Received: 16 April 2019 | Accepted: 28 April 2019

DOI: 10.1111/srt.12726

ORIGINAL ARTICLE

WILEY

Data augmentation in dermatology image recognition using machine learning

1st Lt. Pushkar Aggarwal

University of Cincinnati, Cincinnati, Ohio

Correspondence
1st Lt. Pushkar Aggarwal, University of Cincinnati, 2545 Dennis St #7105, Cincinnati, OH – 45219.
Email: aggarwpr@mail.uc.edu

Abstract

Background: Each year in the United States, over 80 million people are affected by acne, atopic dermatitis, rosacea, psoriasis, and impetigo. Artificial intelligence and machine learning could prove to be a good tool for assisting in the diagnosis of dermatological conditions. The objective of this study was to evaluate the use of data augmentation in machine learning image recognition of five dermatological disease manifestations—acne, atopic dermatitis, impetigo, psoriasis, and rosacea.

Materials and Methods: Open-source dermatological images were gathered and used to retrain TensorFlow Inception version-3. Retraining was done twice—once with and once without data augmentation. Both models were tested with the same images, and R software was used to perform statistical analysis.

Results: The average of each of the statistical measures (sensitivity, specificity, PPV, NPN, MCC, and F1 Score) increased when data augmentation was added to the model. In particular, the average Matthews correlation coefficient increased by 7.7%. Each of the five dermatological manifestations had an increase in area under the curve (AUC) after data augmentation with the average increase in AUC of 0.132 and a standard deviation of 0.033. Atopic dermatitis had the highest increase in AUC of 0.18. With data augmentation, the lowest AUC was 0.87 for psoriasis and the highest was 0.97 for acne, indicating that the model performs well.

Conclusion: With a deep learning-based approach, it is possible to differentiate dermatological images with appreciable MCC, F1 score, and AUC. Further, data augmentation can be used to increase the model's accuracy by a significant amount.

KEYWORDS

acne, artificial intelligence, atopic dermatitis, image recognition, impetigo, psoriasis, rosacea

1 | INTRODUCTION

Almost 85 million Americans (about 27% of the population) were evaluated by a physician for at least one skin disease in 2013. In people 65 years and older, the prevalence of skin diseases increases to 50% with an average of slightly more than two skin diseases per person. The cost for work and life productivity loss for skin diseases was estimated to be more than $10 billion annually.[1] Studies also indicate that there is a shortage of dermatologists.[2,3] With a projected growth in the US population of those over 65 years of age by about 30 million by 2030 and a projected minimal growth in number of physicians, the patient to dermatologist ratio imbalance is likely to grow.[1]

The use of technology to assist physicians in diagnoses and treatments is not new. The adoption of an electronic health record, which includes clinical decision support, computerized physician order entry, and health information exchange by a majority of hospitals in the past two decades has resulted in a reduction in medical errors

from 18.2% to 8.2% in an outpatient setting and a 24% reduction in redundant laboratory tests in a hospital setting.[4-6] Remote monitoring tools such as pacemakers that automatically transmit data to healthcare providers have resulted in a reduction in readmission of patients with heart disease from 25% to 2%.[7] Recently, an automated system has been implemented that analyzes a patient's daily electronic health record looking for signs of sepsis including abnormalities in temperature, blood pressure, and white blood cell count. If the system finds these abnormalities, it can alert the nurse to evaluate the patient for sepsis.[8] This system can be lifesaving to patients and can serve as an extra set of eyes in the hospital. The above and many more technological implementations have led to improved efficiency and accuracy by physicians and overall improved outcome for the patient.

Artificial intelligence and machine learning have made strides in various fields. However, its nascent presence in the field of medicine has abundant room to grow. Specifically, image recognition using neural networks can prove to be a significant aid to physicians in both their accuracy of diagnosis and overall efficiency. Deep convolutional neural networks consist of an architecture that resembles the organization of the human brain and as such can be used to process and analyze data in a similar manner as humans.[9]

Machine learning for image recognition in the field of dermatology has been previously examined.[10] The next step in this area is to develop methods to improve the image recognition accuracy. The objective of this study was to evaluate the use of data augmentation on machine learning image recognition accuracy of five dermatological diseases manifestations—acne, rosacea, psoriasis, atopic dermatitis, and impetigo. Each year in the United States, acne affects about 50 million people, atopic dermatitis affects about 28 million people, rosacea affects about 16 million people, psoriasis affects about 7.5 million people, and impetigo affects about 3 million people.[11-13] These five diseases were chosen for their fairly high incidence in the population and for their potential indiscernibility as they have similarity in clinical skin presentation.

