Biomedical Optics EXPRESS

Model learning analysis of 3D optoacoustic mesoscopy images for the classification of atopic dermatitis

SOJEONG PARK,[1,9] SHIER NEE SAW,[1,2,9] XIUTING LI,[3,9] MAHSA PAKNEZHAD,[1] DAVIDE COPPOLA,[1] U. S. DINISH,[3] AMALINA BINITE EBRAHIM ATTIA,[3] YIK WENG YEW,[4] STEVEN TIEN GUAN THNG,[4] HWEE KUAN LEE,[1,5,6,7,8,10] AND MALINI OLIVO[3,11]

[1] *Bioinformatics Institute, Agency of Science, Technology and Research, A*STAR, 30 Biopolis Street, #07-01 Matrix, 138671, Singapore*
[2] *Current address: Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia*
[3] *Laboratory of Bio-Optical Imaging, Singapore Bioimaging Consortium, A*STAR, 11 Biopolis Way, 138667, Singapore*
[4] *National Skin Centre, 1 Mandalay, 308205, Singapore*
[5] *School of Computing, National University of Singapore, 13 Computing Drive, Singapore, 117417, Singapore*
[6] *Singapore Eye Research Institute (SERI), 11 Third Hospital Ave, Singapore, 168751, Singapore*
[7] *Image and Pervasive Access Laboratory (IPAL), 1 Fusionopolis Way, #21-01 Connexis (South Tower), 138632, Singapore*
[8] *Rehabilitation Research Institute of Singapore, 11 Mandalay Road #14-03, Clinical Sciences Building, 308232, Singapore*
[9] *Co-first authors*
[10] *leehk@bii.a-star.edu.sg*
[11] *malini_olivo@sbic.a-star.edu.sg*

Abstract: Atopic dermatitis (AD) is a skin inflammatory disease affecting 10% of the population worldwide. Raster-scanning optoacoustic mesoscopy (RSOM) has recently shown promise in dermatological imaging. We conducted a comprehensive analysis using three machine-learning models, random forest (RF), support vector machine (SVM), and convolutional neural network (CNN) for classifying healthy versus AD conditions, and sub-classifying different AD severities using RSOM images and clinical information. CNN model successfully differentiates healthy from AD patients with 97% accuracy. With limited data, RF achieved 65% accuracy in sub-classifying AD patients into mild versus moderate-severe cases. Identification of disease severities is vital in managing AD treatment.

1. Introduction

Atopic dermatitis (AD) is a chronic inflammatory skin disease with itch, inflammation and red rashes. The prevalence of AD is approximately 10% and it is more commonly seen in children, especially children under the age of five [1–3]. Nonetheless, an adult who suffers from AD tends to represent a more persistent and severe condition [4]. The cause of AD is unknown but there is evidence that suggests that genetic, allergic and environmental factors can be related to the development of AD [5–7].

The disease progression of AD can be variable and can proceed in three directions: (i) persistent AD, (ii) intermittent AD or (iii) improvement. It is therefore important to diagnose and prognosticate AD so that treatment can be tailored at each stage. Currently, many scoring systems are available to assess the severity of AD, such as Eczema Area Severity Index (EASI) and

#415105 https://doi.org/10.1364/BOE.415105
Journal © 2021 Received 16 Nov 2020; revised 15 Feb 2021; accepted 9 Mar 2021; published 27 May 2021

Scoring Atopic Dermatitis (SCORAD), modified EASI (mEASI) and others [8–10]. However, these scoring systems are semi-quantitative at the best as they are designed based on metrics such as itchiness, redness and scale of the affected skin region, while others include the quality of patients' life. In addition, these scoring systems are based on visual inspections and it has been reported that visual skin assessments can only differentiate severities of AD in 25% of cases when self-assessed by patients [11]. Furthermore, it requires experience and training for clinicians to make visual assessments, subjecting these scoring systems to inter-rater variability [12]. It is desirable to have a non-invasive objective scoring tool that reflects the true AD severity throughout the therapeutic intervention and AD clinical trial, especially for mild and non-mild severities.

Raster-Scanning Optoacoustic Mesoscopy (RSOM), first introduced in 2013, is an emerging hybrid optical and ultrasound imaging technique that offers non-invasive, deep penetration imaging and provides high-resolution images [13]. RSOM imaging provides deep skin structural imaging up to 1–2 mm beneath the skin surface with high resolutions up to ~7 μm axial and ~30 μm lateral resolution [14]. With these resolutions, the vascular remodeling of the skin in various clinical severities of AD can be detected by RSOM imaging [12,15,16]. Differential diagnosis is thus critical to achieve accurate AD diagnosis [17]. Considerable efforts have been made to explore the efficacy of RSOM in assessing skin inflammatory diseases. For example, skin-specific metrics derived from RSOM such as total blood volume (TBV) and epidermis thickness (ET) have shown a substantial difference between control and skin inflammatory conditions [,14, 18]. In another study by Li et al., the feasibility of using RSOM derived skin-specific metrics in different skin phenotypes populations was investigated [19].

Machine learning models have shown significant success in the classification of skin disease diagnosis using dermatological images of superficial skin conditions [20–27]. However, these models have not been applied on RSOM images except for the work by Yew et al. [18]. In the work by Yew et al., the authors proposed an objective AD severity evaluation metrics – the Eczema Vascular Severity Index (EVSI) using Support Vector Machine (SVM) [18]. Handcrafted skin-specific features derived from RSOM images such as TBV, low-high frequency ratio (LHFR) and ET were used as features to train the model [18]. However, Convolutional Neural Networks (CNNs) were not utilized to automatically extract useful features from 3D RSOM images. In this study, we explore the utilization of CNNs for automatic extraction of useful features from 3D RSOM images and combine these features with handcrafted features proposed by Yew et al. [18].

We conducted a comprehensive analysis using three machine learning (ML) methods, SVM, Random Forest (RF) and CNNs in classifying healthy and various AD conditions. We performed two analyses (i) Healthy vs. AD and (ii) Mild vs. Moderate-Severe AD conditions. The motivation for conducting the second analysis in classifying between mild and more serious AD conditions is that patient-specific clinical care can be provided accordingly for better treatment outcomes.

