Automated grading of acne vulgaris by deep learning with convolutional neural networks

Ziying Vanessa Lim*1, Farhan Akram*2, Cuong Phuc Ngo2,3, Amadeus Aristo Winarto2,3, Wei Qing Lee4, Kaicheng Liang2, Hazel Hweeboon Oon1, Steven Tien Guan Thng1,5, Hwee Kuan Lee2,4,6,7

1 National Skin Centre, Singapore
2 Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore
3 Hwa Chong Institution, Singapore
4 School of Computing, National University of Singapore, Singapore
5 Skin Research Institute Singapore, A*STAR, Singapore
6 Image and Pervasive Access Lab, CNRS, Singapore
7 Singapore Eye Research Institute, Singapore

*Joint authorship: these authors contributed equally to the manuscript

Acknowledgments

The authors wish to acknowledge the following for their contributions: research coordinators Jasmine Lee, Siti Umairah Mohammed Adnan, Veron Lu, Yan Qin Mak, and Jiesi Kwa from the National Skin Centre, Singapore; and Nurjehan Binte Mohamad Yasin from the Skin Research Institute of Singapore for her work on labeling.

Running head: Automated grading of acne by deep learning

Manuscript word count: 1,967
Table count: 1
Figure count: 5

Funding: None

Conflicts of interest: HH Oon has served as a speaker, advisory board member and researcher for Galderma, is a clinical investigator for Janssen, Novartis and Pfizer, and an advisory board member for AbbVie, Eli Lilly, Janssen, Novartis and LEO Pharma.
Corresponding authors:
Ziying Vanessa Lim
Address: National Skin Centre
1 Mandalay Road, Singapore 308205
Telephone: +65 6253 4455
Email: drvanessalim@gmail.com

Farhan Akram
Address: Bioinformatics Institute, A*STAR
30 Biopolis Street, #07-01 Matrix
Singapore 138671
Telephone: +65 6478 8298
Email: farhan@vim.cau.ac.kr
Abstract

Background
The visual assessment and severity grading of acne vulgaris by physicians can be subjective, resulting in inter- and intra-observer variability.

Objective
To develop and validate an algorithm for the automated calculation of the Investigator’s Global Assessment (IGA) scale, to standardize acne severity and outcome measurements.

Materials and Methods
472 photographs (retrieved 01/01/2004-04/08/2017) in the frontal view from 416 acne patients were used for training and testing. Photographs were labeled according to the IGA scale in three groups of IGA clear/almost clear (0-1), IGA mild (2), and IGA moderate-to-severe (3-4). The classification model used a convolutional neural network, and models were separately trained on three image sizes. The photographs were then subjected to analysis by the algorithm, and the generated automated IGA scores were compared to clinical scoring. The prediction accuracy of each IGA grade label, and the agreement (Pearson correlation) of the two scores were computed.

Results
The best classification accuracy was 67%. Pearson correlation between machine-predicted score and human labels (clinical scoring and researcher scoring) for each model and various image input sizes was 0.77. Correlation of predictions with clinical scores was highest when using Inception v4 on the largest image size of 1200x1600. Two sets of human labels showed a high correlation of 0.77, verifying the repeatability of the ground truth labels. Confusion matrices show that the models performed sub-optimally on the IGA 2 label.

Conclusion
Deep learning techniques harnessing high-resolution images and large datasets will continue to improve, demonstrating growing potential for automated clinical image analysis and grading.

Keywords: acne vulgaris, severity grading, automated grading, deep learning, convolutional neural network
Introduction

Acne vulgaris is a common disorder affecting approximately 9% of the global population, most commonly in the younger age groups. Facial acne and the scarring that results from it may negatively impact self-esteem and quality of life.

One of the methods currently used for the evaluation of acne severity is the Investigator’s Global Assessment (IGA) with five ordinal grades (clear, almost clear, mild, moderate, and severe). In order to derive this score, a physician needs to assess the lesion types, their quantity and the density of involvement. This manual method can be time consuming and tedious to perform in a busy clinic consultation. As this grading method relies on approximation of the number of lesions, it results in a subjective assessment with both inter and intra-observer variability; a European study examining inter-observer variability of a 3-point acne scale showed very poor agreement between readers (kappa value between 0.2 and 0.6), with only the most severe cases agreed upon. It is important to accurately assess acne grade to evaluate its severity as this influences treatment options and assessment of response to therapy.