2 | MATERIALS AND METHODS

2.1 | Collection of images

Open-source dermatological images captured through DermNet NZ,[14] Dermatology Atlas,[15] Hellenic Dermatological Atlas,[16] and Google Images were downloaded. After compiling these images, each image was analyzed and cropped to remove suspected noise. Examples of noise included visible clothing on unaffected skin areas or a background object. During this image selection process, some images were also discarded due to insufficient image size or poor focus. After the above image processing and stratifying, a total of 332, 92, 138, 280, and 96 images for acne, atopic dermatitis, impetigo, psoriasis, and rosacea, respectively, were selected for model development, testing, and evaluation using a neural network. 30 images from each dermatological manifestation image total were set aside for the testing phase.

2.2 | Deep convolutional neural network

TensorFlow™,[17] an open-source software library by Google, was used as a deep learning framework and was used to retrain Inception version 3 (v-3). Inception (v-3) is a deep convolutional neural network. This neural network consists of a hierarchy of multiple computational layers that each has an input and output. For image recognition, the first layers of the neural network process basic information such as lines and curves while the higher layers in the hierarchy process complex and abstract information.[18] All layers except the final layer of this neural network are pre-trained with more than 1.2 million images. The final layer of the neural network was retrained with the gathered dermatological images. During the retraining process, the neural network underwent both a training and validation step. In the training step, the inputted images were used to train the neural network. In the validation step, inputted naïve images that had not been seen previously by the neural network were used to iteratively assess training accuracy.[19]

2.3 | Statistical analysis

After the model had been retrained (trained and validated), a user-inputted testing/assessment step was performed in which thirty images for each of the five dermatological manifestations were inputted and the results were statistically analyzed. The program assessment output is percentages of the probability of each of the dermatological manifestations for each testing image inputted. R software (2017)[20] was used to perform the statistical analysis. Sensitivity, specificity, positive predictive value, negative predictive value, Matthews correlation coefficient (MCC), and F1 score were calculated for each dermatological manifestation. The F1 score is the harmonic average of the sensitivity and positive predictive value. MCC takes into account true positives and negatives and false positive and negatives and is a balanced measure even if the classes are different sizes. MCC is widely used as a performance metric in bioinformatics.[21]

All statistical values calculated except MCC have a range from zero to one with a better model having a value closer to 1. MCC ranges from −1 to 1 with −1 indicating a complete disagreement between prediction and correct answer, 0 indicating randomness, and 1 indicating complete agreement between prediction and correct answer.

In addition, receiver operating curves were generated for each of the five dermatological manifestations based on incremental increases in the percentage cutoff for the correct dermatological manifestation in the model output. Areas under the curve were calculated for each of these curves to quantify the model's performance. A confusion matrix was constructed from the results of the testing images. The diagonal cells in the matrix contain the recall/sensitivity value, which is the true positives divided by the sum of the true positives and false negatives. The remaining cells show what percent of the 30 testing images for each dermatological

manifestation were incorrectly labeled, stratified by the remaining dermatological manifestations.

2.4 | Overfitting and data augmentation

Overfitting is one challenge present in machine learning. When a deep convolutional neural network overfits, it works extremely well on training data but poorly on data it has never seen before. This is especially important in the field of dermatology because of the variability that exists in the images that the neural network will be analyzing. For example, each image of a dermatological manifestation will have variability in the grid location of the dermatological manifestation, angle at which the image is taken, and the size of the dermatological manifestation. Two steps were taken to reduce overfitting. First, a dropout layer was added and set to 0.5. This results in 50% of the neurons to be randomly turned off during the training process and therefore reduce the likelihood of overfitting.

