Received: 15 April 2021 | Accepted: 31 July 2021

DOI: 10.1111/srt.13086

ORIGINAL ARTICLE

WILEY

Discriminative deep learning based benignity/malignancy diagnosis of dermatologic ultrasound skin lesions with pretrained artificial intelligence architecture

Alexandra Laverde-Saad[1] | Abdulhadi Jfri[1] | Rubén García[2] | Irene Salgüero[3] | Constanza Martínez[3] | Hirune Cembrero[3] | Gastón Roustán[3] | Fernando Alfageme[3]

[1] Dermatology Department, McGill University, Montreal, Quebec, Canada

[2] Dermatology Department, Hospital Universitario de Salamanca, Salamanca, Spain

[3] Dermatology Department, Hospital Universitario Puerta de Hierro Majadahonda, Madrid, Spain

Correspondence
Alexandra Laverde-Saad, Dermatology Department, McGill University, Montreal, Quebec H3A 0G4, Canada.
Email: alexandra.laverde-saad@mail.mcgill.ca

Abstract

Background: Deep-learning algorithms (DLAs) have been used in artificial intelligence aided ultrasonography diagnosis of thyroid and breast lesions. However, its use has not been described in the case of dermatologic ultrasound lesions. Our purpose was to train a DLA to discriminate benign form malignant lesions in dermatologic ultrasound images.

Materials and methods: We trained a prebuilt neural network architecture (EfficientNet B4) in a commercial artificial intelligence platform (Peltarion, Stockholm, Sweden) with 235 color Doppler images of both benign and malignant ultrasound images of 235 excised and histologically confirmed skin lesions (84.3% training, 15.7% validation). An additional 35 test images were used for testing the algorithm discrimination for correct benign/malignant diagnosis. One dermatologist with more than 5 years of experience in dermatologic ultrasound blindly evaluated the same 35 test images for malignancy or benignity.

Results: EfficientNet B4 trained dermatologic ultrasound algorithm sensitivity; specificity; predictive positive values, and predicted negative values for validation algorithm were 0.8, 0.86, 0.86, and 0.8, respectively for malignancy diagnosis. When tested with 35 previously unevaluated images sets, the algorithmt's accuracy for correct benign/malignant diagnosis was 77.1%, not statistically significantly different from the dermatologist's evaluation (74.1%).

Conclusion: An adequately trained algorithm, even with a limited number of images, is at least as accurate as a dermatologic-ultrasound experienced dermatologist in the evaluation of benignity/malignancy of ultrasound skin tumor images devoid of clinical data.

KEYWORDS
artificial intelligence, deep learning, dermatologic ultrasound, skin ultrasound

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *Skin Research and Technology* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Artificial intelligence and, more specifically, deep-learning algorithms (DLAs), are rapidly permeating image-based diagnostic modalities such as X-rays, films, CT, and ultrasound, showing high diagnostic accuracy and performance.[1–4] However, dermatologic ultrasound, a recent application of high-frequency ultrasound to the superficial structures of skin and appendages,[5,6] has not been explored as an image source for DLAs. The purpose of our investigation was to train a prebuilt deep-learning architecture with a limited dermatologic ultrasound database of skin tumors and evaluate its efficacy and efficiency in comparison with expert human diagnosis regarding benignity or malignancy diagnosis.

2 | MATERIALS AND METHODS

2.1 | Study design

This was a study of DLA performance, from training to validation and testing on common dermatologic ultrasound (dermUS) skin tumor images. The DLA was used to classify images as benign or malignant.

The Institutional Review Board exempted the study and informed consent was waived, as no patient data were used in the creation or testing of the DLA.

2.2 | Data

DermUS images were obtained from the institutional dermatologic ultrasound archive. Single scan color Doppler images acquired according to DERMUS guidelines[7] of 235 surgically excised or biopsied skin lesions were selected. Diagnosis of the lesions was restricted to seven categories as seen in Table 1. These diagnoses represent the most prevalent benign and malignant skin lesions in everyday clinical practice.[8]

Researchers confirmed the absence of patient identifiers prior to image extraction. Validation, training, and testing scans were scanned with a 10–22 MHz and 18 MHz lineal probe in a single machine (My Lab class C, Esaote, Geneva) for image homogeneity.