To the best of the authors' knowledge, this study is the first effort to employ raw 3D RSOM images for the classification of AD conditions using a deep learning model. The objective of the study is to evaluate the performance of SVM, RF and CNNs in classifying healthy vs. AD conditions and mild vs. moderate-severe AD conditions. We designed an optimal neural network architecture that receives 3D RSOM images and other handcrafted features and successfully combines them in the network.

2. Methods and materials

2.1. Overview

In this study, we performed a thorough analysis by applying three different ML models on 3D RSOM images and compared the performance of each model using different combinations of inputs to the model. We utilized raw 3D RSOM images and four handcrafted features as inputs to train ML models. Three handcrafted features were derived from RSOM images, proposed by Yew et al. [17], namely TBV, ET and LHFR. The fourth feature is trans-epidermal water loss (TEWL), which reflects skin barrier dysfunction and is shown to be affected in AD condition [28].

The workflow of analysis is as follows: First, we evaluated the performance of traditional ML models such as SVM and RF using different combinations of the following features: TBV, ET, LHFR and TEWL. Secondly, we adopted CNN and used raw 3D RSOM images as inputs to train the network. Thirdly, we employed both 3D RSOM images and handcrafted features information to train the CNN model. The performance of models for every combination of inputs was compared and reported.

2.2. Subjects

This study was approved by the Domain Specific Review Board (DSRB) of the National Health Group, Singapore (Ref No. 2017/00932). Patients were imaged in compliance with our institutional approvals and informed consent was obtained. Study participants were recruited from AD patients visiting the National Skin Centre, Singapore. The diagnosis of AD was made based on the Hanifin and Rajka diagnostic criteria [29]. This study also included healthy controls who were defined as not having AD, any form of inflammatory skin diseases and any atopic co-morbidities such as asthma, allergic rhinitis and allergic conjunctivitis. 76 participants were recruited for this study, 53 were AD patients and 23 were healthy controls. All 53 AD participants had their disease severity assessed by an experienced dermatologist using SCORAD. There were 19, 26 and 8 patients suffering from mild, moderate and severe AD, respectively. The criteria of AD severity using SCORAD is as follows: below 25 is defined as mild, between 25 and 50 as moderate and greater than 50 as severe [8].

2.3. Image acquisition

The 3D RSOM images were collected by using RSOM Explorer C50 system (iThera Medical GmbH, Germany). RSOM system was implemented with one diode-pumped solid-state (DPSS, Nd:YAG, 532 nm) to provide < 1 ns pulse and a per-pulse energy up to 125 µJ with a laser's repetition rate of 270 Hz. The flexible articulated arm of RSOM allows raster scanning of 5 mm × 3 mm area on the skin in about 2.5 minutes. An in-tandem illumination-detector element is located at the focal point of the transducer, which raster-scans the two-dimensional (2D) region of interest (ROI) on the skin in a regularly-spaced acquisition grid and collects the ultra sound signals from 11 to 99 MHz. This non-invasive RSOM system can provide good quality 3D images with high resolution and deep penetration from the skin surface. The 3D RSOM images were visualized with two frequency sub-bands, high-frequency (HF) (33–99 MHz) in green and low-frequency (LF) (11–33 MHz) in red, representing the small and big vascular structure, respectively (Fig. 1).

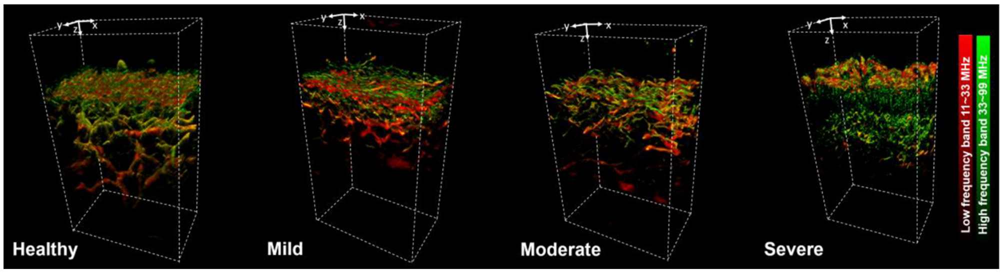


Fig. 1. Representative 3D RSOM images for healthy, mild, moderate and severe AD conditions are shown in the figure.

2.4. Features

In addition to raw voxel values, three handcrafted features extracted from RSOM images and one feature measured using VapoMeter, were used in the evaluation: 1) Total blood volume (TBV), 2) Epidermis thickness (ET), 3) The ratio of low and high frequency signals (LHFR) and 4) Transepidermal water loss (TEWL). TEWL was measured using VapoMeter (Delfin Technologies, Kuopio, Finland). TBV, ET and LHFR were extracted from 3D RSOM images using a similar algorithm as described in Li's study [19]. Briefly, TBV was computed as TBV = ΣN × dV, where N is the number of pixels with values above a threshold (20% of the maximum value in the dermis region) and dV is a certain voxel's volume. ET was computed as the distance between the skin surface and the melanin layer, which is the basal layer of the epidermis. The melanin layer was detected by averaging pixel-values in the x-y plane along the z-axis and finding the region with overlapping peaks for both high and low frequency optoacoustic profile, representing the center of the melanin layer. The full width at half maximum (FWHM) was performed to obtain the boundary of the melanin layer. LHFR was computed as the ratio of the mean value of the pixels in the LF band image and the HF band image in the dermis region.

2.5. Machine learning (ML) models

Two traditional ML approaches, namely SVM and RF were adopted in classifying (i) healthy vs. AD conditions and (ii) mild AD vs. moderate-severe AD conditions. A linear kernel was employed for the SVM. The RF consists of an ensemble of 25 decision trees, allowed to grow up to a depth of 5 using the entropy criterion. The implementation employed the python library scikit-learn [30]. For both approaches, four features (TBV, ET, LHFR and TEWL) were used as inputs. Features were normalized between 0 and 10. Various combinations of these features were explored to determine the combination that yields the highest accuracy. "Balanced" mode was enabled to account for data balancing during the model training [30]. This mode adjusts the weights given to each sample to be inversely proportional to the class frequency in the training data, which was similar to balancing the data in each severity class during training.