The second step taken to reduce overfitting was to use data augmentation. In data augmentation, the images are modified to account for some of the variability that exists in image taking. To account for grid location and size of the dermatological manifestation and the angle of the image, the training images given to the model were altered using rotation, zoom, shear, and horizontal and vertical flipping. Not only does this reduce overfitting, but it also increases the number of training images which was especially important in this analysis because of the limited number of images that were selected for input into the model. The model was run once without data augmentation but with dropout and once with data augmentation and dropout.

3 | RESULTS

2 × 2 tables were constructed based on the output of the models during the testing phase. The statistical measures calculated using these 2 × 2 tables is displayed in Table 1. The average of each of the statistical measures (sensitivity, specificity, PPV, NPN, MCC, and F1 Score) increased when data augmentation was added to the model. In particular, the average MCC increased by 7.7%. Further, there was an increase in MCC and F1 score for each of the five dermatological manifestations with data augmentation. Atopic dermatitis, which had the lowest MCC and F1 score without data augmentation, had the highest increase in MCC and F1 score with data augmentation. Since atopic dermatitis had the fewest number of inputted images, these results may indicate that data augmentation is especially beneficial when the number of training images is limited.

The confusion matrices for the model without and with data augmentation are shown in Figures 1 and 2, respectively. Psoriasis and atopic dermatitis were mixed up most often in the neural network. Specifically, atopic dermatitis images were labeled as psoriasis in 33% of the tested atopic dermatitis images and psoriasis images were labeled as atopic dermatitis in 27% of the tested psoriasis images when the model was run without data augmentation. These results make sense as psoriasis and atopic dermatitis have very similar physical findings. With data augmentation, the 33% from above decreased to 30% while the 27% stayed the same. The number of correctly labeled images for both atopic dermatitis and psoriasis increased with data augmentation. In fact, for all dermatological conditions except acne, the number of correctly labeled images increased while the number of incorrectly labeled images decreased. Acne had a relatively high percentage of correctly labeled images (73%) without data augmentation, and this percentage stayed the same with data augmentation.

Receiver operating characteristic curves for each of the dermatological manifestations with and without data augmentation are shown in Figure 3. The average area under the curve (AUC), which can range from 0 to 1, was also calculated. Each of the five dermatological manifestations had an increase in area under the curve after data augmentation. The average increase in AUC was 0.132 with a

TABLE 1 Statistical analysis for each dermatological condition with and without data augmentation

	Sensitivity	Specificity	PPV	NPV	MCC	F1 score
Without data augmentation						
Acne	0.733	0.900	0.647	0.931	0.605	0.688
Atopic Dermatitis	0.500	0.875	0.500	0.874	0.375	0.500
Impetigo	0.567	0.933	0.680	0.896	0.537	0.618
Psoriasis	0.633	0.875	0.559	0.905	0.486	0.594
Rosacea	0.567	0.892	0.567	0.892	0.458	0.567
Average ± SD	0.600 ± 0.079	0.895 ± 0.021	0.590 ± 0.065	0.900 ± 0.019	0.492 ± 0.077	0.593 ± 0.062
With data augmentation						
Acne	0.733	0.950	0.786	0.934	0.701	0.759
Atopic Dermatitis	0.633	0.875	0.559	0.905	0.486	0.594
Impetigo	0.633	0.933	0.704	0.911	0.590	0.667
Psoriasis	0.667	0.892	0.606	0.915	0.539	0.635
Rosacea	0.600	0.917	0.643	0.902	0.530	0.621
Average ± SD	0.653 ± 0.045	0.913 ± 0.027	0.660 ± 0.079	0.913 ± 0.011	0.569 ± 0.074	0.655 ± 0.057