2.3 | Data manipulation

All images were cropped to 224 × 224 pixels to eliminate ultrasound exploration metadata. Images were placed in a separate older file and a csv index file with histological diagnosis; sex, age, benignity, or malignancy and probe resolution was matched with the image folder's corresponding image. Images that were blank, such as from loss of transducer-skin contact, were deleted. No other images were deleted, including those without optimal image quality, to increase model robustness and applicability.

2.4 | Deep-learning architecture and platform tested

EfficientNet B4 is built around 2D Depthwise convolution blocks,[9] which have been shown to be extremely cost-efficient and are also the basis of the MobileNetV2 network.[10] However, the exact architecture was not designed by hand but instead is the result of Neural Architecture Search.[11] This is an optimization procedure that searches for the network architecture with the highest possible accuracy given fixed computational resources to run this network.[12] EfficientNet B4 was trained with dermUS preprocessed images in a commercial platform (Peltarion, Stockholm, Sweden) as prebuilt architecture, with the possibility of a parameter optimization called *snippet*. Augmentation of images was tested and a determination was made not to use it,

TABLE 1 Diagnoses of training dataset and ultrasound criteria for discrimination between benign and malignant lesions

Histological diagnosis	Number	US criteria for benign vs. malignant lesions
Benign	151	–Well circumscribed
Cyst	71	–Symmetric, homogenous appearance
Lipoma	40	–Characteristic findings of a specific benign lesion (e.g., sonographic punctum and tract of epidermal cyst)
Dermatofibroma	25	
Seborrheic keratosis	14	
Malignant	84	–Internal hyperechoic spots
Basal cell carcinoma	43	–Irregular borders
Squamous cell carcinoma	24	–Asymmetric, lobulated, heterogenous appearance
Actinic keratosis	17	–Involvement of adjacent structures (bone, cartilage, muscle)

Note: Echogenicity and vascularity are not sufficient characteristics to discriminate between benign versus malignant lesions alone. While most cutaneous malignancies are hypoechoic dermal or subdermal structures, this is frequently the case for benign lesions as well. Further, basal cell carcinomas (BCCs) show low to moderate vascularity and squamous cell carcinomas (SCCs) tend to show higher vascularity, while benign vascular lesions can also show high flow.

as it did not increase model accuracy. Initial weights were trained on ImageNet.[13]

2.5 | Algorithm training and testing

EfficientNet B4 was trained with the aforementioned dermUS image folder and index csv file assigning randomly 80% of the images for training the model and 20% for algorithm validation. The algorithm was trained for 20 epochs, which represent a whole forward propagation and backward propagation of images for algorithm weights adjustment. Optimal weights were obtained in Epoch 14 with a training time of 93 min. Consistent with best practices suggested by a recent critique of medical DL image interpretation algorithm flaws,[14] 35 new testing images, 5 for each of the 7-training diagnosis, were picked from the same institutional database, unrelated to the initial 235 images.

2.6 | Human expert diagnostic testing and comparison with algorithm diagnosis

A dermatologist certified in dermatologic ultrasound by both Spanish Society of Ultrasound and the European Federation of Ultrasound in Medicine and Biology evaluated the same 35 new images. The dermatologist had more than 5 years of experience in dermatologic ultrasound and was blind to clinical images of the lesion or any other clinical data. The correct diagnosis accuracy of both the algorithm and the dermatologist was compared with χ^2 test, with $p < 0.01$ for statistically significant difference.

3 | RESULTS

3.1 | Internal validation test

The internal test set was generated at random by our system from the original 235 image set and comprised 37 dermUS images of skin tumors (22 benign, 15 malignant). Of the 15 malignant tumors, 12 were classified as malignant by the algorithm (sensitivity, 80%). Of the 22 benign lesions, 19 were predicted as benign by the algorithm (specificity, 86.3%). Of the 22 lesions that the algorithm classified as benign, 19 were benign (negative predictive value (NPV), 86.3%). Of the 15 lesions algorithm classified as malignant, 12 were malignant (positive predictive value (PPV), 80%).