A standardized, reproducible way to assess acne and their response to treatment would allow patients to self-monitor through mobile applications, physicians to track progress remotely through tele-dermatology, and research studies to be compared with each other in meta-analyses. The use of digital image analysis has proven useful in the monitoring of several other skin conditions such as melasma and vitiligo and deep learning has already had an impact in the automated diagnosis of skin cancers. This study leverages on the interdisciplinary collaboration between clinician dermatologists and bioinformatics researchers to develop an algorithm to automatically calculate IGA using digital image analysis for the assessment of acne severity and monitoring of treatment outcomes.

Objective

As computer image analysis of facial photographs would be an easier and more reproducible way of assessing the extent and severity of facial acne as opposed to the traditional method of estimation by clinicians, the aim of this study is to develop and validate an algorithm for the automated calculation of the IGA scale, so as to enable clinicians to standardize acne severity and outcome measurements.
Methods

The algorithm for the software was developed in collaboration between dermatologists from the National Skin Centre (NSC), Singapore and image analysis experts from the A*STAR Bioinformatics Institute, Singapore. This study was approved by the National Healthcare Group ethics board (study number: 2017/00719).

Based on the International Classification of Diseases (ICD) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) coded diagnoses of acne vulgaris, retrospective retrieval of all facial photography images of patients with varying severities of facial acne from the period between 01/01/2004 to 04/08/2017 was performed. There were no exclusion criteria. Due to patient confidentiality requirements, detailed demographic data was not collected. The distinguishing facial features of the subjects in the photographs were digitally anonymized so as to maintain patient confidentiality.

A total of 474 frontal facial images from 418 patients were obtained (Fig. 1). Two images were discarded due to the presence of confounding skin conditions, leaving 472 images from 416 patients. The training data used to produce training and validation images consisted of 374 images, and the test set consisted of 98 images. There was no overlap of patients between training and test sets, i.e. patients with multiple images were allocated to either training or test set, not both. Images were of pixel dimensions of ~2000x3000 pixels. The data was divided into three groups, namely IGA Grades 0-1, IGA Grade 2, and IGA Grades 3-4. This grouping was motivated by treatment being similar within each group of IGA grades. Labeling was performed separately by two student trainees and one researcher and then combined by majority vote, with close supervision and verification by a dermatology resident (clinical scoring. Clinical scoring was used as the ground truth for model training. A separate set of labeling was performed by three researchers trained by the dermatology resident (researcher scoring) and used as an alternate set of human labels to validate the ground truth. The image background from both initial training and test sets was removed (‘masking’) (Fig. 2). The initial training set of 374 images was further divided into training and validation set of 314 and 60 images, respectively. In the training set, the groups IGA Grades 0-1, IGA Grade 2, and IGA Grades 3-4 consisted of 131, 132, and 51 images respectively, showing a numerical
imbalance between three groups. In the validation set, each IGA group contained 20 images. In order to minimize the imbalance between groups, which could affect classification performance, both training and validation sets were augmented (Fig. 1). Images were augmented by controlled small distortions of training images with random combinations of left-right flipping, cropping, contrast adjustment, intensity scaling, and shifting/scaling. By using data augmentation to generate transformed images similar to real images, the training set of 314 was increased to 6,248, in which the number of images in IGA Grades 0-1, IGA Grade 2, and IGA Grades 3-4 groups were 2,096, 2,112, and 2,040, respectively. The images in the validation set were also augmented from 60 to 1,200, with each group containing 400 images. The validation set was used to evaluate model performance and hyperparameters during the development process. The test set was used to generate final performance metrics (Fig. 2).

The classification model used a convolutional neural network (CNN), a deep learning architecture that mimics human vision. Convolution mathematical operations are used to extract meaningful features, such as edges, textures and colors, in images. Neural network models inspired by three high performing architectures, DenseNet\textsuperscript{15}, Inception v4\textsuperscript{16}, and ResNet18\textsuperscript{17} were evaluated. Our network architectures are compact versions (with reduced depth) of the original models. (Refer Supplementary Information for details.) Models were separately trained on three image sizes: 600x800 pixels, 750x1000 pixels, and 1200x1600 pixels from 'scratch', to investigate the dependence of model performance on image size. The models that performed best on validation data were saved.

