos Y. LightNet: A Versatile, Standalone Matlab-based Environment for Deep Learning. Proceedings of the 2016 ACM on Multimedia Conference - MM '16. 2016. doi:10.1145/2964284.2973791

Figure captions

Figure. 1. Schematic of the data acquisition process and a sample image of the RSOM system: (a) RSOM raster-scans a single element detector in the x-y plane. (b) Example of a maximum intensity projection of a UWB-RSOM scan from one participant. The combination of low-frequency data (red channel) and high-frequency data (green channel) shows macro- and microvasculature (horizontal scale bar, 500 µm; vertical scale bar, 250 µm)

Figure 2. Schematic of the SkinSeg ML method. 2D images (100 x100 pixels) are extracted from each slice of the 3D image and fed into the ML algorithm. The classification models are trained on descriptive features extracted from each sample. Post-processing is applied on the generated testing labels.

Figure 3. Examples of skin feature classes resolved using UWB-RSOM. Representative images from one participant are shown after reconstruction using only low-frequency (LF) data (14-40 MHz, top row), high-frequency (HF) data (40-120 MHz, middle row), or data of all frequencies (RSOM, bottom row). The columns show images from epidermis, vascular plexus and deeper structures.

Figure 4. Example of the application of the proposed post-processing on a 3-class problem. The outputs of the classifier are visualized with respect to the vertical axes, where each class is represented by a different color. (a) Firstly, linear SVM is applied to separate class 1 from class 2, defining a separating hyperplane that re-assigns the labels of class 1 and 2 based on the maximum-margin criterion, resulting in the new class assignment shown in (b). The process is repeated on the updated labels by applying a second linear SVM to separate the classes 2 and 3, as shown in (c). The final output of the successive application of linear SVMs is the generation of three non-overlapping populations of labels, shown in (d).

Figure 5. Classification accuracy of the best model (Random Forest) as a function of the number of features selected with and without post-processing

Figure 6. Example of UWB-RSOM image classification using the SkinSeg method with feature selection. (a) Cross-sectional UWB-RSOM image from one subject's data, rendered by taking the maximum intensity projection along the y direction. 3D classification map of the same subject's data without post-processing (b) and with post-processing (c). Purple, dead zone; blue, epidermis; green, vascular plexus; yellow, deeper structures. The classification outputs are annotated on the image (a).

TABLE I. Number of sub-images per class

Class	Description	Number of samples per frequency band
C1	Dead zone	3970
C2	Epidermis	7130
C3	Vascular plexus	20600
C4	Deeper structures	7400
	Total	**39100**

TABLE II. Features extracted from each sub-image

ID	Description	Frequency band	Category
1	Maximum value	Low	1st order statistics
2	Minimum value	Low	1st order statistics
3	Standard deviation	Low	1st order statistics
4	Contrast	Low	GLCM
5	Correlation	Low	GLCM
6	Energy	Low	GLCM
7	Homogeneity	Low	GLCM
8	Energy of LL2	Low	Wavelet
9	Energy of LH1	Low	Wavelet
10	Energy of HL1	Low	Wavelet
11	Energy of HH2	Low	Wavelet
12	Energy of LH2	Low	Wavelet
13	Energy of HL2	Low	Wavelet
14	Energy of HH2	Low	Wavelet
15	Maximum value	High	1st order statistics
16	Minimum value	High	1st order statistics
17	Standard deviation	High	1st order statistics
18	Contrast	High	GLCM
19	Correlation	High	GLCM
20	Energy	High	GLCM
21	Homogeneity	High	GLCM
22	Energy of LL2	High	Wavelet
23	Energy of LH1	High	Wavelet
24	Energy of HL1	High	Wavelet
25	Energy of HH2	High	Wavelet
26	Energy of LH2	High	Wavelet
27	Energy of HL2	High	Wavelet
28	Energy of HH2	High	Wavelet
29	Row number		Depth Index

GLCM stands for grey-level co-occurrence matrix
LLk, LHk, HLk, HHk stand for approximation, horizontal, vertical and diagonal details at the k- level analysis of DWT, respectively