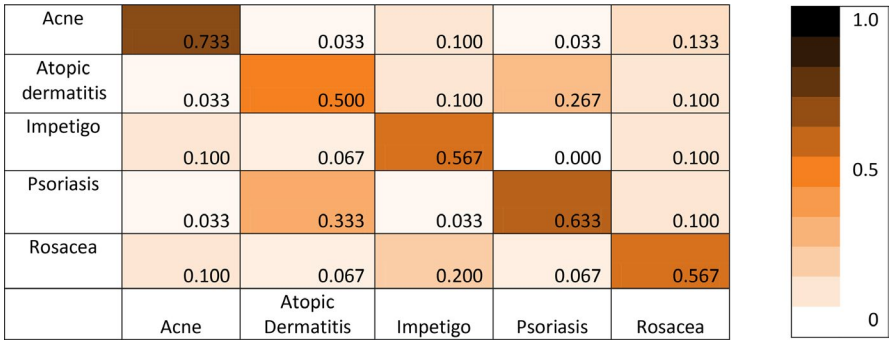

FIGURE 1 Confusion matrix without data augmentation [Colour figure can be viewed at wileyonlinelibrary.com]

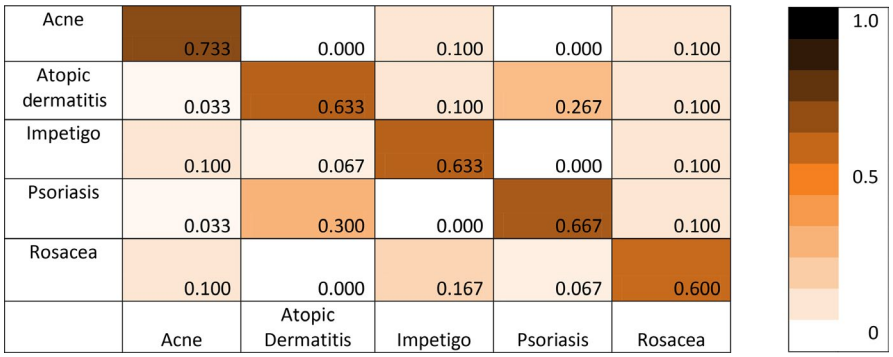

FIGURE 2 Confusion matrix with data augmentation [Colour figure can be viewed at wileyonlinelibrary.com]

standard deviation of 0.033. Atopic dermatitis had the highest increase in AUC of 0.18. With data augmentation, the lowest AUC was 0.87 for psoriasis and the highest was 0.97 for acne, indicating that the model performs well.

4 | DISCUSSION

Overall, the statistical measures, the confounding matrix, and the receiver operating characteristic curves show significant improvements in the accuracy of the model for all five dermatological manifestations with data augmentation. Data augmentation resulted in the largest improvement in atopic dermatitis accuracy which may be because atopic dermatitis had the fewest number of inputted training images. As such, data augmentation can be especially useful when there are a limited number of training images available. The confusion between atopic dermatitis and psoriasis was slightly improved with data augmentation. In order to further decrease confusion between such manifestations that have very similar physical findings, it is likely that increasing the number of training images is needed.

High-level computations, deep learning algorithms, and large image data sets can help exceed or equal human recognition of images. These capabilities are being used by law and immigration authorities to match thumbprints, biometrics, and faces. As per a literature review on PubMed,[22] machine learning has been used to identify skin cancers and was shown to be on par with board-certified dermatologists' identification of the same images.[23] In this skin cancer analysis, data augmentation techniques such as rotation and flipping were used but not zoom and shear. No deep learning approach was found to decipher dermatological diseases based on the images of acne, atopic dermatitis, impetigo, psoriasis, and rosacea.

This approach of image classification using artificial intelligence can also be extended to other dermatological diseases having skin manifestations. A database of dermatologic images with a good number of high-resolution images will be beneficial in developing dermatological image recognition application, though it will still be daunted with development challenges. Automated classification of skin manifestations using images is a cumbersome task owing to variability in the images due to different zoom levels, single or multiple manifestations in the image, no information on the body part of the manifestation, resolution of an image and angle, and lighting of the image, in addition to the disease variability/progression. The negative effect of some of these on the model can be reduced with data augmentation. Further symptoms such as itching, pain, and other clinical symptoms are absent in the image analysis which can help the dermatologist in diagnosing the disease. Deep neural networks are notoriously difficult to train, and overfitting to data is a major challenge, since they are nonlinear and have many parameters.[24]

The model application in present form is unable to decipher the severity of the diseases, pathology and disease progression. In future, the model will be amended to have the capability of deciphering the severity and progression of the disease based on the skin manifestations. For deciphering pathology, pathological images having high resolution will be needed and would need to be trained.