3.2 | External test

The external test set was comprised of 35 new lesion images (20 benign, 15 malignant) with the same diagnostic categories as the validation set (5 for each diagnostic category outlined in Table 1). Of the 15 malignant lesions, 11 were classified as malignant by the algorithm (sensitivity, 73.3%). Of the 20 benign lesions, 16 were predicted as benign by the algorithm (specificity 80%). Of the 20 lesions that the algorithm classified as benign, 16 were benign (NPV, 86.9%). Of the 15 lesions algorithm classified as malignant, 11 were malignant (PPV, 73.3%). Accuracy of DLA for correctly classified lesions as benign or malignant was 77.1%. Figures 1 and 2 are, respectively, examples of benign lesions diagnosed as malignant by DLA (false positive) and malignant lesion diagnosed as benign by DLA (false negative).

With respect to the expert dermUS human evaluator, sensitivity, specificity, PPV, NPV was 73.3%, 75%, 78.9%, and 68.8%, respectively. The accuracy of the human expert was 74.2%, which was not significant different from the DLA accuracy ($p = 0.34$).

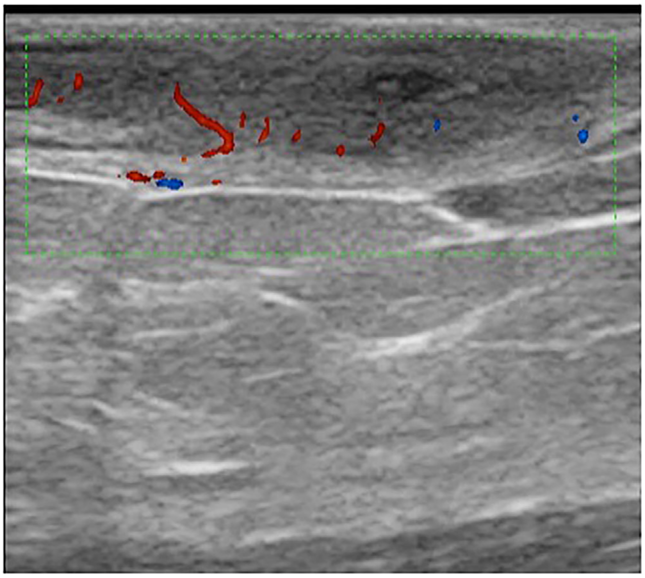

FIGURE 1 False positive; dermatofibroma classified as a malignant lesion

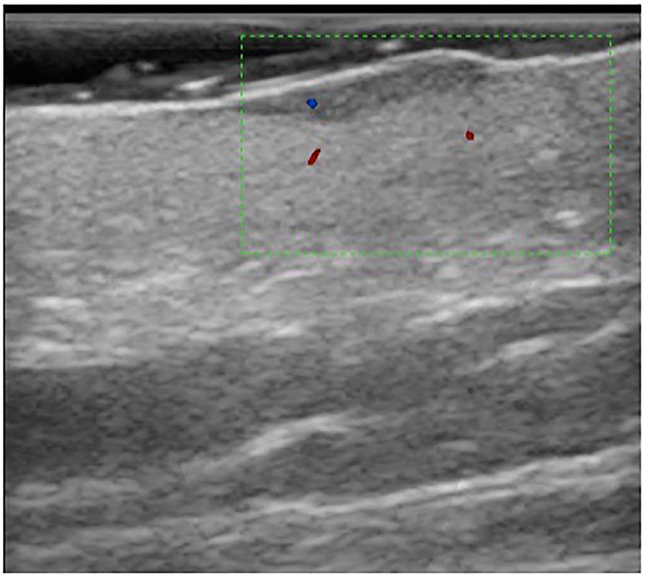

FIGURE 2 False negative; basal cell carcinoma classified as a benign lesion

4 | DISCUSSION

Use of both artificial intelligence and deep learning (machine learning based on patterns identified by neural networks) is increasing in the fields of ultrasound and dermatology. Previous applications in which ultrasound has proven to be helpful in distinguishing benign and malignant lesions include thyroid nodules and breast nodules.[14,15] Regarding dermatology, there have been multiple published "challenges" in which dermatologists compete with data scientists and their algorithms in the diagnosis of dermoscopy images. These have yielded great success for the algorithms when measured against average dermatologists, but not against world class experts.[16]