All the analyses were performed in a six-fold validation fashion. Patients were split into two groups (training and validation) in a stratified manner. Since we had limited datasets, we did not have testing datasets. Therefore, the number of samples used in this ML analysis consists of 53 AD and 17 healthy subjects. 80% of patients from each AD severity were assigned as training data and 20% of patients belonging to each AD severity were assigned to validation data.

2.6. Deep learning (DL) models

2.6.1. Network architecture

Our CNN model consists of nine layers as shown in Fig. 2. It is made of five alternating repeated convolutional layers and max-pooling layers followed by five fully-connected (FC) layers to reduce the model complexity [31]. The inputs to the models were 3D LF and HF RSOM images, with a shape of 424 × 64 × 64, stacked together, resulting in a size of 2 × 424 × 64 × 64. The first layer in the model contains 64 convolutional filters of size 3 × 3 × 3 with a stride of 2 × 2 × 2. ReLU (Rectified Linear Unit) activation was applied and was followed by a max-pooling layer as shown below. A softmax was used as the activation function for the output layer.

We trained two CNN models: (i) using only LF and HF 3D RSOM images (without handcrafted features and (ii) using LF and HF 3D RSOM images with handcrafted features (TBV, ET, LHFR and TEWL). For the second CNN models trained with handcrafted features, the four features were concatenated at the bottleneck layer as shown in Fig. 2. All features were normalized between 0 and 10 to improve CNN model stability and modeling performance.

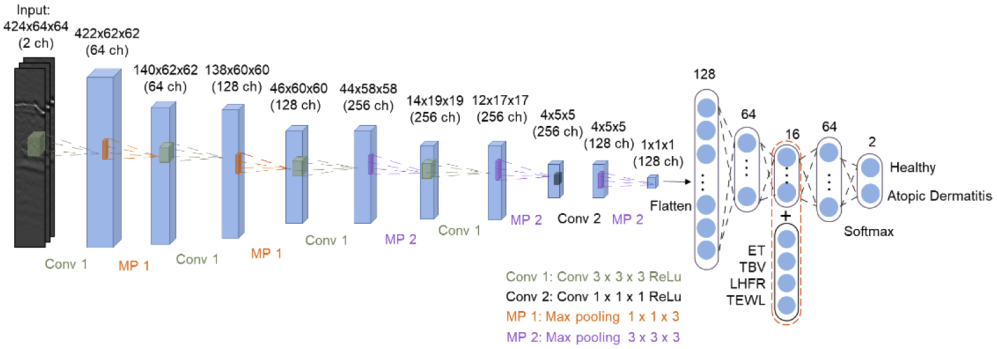


Fig. 2. Figure shows the CNN architecture that was trained to classify 3D RSOM images. Low and high frequency of RSOM images were stacked together, resulting in an input size of 2 × 424 × 64 × 64. Images underwent a series of convolutional (Conv) and max-pooling (MP) layers for feature extraction. ET (epidermis thickness), TBV (total blood volume), LHFR (low and high frequency ratio), and TEWL (transepidermal water loss) were added at the bottleneck layer.

2.6.2. Data augmentation and balancing

Similar to the ML experiment, CNN analyses were carried out in a six-fold validation fashion. In each fold, data augmentation was performed to balance the data in each severity class. The data augmentation was performed in the training and validation data sets at the patient-level to avoid information leakage [32,33].

The size of 3D RSOM images was 500 × 150 × 250 initially with an axial resolution of 7 µm and a lateral resolution of 30 µm. We first cropped away the blank regions in 3D RSOM images, which did not contain any skin information, resulting in a fixed-size of 424 × 150 × 250. All 3D RSOM images were checked carefully to ensure the resultant cropped 3D RSOM images covered the dermal and epidermal layers. We performed a data augmentation process, as shown in Fig. 3, including flipping (left-right and front-back) and rotating the 3D image stack about the z-axis (45°, 90°, 135°, and 180°). Then, we cropped regions with fixed-size equal to 424 × 64 × 64 pixels from the rotated or flipped images. 64 × 64 pixels were carefully chosen to include sufficient information for the CNNs model to extract adequate features. Limited by GPU memory, we could only use a cropped region as big as 424 × 64 × 64 pixels in our analysis. This cropping process was repeated until the number of samples in each severity class matched (refers to next paragraph for details). The same cropping was done on both HF and LF images and cropped regions were used as a pair during the training process.

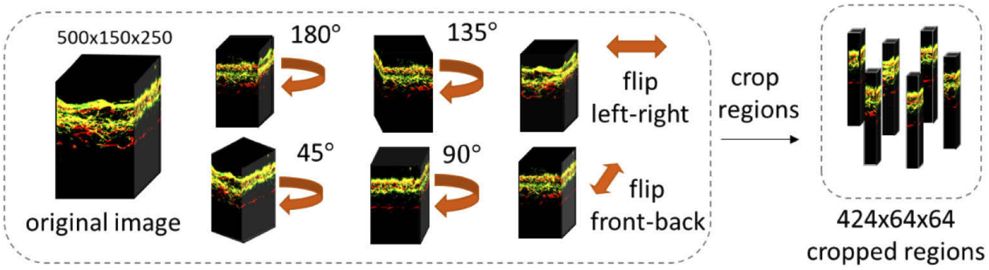


Fig. 3. Data Augmentation Process: To generate a sample, the original 3D RSOM images were rotated about the z-axis or flipped left-right or front-back and a small region of fixed size 424 × 64 × 64 was randomly cropped from the rotated or flipped image.

Tables 1 and 2 show the number of cropped regions per patient for one of the cross-validation datasets using 3D RSOM images as input to train the model (23 healthy and 53 AD patients). In the healthy vs. AD analysis, we performed data augmentation and balancing as described above until we had ~200 training samples and ~20 validation samples in each severity class. For example, in Table 1, we have 15 mild AD patients, five regions were randomly cropped after rotating and flipping the patient's HF and LF 3D RSOM images giving a total of 75 cropped regions for the mild category. A similar number of regions were augmented for patients with moderate and severe AD conditions giving a total of 210 augmented regions for AD condition (mild, moderate and severe classes), which were comparable to the total number of augmented regions for the healthy subjects (209 regions). Similar data augmentation was performed for validation data and mild vs. moderate-severe analyses. The details of the total number of regions cropped are shown in Table 2.

