Abstract

Assessing Generalizability of Deep Learning Models Trained on Standardized and Nonstandardized Images and Their Performance Against Teledermatologists

Ibukun Oloruntoba[1]; Toan D Nguyen[2], PhD; Zongyuan Ge[3], PhD; Tine Vestergaard[4], MD, PhD; Victoria Mar[5], MBBS, FACD, PhD

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

[2]Monash University, Melbourne, Australia

[3]Monash eResearch Centre, Monash University, Melbourne, Australia

[4]Department of Dermatology and Allergy Centre, Odense University Hospital, Odense, Denmark

[5]Victorian Melanoma Service, Alfred Health and School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

Corresponding Author:
Ibukun Oloruntoba
School of Public Health and Preventive Medicine
Monash University
Wellington Rd
Clayton
Melbourne, VIC 3800
Australia
Phone: 61 3 9905 4000
Email: aolo0001@student.monash.edu

Abstract

Background: Convolutional neural networks (CNNs) are a type of artificial intelligence that show promise as a diagnostic aid for skin cancer. However, the majority are trained using retrospective image data sets of varying quality and image capture standardization.

Objective: The aim of our study is to use CNN models with the same architecture, but different training image sets, and test variability in performance when classifying skin cancer images in different populations, acquired with different devices. Additionally, we wanted to assess the performance of the models against Danish teledermatologists when tested on images acquired from Denmark.

Methods: Three CNNs with the same architecture were trained. CNN-NS was trained on 25,331 nonstandardized images taken from the International Skin Imaging Collaboration using different image capture devices. CNN-S was trained on 235,268 standardized images, and CNN-S2 was trained on 25,331 standardized images (matched for number and classes of training images to CNN-NS). Both standardized data sets (CNN-S and CNN-S2) were provided by Molemap using the same image capture device. A total of 495 Danish patients with 569 images of skin lesions predominantly involving Fitzpatrick skin types II and III were used to test the performance of the models. Four teledermatologists independently diagnosed and assessed the images taken of the lesions. Primary outcome measures were sensitivity, specificity, and area under the curve of the receiver operating characteristic (AUROC).

Results: A total of 569 images were taken from 495 patients (n=280, 57% women, n=215, 43% men; mean age 55, SD 17 years) for this study. On these images, CNN-S achieved an AUROC of 0.861 (95% CI 0.830-0.889; *P*<.001), and CNN-S2 achieved an AUROC of 0.831 (95% CI 0.798-0.861; *P*=.009), with both outperforming CNN-NS, which achieved an AUROC of 0.759 (95% CI 0.722-0.794; *P*<.001; *P*=.009). When the CNNs were matched to the mean sensitivity and specificity of the teledermatologists, the model's resultant sensitivities and specificities were surpassed by the teledermatologists. However, when compared to CNN-S, the differences were not statistically significant (*P*=.10; *P*=.05). Performance across all CNN models and teledermatologists was influenced by the image quality.

Conclusions: CNNs trained on standardized images had improved performance and therefore greater generalizability in skin cancer classification when applied to an unseen data set. This is an important consideration for future algorithm development,

regulation, and approval. Further, when tested on these unseen test images, the teledermatologists *clinically* outperformed all the CNN models; however, the difference was deemed to be statistically insignificant when compared to CNN-S.

Conflicts of Interest: VM received speakers fees from Merck, Eli Lily, Novartis and Bristol Myers Squibb. VM is the principal investigator for a clinical trial funded by the Victorian Department of Health and Human Services with 1:1 contribution from MoleMap.

(iproc 2021;7(1):e35391) doi: 10.2196/35391

KEYWORDS

teledermatology; CNN; artificial intelligence; skin cancer; Denmark; Australia; New Zealand; image standardization; generalizability; classification

Multimedia Appendix 1

Receiver operating characteristic (ROC) curves for the three convolutional neural network (CNN) models and the performances of the teledermatologists on the Danish test set. The ROC and the area under the curve of the ROC of the CNN models in relation to the sensitivity and 1-specificity of the teledermatologists when tested on the 569 Danish test images. The teledermatologist's performance was greater than all of the CNN models.
[PNG File , 341 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Table 1: Sensitivity and specificity of the convolutional neural network models when matched to the average performance of the teledermatologists.
[PNG File , 398 KB-Multimedia Appendix 2]

Abbreviations

AUROC: area under the curve of the receiver operating characteristic
CNN: convolutional neural network

Edited by T Derrick; this is a non–peer-reviewed article. Submitted 02.12.21; accepted 03.12.21; published 10.12.21.

Please cite as:
Oloruntoba I, Nguyen TD, Ge Z, Vestergaard T, Mar V
Assessing Generalizability of Deep Learning Models Trained on Standardized and Nonstandardized Images and Their Performance Against Teledermatologists
iproc 2021;7(1):e35391
URL: https://www.iproc.org/2021/1/e35391
doi: 10.2196/35391
PMID: 27762282