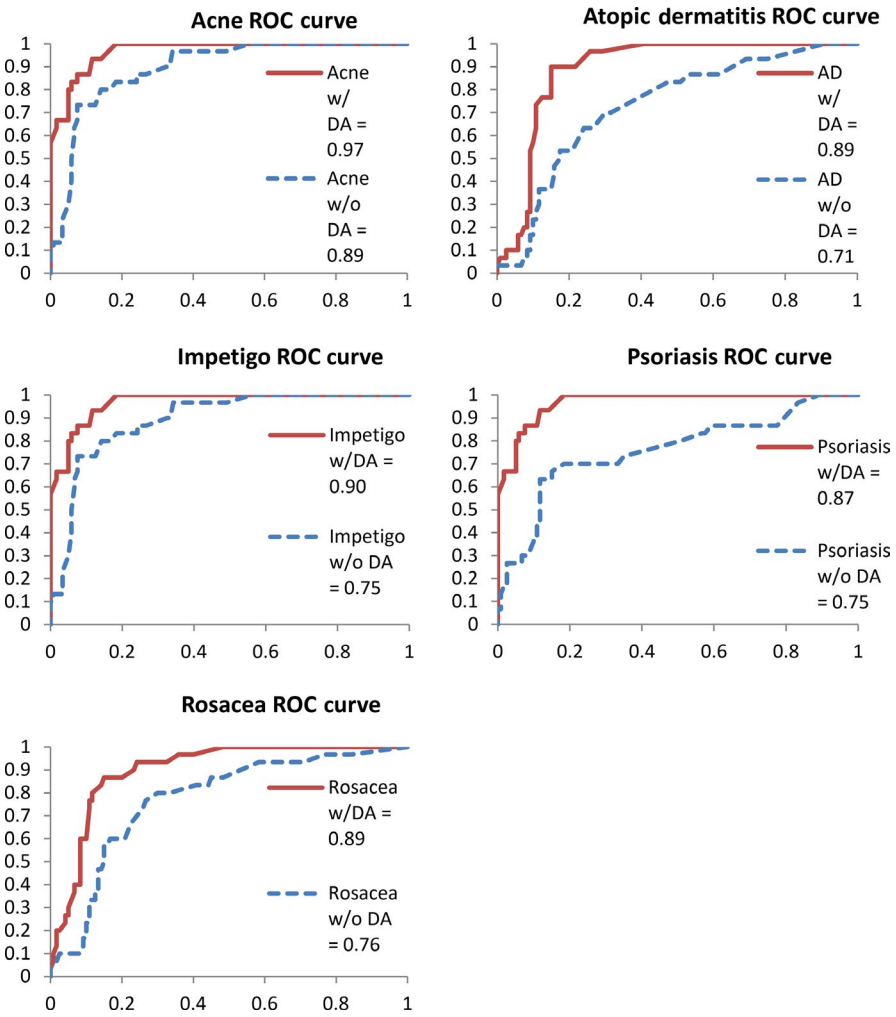

FIGURE 3 Receiver operating curves and area under the curves with and without data augmentation (DA) for each dermatological condition [Colour figure can be viewed at wileyonlinelibrary.com]

5 | CONCLUSION

With a deep learning-based approach using TensorFlow™[17] and Inception Version 3, it is possible to differentiate dermatological images with appreciable MCC, F1 score, and AUC. Using data augmentation, the MCC, F1 score, and AUC can be increased by a significant amount. Data augmentation along with a larger image database containing higher resolution images and along with a clinical support system which includes symptoms and the place/s of occurrence of the skin symptoms on the body can help in deciphering the image with even higher probability. The model can support clinical decisions of dermatologists and give them an extra pair of eyes for diagnosis of the dermatological disease. Future model of the application can monitor the progression of the disease before and after treatment/s.

CONFLICTS OF INTEREST

There was no requirement of IRB as the images were not identified to a person. The author has no ethical conflicts to disclose. The author has no conflicts of interest to declare.