Table 1. Table shows the number of subjects in the experiment of healthy vs atopic dermatitis, the number of cropped regions per subject and total number of samples generated at the end of the augmentation process (stated in the total number of cropped regions column) for training and validation data sets. The table shows one of the cross-validation fold.

Healthy vs. Atopic Dermatitis Analysis						
	Training Data			*Validation Data*		
Severity	Num. of Subjects	Num. of Cropped Regions per Subject	Total Num. of Cropped Regions	Num. of Subjects	Num. of Cropped Regions per Subject	Total Num. of Cropped Regions
Mild	15	5	75	4	1	4
Moderate	20	3	60	6	1	6
Severe	6	12	72	2	1	6
Healthy	19	11	209	4	5	20

Table 2. Table shows the number of patients in the experiment of mild vs moderate-severe, the number of cropped regions per patient and total number of samples generated at the end of the augmentation process (stated in the total number of cropped regions column) for training and validation data sets. The table shows one of the cross-validation fold.

Mild vs. Moderate-Severe Analysis						
	Training Data			*Validation Data*		
Severity	Num. of Patients	Num. of Cropped Regions per Patients	Total Num. of Cropped Regions	Num. of Patients	Num. of Cropped Regions per Patients	Total Num. of Cropped Regions
Mild	15	5	225	4	5	20
Moderate	20	5	100	6	2	12
Severe	6	18	108	2	5	10

2.6.3. Model training and evaluation

The network was trained for 200 epochs with learning rate 1×10^{-5}, learning rate decay of 0.05 and learning step decay of two with Adam optimizer. The batch size was set to be four. All computations were carried out on a Linux workstation with Intel Core i7-4790 CPU with 3.6 GHz clock speed, 16 GB RAM and a GeForce GTX TITAN X. It took approximately 9 min for one epoch and a total of 30 hours to train the model using the above-mentioned workstation. Tensorflow 1.12 [34] implementation was used in our study. Three CNN models were trained using different combinations of inputs.

The first CNN model trained used only LF and HF 3D RSOM images (23 healthy and 53 AD cases) as inputs. The second CNN model trained used 3D RSOM images, and three features

(TBV, LHFR, and TEWL) added at the bottleneck layer (Fig. 2). The third CNN model used 3D RSOM images and four features (ET, TBV, LHFR, and TEWL). For the second and third analyses, 6 cases out of the 23 healthy cases did not have complete feature information, making the number of samples to be 53 AD and 17 healthy cases. Validation data was used to evaluate the models' prediction accuracy. Since one patient would have more than one sample due to the cropping pipeline, majority voting was performed to determine the final prediction for that patient. If there is a patient without a final prediction due to having an equal number of prediction outcomes, one additional sample was randomly cropped from the 3D RSOM images and evaluated to obtain the final prediction.

3. Results

Table 3 tabulates the average and standard deviation of the validation accuracy of RF, SVM and CNN for six-fold cross-validation results. Figure 4 shows the confusion matrices for the three models evaluated on validation datasets. The confusion matrices shown are for models that yielded the highest validation accuracy as reported in Table 3.

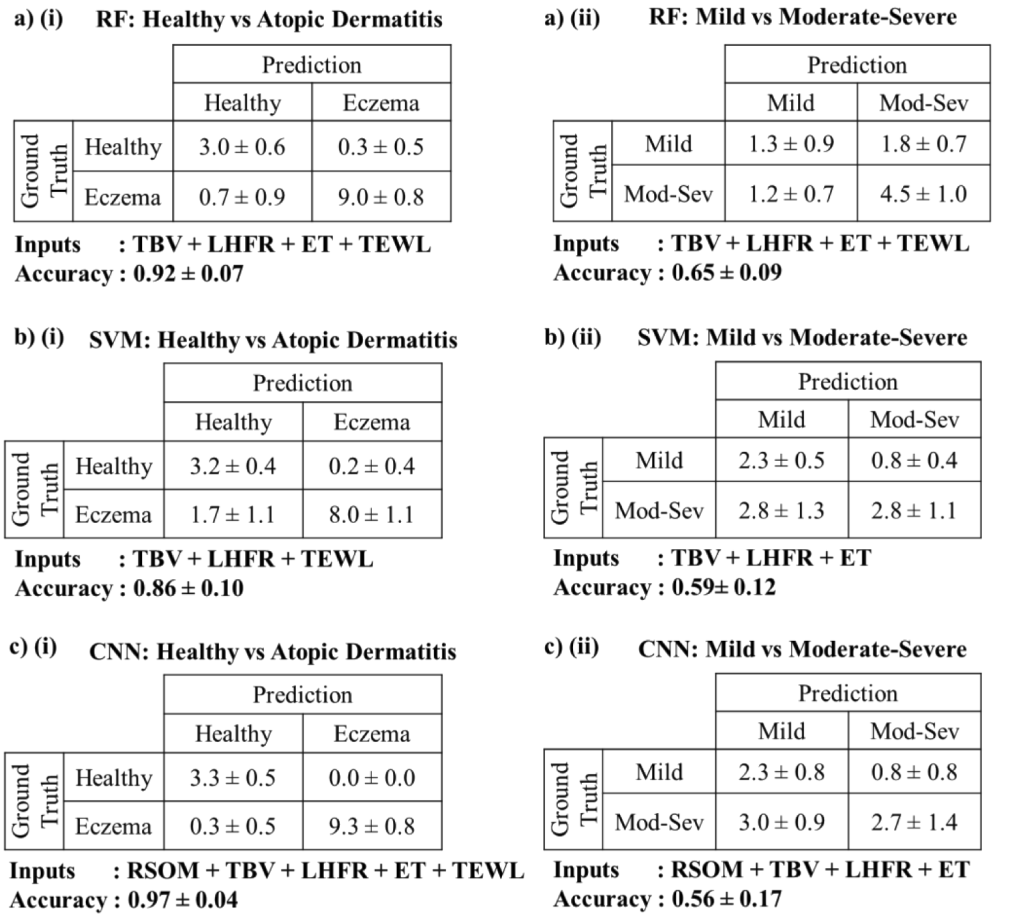
Table 3. Validation accuracy for two analyses using three models with specified inputs to the models. Values shown are average and standard deviation for six-fold cross-validation. The highest validation accuracy for each particular model is shown in bold. TBV: Total Blood Volume, LHFR: Low High Frequency Ratio, ET: Epidermis Thickness, TEWL: Transepidermal water loss measured from VapoMeter.