However, dermatologic ultrasound images have not been addressed as a possible field of deep learning. The superficiality of dermal-epidermal lesions makes their dermUS images more difficult to assess than deep subcutaneous lesions. As these images start being acquired using higher frequency equipment,[17] higher resolution images will yield more information that might be advantageous for building algorithms. Because of this, although most reports on diagnosis use B mode images, we chose to use color Doppler images that contain B mode information (shape, borders, echogenicity) and vascularization pattern information. This information is also used in clinical practice for the differential diagnosis of benign and malignant lesions.[18] It is worth noting that false-positive and false-negative lesions in our study largely demonstrated high vascularity and had ill-defined margins, reflecting that these findings are not always specific to benign or malignant lesions.

DLA selection is as important as data curation. In a recent paper, Blaivas and Blaivas[19] compare different neural networks architectures and demonstrate that performance varies in efficiency (time to train and accuracy) depending on the DL architecture. As these authors state, more recently created architecture is not always more efficient than "older" architectures, and it may be that some architectures are better depending on the type of exploration we are dealing with. In our case we chose EfficientNet B4, a modern architecture, which optimizes parameters with minimum training time in comparison with other more complex architectures. The main advantage of "lighter" architectures is the possibility of more efficient training time and the possibility of transferring them to mobile apps in smartphones or tablets, which have limited computational capability but optimal accessibility.[9–11]

In our study, we used a commercial DL platform instead of one that was homemade or ad hoc from scratch programed algorithms. This provided us the possibility of access to adapting a prebuilt architecture to our dataset, fine-tuning parameters to optimize predictive capability of our DLA. The main advantage of using this kind of no-code or low-code platforms main advantage is accessibility to users without the requirement of a deep knowledge in programming languages.[19].

DLA algorithms are usually pretrained with nonmedical images.[20] Large image databases such as Imagenet[12] permit initial weights preadjustment in order to make borders or other basic features of images more easily recognizable by algorithms.[20] However, in the case of ultrasound, some authors suggest that initial weight adjustments should be made with the same kind of images to improve system accuracy.[21] This is a controversy, which could foster the necessity of creating largely public ultrasound images databases.

Another controversial aspect of training DLAs is the quantity and quality of images necessary. Most publications suggest the use of a large number of images in order to fine-tune algorithms.[20] In our DLA, we used a limited number of images of seven categories and performance was similar to expert diagnosis. The aphorism of "garbage in-garbage out" is in no way improved by increasing the number of inputs without any sequential logic. In that sense, training of specific application DLAs should be done with the help of expert trainers in that application in order to make machine learning more systematic and to interpret failures of DLAs.[20–21]

The main limitations of our study are the limited number of images used, the restricted categories for DLA training (which exclude other diagnoses), and the possibility that other architectures not trained would be more efficient than EfficientNet B4. However, taking into account that these categories are the most frequently diagnosed in general dermatologic ultrasound, and that even with a limited number of images, efficiency for malignancy/benignity diagnosis was comparable to expert malignancy/benignity diagnosis. This exploratory investigation is a starting point for further investigations in this field.

5 | CONCLUSIONS

DLA can use dermatologic ultrasound images as a source for malignancy/benignity diagnosis, even with a limited number of images, with accuracy equivalent to that of a dermatologic ultrasound human expert. Including more diagnostic categories, more cases, and more image features such as epidemiological or clinical variables in optimized architectures, may be key for a future of automated diagnosis in dermatologic ultrasound. However, the supervision of human experts to train, design and control quality of results will also be necessary.

ACKNOWLEDGMENTS

We would like to thank Matteo Barbieri and Andreas Vrålstad from Peltarion for their help, and generous training, project discussion and platform access.