ORCID

1st Lt. Pushkar Aggarwal https://orcid.org/0000-0001-7549-8876

REFERENCES

1. Lim HW, Collins S, Resneck JS, et al. The burden of skin disease in the United States. *J Am AcadDermatol*. 2017;76(5):958-972.e2.
2. Industry Trends Indicate A Shortage of Dermatologists. *The Florida Society of Dermatology & Dermatologic Surgery*. https://fsdds.org/industry-trends-indicate-a-shortage-of-dermatologists/. Published October 4, 2017. Accessed February 4, 2019.
3. Suneja T, Smith ED, Chen GJ, Zipperstein KJ, Fleischer, Jr AB, Feldman SR. Waiting times to see a dermatologist are perceived as too long by dermatologists: implications for the dermatology workforce. *Arch Dermatol*. 2001;137(10):1303-1307.
4. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. *Risk ManagHealthc Policy*. 2011;4:47-55.
5. Devine EB, Hansen RN, Wilson-Norton JL, et al. The impact of computerized provider order entry on medication errors in a multispecialty group practice. *J Am Med Inform Assoc*. 2010;17(1):78-84.
6. Bates DW, Kuperman GJ, Rittenberg E, et al. A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. *Am J Med*. 1999;106(2):144-150.

7. Biggest Technological Advancements for Healthcare in the Last Decade. *Becker's Hospital Review*. https://www.beckershospitalreview.com/healthcare-information-technology/10-biggest-technological-advancements-for-healthcare-in-the-last-decade.html. Published January 28, 2014. Accessed February 4, 2019.
8. Harris R. *Synergy Between Nurses And Automation Could Be Key To Finding Sepsis Early*. NPR. https://www.npr.org/sections/health-shots/2018/02/22/583846656/synergy-between-nurses-and-automation-could-be-key-to-finding-sepsis-early. Published February 22, 2018. Accessed February 4, 2019.
9. Yates EJ, Yates LC, Harvey H. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *ClinRadiol*. 2018;73(9):827-831.
10. Singh N, Gupta SK. Recent advancement in the early detection of melanoma using computerized tools: An image analysis perspective. *Skin Res Technol*. 2019;25(2):129-141.
11. Gaille B. *23 Dermatology Industry Statistics and Trends*. *BrandonGaille*. https://brandongaille.com/23-dermatology-industry-statistics-and-trends/. Published June 12, 2018. Accessed February 4, 2019.
12. Skin conditions by the numbers. *American Academy of Dermatology*. https://www.aad.org/media/stats/conditions/skin-conditions-by-the-numbers. Accessed February 4, 2019.
13. How to Treat Impetigo and Control This Common Skin Infection. *US Food and Drug Administration*. https://www.fda.gov/ForConsumers/ConsumerUpdates/ucm048837.htm. Published November 1, 2016. Accessed February 4, 2019.
14. Dermatology Image Library. *DermNet NZ*. https://www.dermnetnz.org/image-library/. Accessed February 4, 2019.
15. da Silva SF. *Dermatology Atlas*. http://www.atlasdermatologico.com.br/. Accessed February 4, 2019.
16. Verros CD. *Hellenic Dermatology Atlas*. http://www.hellenicdermatlas.com/en/. Accessed February 4, 2019.
17. Abadi M, Barham P, Chen J, et al. *TensorFlow: A system for large-scale machine learning*. arXiv:1605.08695.
18. Rampasek L, Goldenberg A. TensorFlow: biology's gateway to deep learning? *Cell Syst*. 2016;2(1):12-14.
19. Zhang YC, Kagen AC. Machine learning interface for medical image analysis. *J Digit Imaging*. 2016;30(5):615-621.
20. R Core Team. *R A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
21. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE*. 2017;12(6):e0177678.
22. US National Library of Medicine, *National Institutes of Health*. PubMed [database]. http://www.ncbi.nlm.nih.gov/pubmed. Accessed December 31, 2018.
23. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.
24. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *MolSyst Biol*. 2016;12(7):878.

How to cite this article: Aggarwal 1LP. Data augmentation in dermatology image recognition using machine learning. *Skin Res Technol*. 2019;25:815–820. https://doi.org/10.1111/srt.12726