Model	Inputs to Model					Validation Accuracy	
	3D RSOM	Features derived from RSOM			TEWL	Healthy vs. Atopic Dermatitis	Mild vs. Moderate-Severe
		TBV	LHFR	ET			
RF		✓	✓	✓		0.81 ± 0.08	0.63 ± 0.12
		✓	✓	✓	✓	**0.92 ± 0.07**	**0.65 ± 0.09**
		✓	✓		✓	0.91 ± 0.06	0.59 ± 0.14
SVM		✓	✓	✓		0.77 ± 0.08	**0.59 ± 0.12**
		✓	✓	✓	✓	0.82 ± 0.11	0.58 ± 0.18
		✓	✓		✓	**0.86 ± 0.10**	0.49 ± 0.18
CNN	✓					0.48 ± 0.13	0.47 ± 0.16
	✓	✓	✓	✓		0.94 ± 0.10	**0.56 ± 0.17**
	✓	✓	✓	✓	✓	**0.97 ± 0.04**	0.54 ± 0.27

In healthy vs. AD analysis, CNN achieved the highest performance among the three models, giving a validation accuracy of 97%, using all four features. However, when using only LF and HF 3D RSOM images, CNN yielded only 48% validation accuracy in classifying healthy vs. AD condition. Adding three handcrafted features (TBV, LHFR and ET) to the model increased the validation accuracy to 94%. The performance of CNN was further improved by 3% when TEWL was added to the model, achieving 97% accuracy. For ML models, RF performed better than SVM in all the analyses performed for different combinations of features.

In mild vs. moderate-severe analysis, RF gave the highest accuracy of 65% in severity score prediction among all three models using all four features. SVM model, on the other hand, showed a validation accuracy of 59% when TBV, LHFR, and ET were used. Lastly, CNN exhibited slightly lower accuracy at 56% in predicting severity compared to RF and SVM, using 3D RSOM images and three handcrafted features derived (TBV, LHFR, ET) from RSOM images as inputs.

Figure 5 shows the representative RSOM images that were correctly and wrongly classified by CNN in the analyses of healthy vs. AD conditions. While the structural differences between healthy and AD conditions are apparent, the moderate AD condition that was wrongly classified



a) (i) RF: Healthy vs Atopic Dermatitis

		Prediction	
		Healthy	Eczema
Ground Truth	Healthy	3.0 ± 0.6	0.3 ± 0.5
	Eczema	0.7 ± 0.9	9.0 ± 0.8

Inputs : TBV + LHFR + ET + TEWL
Accuracy : 0.92 ± 0.07

a) (ii) RF: Mild vs Moderate-Severe

		Prediction	
		Mild	Mod-Sev
Ground Truth	Mild	1.3 ± 0.9	1.8 ± 0.7
	Mod-Sev	1.2 ± 0.7	4.5 ± 1.0

Inputs : TBV + LHFR + ET + TEWL
Accuracy : 0.65 ± 0.09

b) (i) SVM: Healthy vs Atopic Dermatitis

		Prediction	
		Healthy	Eczema
Ground Truth	Healthy	3.2 ± 0.4	0.2 ± 0.4
	Eczema	1.7 ± 1.1	8.0 ± 1.1

Inputs : TBV + LHFR + TEWL
Accuracy : 0.86 ± 0.10

b) (ii) SVM: Mild vs Moderate-Severe

		Prediction	
		Mild	Mod-Sev
Ground Truth	Mild	2.3 ± 0.5	0.8 ± 0.4
	Mod-Sev	2.8 ± 1.3	2.8 ± 1.1

Inputs : TBV + LHFR + ET
Accuracy : 0.59± 0.12

c) (i) CNN: Healthy vs Atopic Dermatitis

		Prediction	
		Healthy	Eczema
Ground Truth	Healthy	3.3 ± 0.5	0.0 ± 0.0
	Eczema	0.3 ± 0.5	9.3 ± 0.8

Inputs : RSOM + TBV + LHFR + ET + TEWL
Accuracy : 0.97 ± 0.04

c) (ii) CNN: Mild vs Moderate-Severe

		Prediction	
		Mild	Mod-Sev
Ground Truth	Mild	2.3 ± 0.8	0.8 ± 0.8
	Mod-Sev	3.0 ± 0.9	2.7 ± 1.4

Inputs : RSOM + TBV + LHFR + ET
Accuracy : 0.56 ± 0.17

Fig. 4. Confusion Matrix for all models in the two analyses. The confusion matrices shown is generated using the model that yielded the highest validation accuracy as reported in Table 3. Values shown are mean and standard deviation for six-fold cross-validations. Inputs are the images/features used to train the model. RF: Random Forest, SVM: Support Vector Machine, CNN: Convolutional Neural Networks, TBV: Total Blood Volume, LHFR: Low High Frequency Ratio, ET: Epidermis Thickness, TEWL: Transepidermal Water Loss.

as healthy may have similar structural features to the healthy case, such as the absence of capillary loops, and intact epidermis. Figure 6 shows the representative RSOM images that were wrongly classified by CNN in the analysis of mild vs. moderate-severe AD conditions. In other words, the mild AD case was classified as a moderate-severe AD case and vice versa.

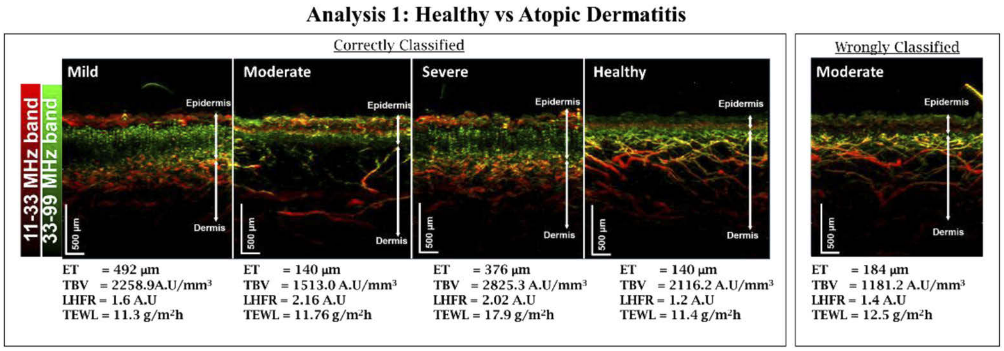


Fig. 5. Representative RSOM cross-sectional images of correctly and wrongly classified by CNN in the experiment of classifying healthy vs. AD condition.