ORCID

Alexandra Laverde-Saad https://orcid.org/0000-0002-7206-9811
Constanza Martínez https://orcid.org/0000-0002-2824-4205
Fernando Alfageme https://orcid.org/0000-0002-0962-9783

REFERENCES

1. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopa VK. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018;392:2388–96.
2. Singh R, Kalra MK, Nitiwarangkul C, Patti JA, Homayounieh F, Padole A, et al. Deep learning in chest radiography: detection of findings and presence of change. PLoS One. 2018;13:e0204155.
3. Fresilli, D, Grani, G, De Pascali, ML, Alagna G, Tassone E, Ramundo V, et al. Computer-aided diagnostic system for thyroid nodule

sonographic evaluation outperforms the specificity of less experienced examiners. J Ultrasound. 2020;23:169–74. https://doi.org/10.1007/s40477-020-00453-y

4. Chiappa, V, Bogani, G, Interlenghi, M, Salvatore C, Bertolina F, Sarpietro G, et al. The adoption of radiomics and machine learning improves the diagnostic processes of women with ovarian masses (the AROMA pilot study). J Ultrasound. 2018. https://doi.org/10.1007/s40477-020-00503-5
5. Wortsman X, Wortsman J. Clinical usefulness of variable-frequency ultrasound in localized lesions of the skin. J Am Acad Dermatol. 2010;62:247–56.
6. Wortsman X, Alfageme F, Roustan G, Arias-Santiago S, Martorell A, Catalano O, et al. Guidelines for performing dermatologic ultrasound examinations by the DERMUS group. J Ultrasound Med. 2016;35:577–80.
7. Wortsman X, Wortsman J. Clinical usefulness of variable-frequency ultrasound in localized lesions of the skin. J Am Acad Dermatol. 2010;62:247–56.
8. Vidal D, Ruiz-Villaverde R, Alfageme F, Roustan G, Mollet J, Ruiz-Carrascosa JC, et al. Use of high frequency ultrasonography in dermatology departments in Spain. Dermatol Online J. 2016;17; 22.
9. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, Long Beach, CA; 2019. p. 6105–14.
10. Sandler M, Howard A., Zhu M, Zhmoginov A, Chen L. MobileNetV2: inverted residuals and linear bottlenecks. Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT; 2018.
11. Tan M, Chen B., Pang R, et al. MnasNet platform-aware neural architecture search for mobile. Conference on Computer Vision and Pattern Recognition, Long Beach, CA; 2019.
12. ImageNet Website. http://image-net.org (2020). Accessed 2 Nov 2020.
13. Blaivas M, Arntfield R, White M. DIY AI, deep learning network development for automated image classification in a point-of-care ultrasound quality assurance program. J Am Coll Emerg Physicians Open. 2020;1:124131.
14. Xu Y, Wang Y, Yuan J, Cheng Q, Wang X, Carson PL. Medical breast ultrasound image segmentation by machine learning. Ultrasonics. 2019;91:1–9.
15. Ko SY, Lee JH, Yoon JH, Na H, Hong E, Han K, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. Head Neck. 2019;41:885–91.
16. Alfageme F, Wortsman X, Catalano O, Roustan G, Crisan M, Crisan D, et al. European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) position statement on dermatologic ultrasound. Ultraschall Med. 2019. https://doi.org/10.1055/a-1161-8872
17. Shokoohi H, LeSaux MA, Roohani YH, Blaivas M. Enhanced point-of-care ultrasound applications by integrating automated feature-learning systems using deep learning. J Ultrasound Med. 2019;38:1887–97
18. Crisan D, Crisan M, Badea R, Rastian I. Integrative analysis of cutaneous skin tumours using ultrasonogaphic criteria. Preliminary results. Medical Ultrasonography. 2014;16:285–90.
19. Blaivas M, Blaivas L. Are all deep learning architectures alike for point-of-care ultrasound? Evidence from a cardiac image classification model suggests otherwise. J Ultrasound Med. 2020;39:1187–94
20. Miner R. Developing an AI project J Med Imaging Radiat Sci. 2020. https://doi.org10.1016/j.jmir.2020.06.010
21. Blaivas L, Blaivas M. Are convolutional neural networks trained on ImageNet images wearing rose-colored glasses? A quantitative comparison of ImageNet, computed tomographic, magnetic resonance, chest X-ray, and point-of-care ultrasound images for quality. J Ultrasound Med. 2020. https://doi.org10.1016/j.jmir.2020.06.010

How to cite this article: Laverde-Saad A., Jfri A, García R, Salgüero I, Martínez C, Cembrero H, et al. Discriminative deep learning-based benignity/malignancy diagnosis of dermatologic ultrasound skin lesions with pretrained artificial intelligence architecture. Skin Res Technol. 2022;28:35–39. https://doi.org/10.1111/srt.13086