TABLE III. Selected CNN architecture

layer	Type	Description
1	Image Input	100x100x2 images with 'zerocenter' normalization
2	Convolution	8 5x5 convolutions with stride [1 1] and padding [1 1 1 1]
3	Batch Normalization	-
4	ReLU	-
5	Max Pooling	2x2 max pooling with stride [2 2] and padding [0 0 0 0]
6	Convolution	16 5x5 convolutions with stride [1 1] and padding [1 1 1 1]
7	Batch Normalization	-
8	ReLu	-
9	Max Pooling	2x2 max pooling with stride [2 2] and padding [0 0 0 0]
10	Convolution	32 5x5 convolutions with stride [1 1] and padding [1 1 1 1]
11	Batch Normalization	-
12	ReLu	-
13	Max Pooling	2x2 max pooling with stride [2 2] and padding [0 0 0 0]
14	Convolution	64 5x5 convolutions with stride [1 1] and padding [1 1 1 1]
15	Batch Normalization	-
16	ReLu	-
17	Fully Connected	4 Fully Connected layer
18	Softmax	-
19	Classification Output	Cross entropy

TABLE IV. Skin layer classification performance for different algorithms

	CLASSIFIER (HYPERPARAMETERS)	Accuracy (%)	Accuracy after post-processing (%)
Machine Learning	Decision trees (minimum leaf size: 5, Split criterion: Gini's index, Maximal number of decision splits: 7)	77.56 [a,b]	82.01 [a,b]
	LDA	76.42 [a, b]	80.15 [a, b]
	Linear SVM (C= 104)	67.98 [a, b]	74.46 [a, b]
	Non-linear Gaussian SVM (C=80, sigma =0.15)	71.56 [a, b]	78.70 [a, b]
	KNN1	76.15 [a, b]	80.16 [a, b]
	Weighted KNN-10	78.18 [a, b]	80.05 [a, b]
	AdaBoost (number of weak learners: 50, best weak learner: DT)	80.99 [a, b]	83.23 [a, b]
	Random Forest (number of weak learners: 35, best weak learner: LDA)	**85.66**	**86.13**
	RUS-Boost (number of weak learners: 24, best weak learner: DT)	81.55 [a, b]	84.81[a]
	Three-layer Neural Networks (Adam optimization, ReLU functions and adaptive learning rate)	79.11 [a, b]	82.42 [a, b]
Learning	CNN	82.50 [a, b]	84.95 [a]
	AlexNet + PCA + RF (fc6, 10 principal comp.)	78.44 [a, b]	81.20 [a, b]
	AlexNet + PCA + RF (fc8, 10 principal comp.)	69.76 [a, b]	74.23 [a, b]
	ResNet50 + PCA + RF (fc1000, 10 principal comp.)	68.48 [a, b]	73.12 [a, b]

a. Significantly different from random forest ($p < 0.05$) by applying t-tests on the LOOCV accuracies over the 12 data folds

b. Significantly different from random forest ($p < 0.01$) by applying t-tests on the LOOCV accuracies over the 12 data folds

TABLE V. Confusion matrix of the best model (Random Forest) with post-processing using the entire feature set

	Class 1	Class 2	Class 3	Class 4	Per-class accuracy (%)
Class 1	1651	149			91.72
Class 2	229	928	243		66.29
Class 3		191	2735	174	88.23
Class 4			401	3299	89.16
			Total		**86.13**

TABLE VI. First four most informative features selected

ID	Description	Frequency band	Explanation
29	Row number	-	Depth
19	Correlation/GLCM	High	Linear dependency between neighboring pixels in high-frequency data
16	Minimum value/1st order	High	Darkness of the area
5	Correlation/GLCM	Low	Linear dependency between neighboring pixels in low-frequency data

TABLE VII. Confusion matrix of the best model (Random Forest) without post-processing using the best four features

	Class 1	Class 2	Class 3	Class 4	Per class accuracy (%)
Class 1	1640	149	11		91.11
Class 2	232	922	246		65.86
Class 3	2	195	2709	194	87.39
Class 4		34	399	3267	88.30
	Total				**85.38**

TABLE VIII. Confusion matrix of the best model (Random Forest) with post-processing using the best four features

	Class 1	Class 2	Class 3	Class 4	Per-class accuracy (%)
Class 1	1689	111			93.83
Class 2	202	997	201		71.21
Class 3		135	2824	141	91.09
Class 4			399	3301	89.21
	Total				**88.11**