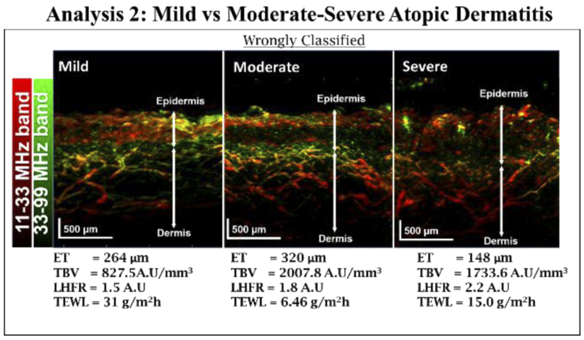


Fig. 6. Representative RSOM cross-sectional images of wrongly classified samples by the CNN in the experiment of classifying mild vs. non-mild AD (moderate-severe) conditions.

4. Discussion

In recent years, there have been a number of studies deploying deep learning for skin disease classification. These studies applied deep learning on dermoscopic images [26,35,36] but not on 3D optoacoustic images. In this paper, we deploy deep learning for classification of 3D RSOM images.

Using a small data set (76 patients), we could achieve a better diagnostic accuracy compared to the model by Liu et al. which was trained on a large data set (∼16 k dermoscopic images) [37]. Using RSOM images, we were able to boost the model's performance with a smaller data set. These results suggested that the ML model's performance could be largely affected by the data information presented to the model. As data labeling is a laborious task for preparing large data sets of medical imaging for ML, highly efficient models on small data sets can be a boon in timely analyses [38–40].

We extended our work based on previous studies on RSOM image analysis. In the current study, we adopted the ML model to classify different AD severity conditions – mild vs. moderate-severe condition, which was not performed in the past. Furthermore, we performed comprehensive ML training and predictive validation analyses, where we did cross-validation for the generalization of our model. It is particularly important to obtain more reliable results especially when the datasets are small. Our cross-validation results showed that CNNs had higher prediction accuracies than RF and SVM, with an accuracy of 97%, achieving 11% improvement over SVM, in classifying healthy and AD conditions [18].

Using all four handcrafted features, traditional ML models such as RF and SVM could achieve a diagnostic accuracy of 92% and 86% in classifying healthy and AD condition, respectively (refers to Table 3). When raw 3D RSOM images were added to the pipeline using CNN, the diagnostic accuracy was improved to 97% and the model demonstrated more stability in prediction compared to RF and SVM, judging from the low standard deviation. This suggested that the CNN model indeed extracted useful features from 3D RSOM images, which aided in enhancing the CNN's diagnostic accuracy. Even though CNN showed very high diagnostic accuracy in classifying healthy and AD conditions, it did not achieve similar performance in classifying mild vs. moderate-severe AD conditions. The CNN's highest diagnostic accuracy for mild vs moderate-severe AD classification was 56%. A similar prediction accuracy was observed in RF and SVM, where the average diagnostic accuracy was ~60%.

From Fig. 5, the RSOM cross-sectional images for healthy and AD cases were visibly distinguishable. CNN model thus was able to extract useful features in classifying healthy and AD cases even though the datasets were small. For the mild vs. moderate-severe AD conditions, using both raw 3D RSOM images and handcrafted features did not improve the CNN model's accuracy as what we had observed in the healthy vs. AD classification. We believe this was because it is challenging to differentiate between mild and moderate-severe AD RSOM images, as shown in Figs. 5 and 6. There are several reasons for the erroneous classification in Fig. 6. Firstly, since pathological and physiological features form the basis for determining the severity of AD in this study, any deviation in the features will affect the CNN's prediction accuracy. Notably, the mild representative case in Fig. 6 exhibited a TEWL value of 31 g/m^{2}h, far higher than that of severe AD cases. Similarly, the TEWL value of the moderate representative case in Fig. 6 was lower than that of healthy subjects, possibly rendering the wrong classification of the case to be 'mild'. Secondly, if the structural features of the RSOM images are lost due to skin barrier dysfunction in severe AD cases, the feature quantification is challenging since the boundary between epidermis and dermis region is not delineated. As in the severe case in Fig. 6, the ET calculation yields a value of 148 µm, similar to that of healthy subjects which leads it to be wrongly classified. The limited amount of training data further adds to the difficulty resulting in inaccurate classification for validation data set. As discussed, there are a total of 41 AD and 19 healthy subjects in our data set for classifying healthy vs AD cases. However, to predict mild vs moderate-severe AD subjects, fewer samples are available including 15 mild AD subjects and 26 moderate-severe AD subjects. The limited number of samples is another reason for why classification of mild AD vs moderate-severe AD subjects is harder. CNN models in general require many more samples in order to learn to extract useful features for classification. Retraining the model with a larger data set will mitigate this problem.

During ML training, data balance between classes (e.g. healthy vs. disease) is important to ensure the number of samples from both classes is similar. It is particularly important to perform data splitting at the patients-level to avoid potential data leakage from training data to validation data [37]. We have successfully developed a CNN-based pipeline, which include data preparation, data augmentation and model training to recognize various AD severity conditions using raw 3D RSOM images, and handcrafted features. This CNN-based pipeline thus will handle the data splitting at patient-level and is not limited only to skin AD disease classification. It is designed in a modularized manner and has the flexibility to be applied for classification of other skin inflammatory diseases such as rosacea, and psoriasis and other 3D optoacoustic images such as optical coherence tomography, multispectral optoacoustic tomography and multispectral optoacoustic mesoscopy. We have shared our code in https://github.com/davidc9320sg/rsom-dermatitis-cnn/.

It is crucial to diagnose AD severity accurately to monitor the treatment response and plan effective clinical care for patients. We successfully proposed an optimal network architecture suitable for 3D optoacoustic images for AD conditions classification, which can be used as an objective evaluation tool for assessing AD conditions in clinics. At the current state, our CNN model is unable to achieve desirable diagnostic accuracy in classifying mild vs. moderate-severe AD conditions. One reason could be that the AD severities were determined from SCORAD scoring which was subjected to inter- and intra-observational variability, the accuracy may therefore suffer from discrepancies from the SCORAD results. While SCORAD or EASI takes into account the presentation and frequency of AD symptoms, the subsurface inflammation physiology of the skin was out of the scoring framework. With the naked eye, it may be possible to observe that superficial symptoms are improved overtime with treatment, but inflammation may persist under the skin that can significantly impact the way we classify the disease severity. With more data collected and consensus among multiple diagnoses for each patient, the model can be re-trained using the current framework as a baseline to further enhance its accuracy.

There are several limitations in this study. Firstly, the size of the data set was small (76 patients) and the population was mainly Asian cohort. Through an on-going collaboration, we are aiming to expand this study by including patients with lower Fitzpatrick scores (I-II). Secondly, the size of the cropped sample was set at 64×64 pixels due to our insufficient GPU memory. A larger cropped sample might aid in improving CNN models since it provides more information to the CNN model.

The AD classification model in this paper can be an adjunctive diagnostic tool to aid in clinical decisions, especially in differentiating between mild and non-mild AD severities. As even clinicians with experience in optoacoustic images may face challenges to interpret the RSOM images, our classification model aims to classify AD severity with higher sensitivity by extracting features from volumetric vascular structure in 3D RSOM images rather than one-plane imaging features. This proposed pipeline provides the foundation for an AI-aided AD diagnosis and treatment platform.

5. Conclusion

To conclude, we have evaluated the performance of three ML models in classifying AD conditions using 3D RSOM images, handcrafted features derived from RSOM images and transepidermal water loss. Our results showed that CNN models yield the highest accuracy (97%) in classifying healthy vs. AD conditions while RF achieve the highest accuracy (65%) in classifying mild vs. moderate-severe AD conditions. This is the first study to classify AD severity using 3D RSOM images. We developed a pipeline to prepare 3D RSOM images for training a CNN model and showed that the use of raw RSOM voxel values can be advantageous over handcrafted features. Our method can easily be extended to other inflammatory skin diseases such as rosacea and psoriasis.

Funding. Biomedical Research Council; National Medical Research Council (NMRC OF-IRG Grant OFIRG18nov-0101).

Acknowledgment. Acknowledgement also goes to SBIC-iThera joint lab for their help in this study.

Disclosures. The authors declare no conflicts of interest.

References

1. J. I. Silverberg and J. M. Hanifin, "Adult eczema prevalence and associations with asthma and other health and demographic factors: A US population–based study," J. Allergy Clin. Immunol. **132**(5), 1132–1138 (2013).
2. G. Pesce, A. Marcon, A. Carosso, L. Antonicelli, L. Cazzoletti, M. Ferrari, A. G. Fois, P. Marchetti, M. Olivieri, P. Pirina, G. Pocetta, R. Tassinari, G. Verlato, S. Villani, and R. de Marco, "Adult eczema in Italy: prevalence and associations with environmental factors," J Eur Acad Dermatol Venereol **29**(6), 1180–1187 (2015).

3. Y. I. Lopez Carrera, A. Al Hammadi, Y.-H. Huang, L. J. Llamado, E. Mahgoub, and A. M. Tallman, "Epidemiology, Diagnosis, and Treatment of Atopic Dermatitis in the Developing Countries of Asia, Africa, Latin America, and the Middle East: A Review," Dermatol Ther (Heidelb) **9**(4), 685–705 (2019).
4. C. Hoare, A. Li Wan Po, and H. Williams, "Systematic review of treatments for atopic eczema," Health Technology Assessment **4**, 1–191 (2000).
5. F. S. Larsen, N. V. Holm, and K. Henningsen, "Atopic dermatitis. A genetic-epidemiologic study in a population-based twin sample," J. Am. Acad. Dermatol. **15**(3), 487–494 (1986).
6. Q. Deng, C. Lu, Y. Li, J. Sundell, and N. Dan, "Exposure to outdoor air pollution during trimesters of pregnancy and childhood asthma, allergic rhinitis, and eczema," Environ. Res. **150**, 119–127 (2016).
7. W. David Boothe, J. A. Tarbox, and M. B. Tarbox, "Atopic Dermatitis: Pathophysiology," in *Management of Atopic Dermatitis: Methods and Challenges*, E. A. Fortson, S. R. Feldman, and L. C. Strowd, eds. (Springer International Publishing, 2017), pp. 21–37.
8. R. Chopra, P. Vakharia, R. Sacotte, N. Patel, S. Immaneni, T. White, R. Kantor, D. Hsu, and J. Silverberg, "Severity strata for Eczema Area and Severity Index (EASI), modified EASI, Scoring Atopic Dermatitis (SCORAD), objective SCORAD, Atopic Dermatitis Severity Index and body surface area in adolescents and adults with atopic dermatitis," Br. J. Dermatol. **177**(5), 1316–1321 (2017).
9. P. P. Vakharia, R. Chopra, R. Sacotte, N. Patel, S. Immaneni, T. White, R. Kantor, D. Y. Hsu, and J. I. Silverberg, "Validation of patient-reported global severity of atopic dermatitis in adults," Allergy **73**(2), 451–458 (2018).
10. Y. Leshem, T. Hajar, J. Hanifin, and E. Simpson, "What the Eczema Area and Severity Index score tells us about the severity of atopic dermatitis: an interpretability study," Br. J. Dermatol. **172**(5), 1353–1357 (2015).
11. C. R. Charman, A. J. Venn, H. Williams, and M. Bigby, "Measuring atopic eczema severity visually: which variables are most important to patients?" Arch. Dermatol. **141**(9), 1146–1151 (2005).
12. A. Bożek and A. Reich, "Assessment of intra-and inter-rater reliability of three methods for measuring atopic dermatitis severity: EASI, objective SCORAD, and IGA," Dermatology **233**(1), 16–22 (2017).
13. M. Omar, J. Gateau, and V. Ntziachristos, "Raster-scan optoacoustic mesoscopy in the 25–125 MHz range," Opt. Lett. **38**(14), 2472–2474 (2013).
14. J. Aguirre, M. Schwarz, N. Garzorz, M. Omar, A. Buehler, K. Eyerich, and V. Ntziachristos, "Precision assessment of label-free psoriasis biomarkers with ultra-broadband optoacoustic mesoscopy," Nat. Biomed. Eng. **1**(5), 0068 (2017).
15. B. Hindelang, J. Aguirre, A. Berezhnoi, T. Biedermann, U. Darsow, B. Eberlein, and V. Ntziachristos, "Quantification of skin sensitivity to ultraviolet radiation using ultra-wideband optoacoustic mesoscopy," *Br. J. Dermatol.* (2020).
16. B. Hindelang, J. Aguirre, A. Berezhnoi, H. He, K. Eyerich, V. Ntziachristos, T. Biedermann, and U. Darsow, "Optoacoustic mesoscopy shows potential to increase accuracy of allergy patch testing," Contact Dermatitis (2020).
17. C. Avena-Woods, "Overview of atopic dermatitis," Am. J. Manag. Care **23**, S115–S123 (2017).
18. Y. W. Yew, U. Dinish, A. H. Y. Kuan, X. Li, K. Dev, A. B. E. Attia, R. Bi, M. Moothanchery, G. Balasundaram, and J. Aguirre, "Raster-scanning optoacoustic mesoscopy (RSOM) imaging as an objective disease severity tool in atopic dermatitis patients," *J. Am. Acad. Dermatol.* (2020).
19. X. Li, U. Dinish, J. Aguirre, R. Bi, K. Dev, A. B. E. Attia, S. Nitkunanantharajah, Q. H. Lim, M. Schwarz, and Y. W. Yew, "Optoacoustic mesoscopy analysis and quantitative estimation of specific imaging metrics in Fitzpatrick skin phototypes II to V," J. Biophotonics **12**, e201800442 (2019).
20. Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, and M. Fujimoto, "Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis," Br. J. Dermatol. **180**(2), 373–381 (2019).
21. J. Premaladha and K. S. Ravichandran, "Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms," J Med Syst **40**(4), 96 (2016).
22. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature **542**(7639), 115–118 (2017).
23. A. Rezvantalab, H. Safigholi, and S. Karimijeshni, "Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms," arXiv preprint arXiv:1810.10348 (2018).
24. S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," J. Invest. Dermatol. **138**(7), 1529–1538 (2018).
25. E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian, "Melanoma detection by analysis of clinical images using convolutional neural network," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC)(2016), pp. 1373–1376.
26. S.-Q. Wang, X.-Y. Zhang, J. Liu, C. Tao, C.-Y. Zhu, C. Shu, T. Xu, and H.-Z. Jin, "Deep learning-based, computer-aided classifier developed with dermoscopic images shows comparable performance to 164 dermatologists in cutaneous disease diagnosis in the Chinese population," Chin Med J (Engl) **133**(17), 2027–2036 (2020).
27. A. Minagawa, H. Koga, T. Sano, K. Matsunaga, Y. Teshima, A. Hamada, Y. Houjou, and R. Okuyama, "Dermoscopic diagnostic performance of Japanese dermatologists for skin tumors differs by patient origin: A deep learning convolutional neural network closes the gap," The Journal of Dermatology n/a.
28. S. G. Danby and M. J. Cork, "The skin barrier in atopic dermatitis," Curr Allergy Asthma Rep **14**(5), 433 (2014).
29. Y. W. Yew, U. Dinish, E. C. E. Choi, R. Bi, C. J. H. Ho, K. Dev, X. Li, A. B. E. Attia, M. K. W. Wong, and G. Balasundaram, "Investigation of morphological, vascular and biochemical changes in the skin of an atopic dermatitis

Biomedical Optics EXPRESS

(AD) patient in response to dupilumab using raster scanning optoacoustic mesoscopy (RSOM) and handheld confocal Raman spectroscopy (CRS)," J. Dermatol. Sci. **95**(3), 123–125 (2019).

30. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," the Journal of machine Learning research **12**, 2825–2830 (2011).
31. Z. Lin, R. Memisevic, and K. Konda, "How far can we go without convolution: Improving fully-connected networks," arXiv preprint arXiv:1511.02580 (2015).
32. N. Bussola, A. Marcolini, V. Maggio, G. Jurman, and C. Furlanello, "Not again! Data leakage in digital pathology," arXiv preprint arXiv:1909.06539 (2019).
33. L. Shi, G. Campbell, W. Jones, F. Campagne, Z. Wen, S. Walker, Z. Su, T. Chu, F. Goodsaid, and L. Pusztai, "The MAQC-II project: a comprehensive study of common practices for the development and validation of microarray-based predictive models," (2010).
34. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*(2016), pp. 265–283.
35. S. Jinnai, N. Yamazaki, Y. Hirano, Y. Sugawara, Y. Ohe, and R. Hamamoto, "The development of a skin cancer classification system for pigmented skin lesions using deep learning," Biomolecules **10**(8), 1123 (2020).
36. A. Minagawa, H. Koga, T. Sano, K. Matsunaga, Y. Teshima, A. Hamada, Y. Houjou, and R. Okuyama, "Dermoscopic diagnostic performance of Japanese dermatologists for skin tumors differs by patient origin: a deep learning convolutional neural network closes the gap," The Journal of Dermatology n/a (2020).
37. Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G. S. Corrado, L. H. Peng, D. R. Webster, D. Ai, S. J. Huang, Y. Liu, R. C. Dunn, and D. Coz, "A deep learning system for differential diagnosis of skin diseases," Nat. Med. **26**(6), 900–908 (2020).
38. N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation," Med. Image Anal. **63**, 101693 (2020).
39. G. Zhang, C.-H. R. Hsu, H. Lai, and X. Zheng, "Deep learning based feature representation for automated skin histopathological image annotation," Multimed Tools Appl **77**(8), 9849–9869 (2018).
40. D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," Nat. Commun. **11**(1), 3673 (2020).