The clinical photographs were subjected to analysis by the algorithm, and the generated automated IGA scores were compared to the clinical scoring. The prediction accuracy of each IGA grade label, and the agreement (Pearson correlation) of the two scores were computed.

As a comparison to techniques previously described in the literature, an acne analysis algorithm based on deep learning recently reported by Microsoft and Nestle Skin Health\textsuperscript{18} was also trained on our dataset, and used to generate performance metrics for comparison with our approach.
Results

The best classification accuracy was 67% from the Inception v4 model trained on either image size 1200x1600 or 600x800 pixels (Table 1). Predicted accuracies for each label (IGA Grades 0-1, IGA Grade 2, and IGA Grades 3-4), as well as performance by the baseline Microsoft-Nestle algorithm, are reflected in the confusion matrices in Figure 3. The Pearson correlations between predictions and clinical scoring are also presented (Figure 4).

The largest image inputs gave the best performance for DenseNet and Inception v4, suggesting that the relatively minute features of individual acne lesions/clusters may be best captured by higher pixel densities. Image size of 1200x1600 pixels was the largest reasonable size possible given the image sizes available in our dataset and computer memory limitations in our present setup. Increased neural network complexity (Inception v4 vs DenseNet or ResNet18) also appeared to perform better. Figure 5 highlights three difficult examples where the model wrongly predicted clinically mild (IGA Grades 0-1) cases as IGA Grades 3-4 instead. Correlation of predictions with clinical scores (Figure 4) was also substantially higher when using Inception v4 on the largest image size of 1200x1600. Correlations of the model predictions with each set of the human labels (clinical and researcher scoring) were similar, suggesting that the model had acceptable agreement with different sets of human readers. The two sets of human labels also showed a high correlation of 0.77, thereby verifying the repeatability of the ground truth labels. Confusion matrices (Fig. 3) show that the models performed sub-optimally on the IGA 2 label (intermediate severity).

Discussion

Making use of computerized systems to analyse digital images removes inherent intra and inter-observer variability of acne severity scoring, and ensures consistency and reproducibility as long as a high-resolution facial photograph is used for the analysis.

Some groups have reported promising results using neural networks for acne classification\textsuperscript{10-12}. One study classified acne into seven categories using a pre-trained model\textsuperscript{10}. However, face data was cropped into 50 pixel patches for training, which
may focus too narrowly on features at a small scale. Another study used features generated by a large set of statistical image metrics that were dimensionally reduced by principal component analysis, and then combined by a shallow neural network for classification of IGA grades\textsuperscript{11}. Instead of using pre-defined image metrics as features, our study uses deep learning and state of the art convolutional neural network architectures (based on DenseNet, Inception v4 and ResNet18) trained from ‘scratch’ to develop complex image features at different scales that are combined to produce an IGA classifier of entire faces. In order to exploit the high-resolution images, we did not use pre-trained models, which typically have constraints on image size. Using features at multiple scales for classification should ensure more holistic assessment that more realistically simulates a clinician’s approach to determining acne severity.

Some limitations of our study include a possible data imbalance in the training data that was retrospectively retrieved from the clinical institution’s image database. Confusion matrices (Fig. 3) show that the models performed sub-optimally on the IGA 2 label (intermediate severity), with a number of plausible reasons: clinical scores for this label were challenging and possibly inaccurate, features of the images labeled with IGA 2 were insufficiently distinctive, or the size of the dataset was insufficient such that the training set did not contain features found in the test set. Figure 4 also demonstrates how our model erroneously under predicted severity of acne. We postulate that this could be due to the darker skin phenotypes V and VI in some images, which may be underrepresented in the training data set. This may be overcome in a future prospective study with larger scale enrollments to generate larger datasets, which are well suited to deep learning techniques. Such data-driven deep learning approaches will continue to improve as studies proceed to enroll more patients, thereby acquiring more data over time. While our algorithm appears to perform similarly to the baseline Microsoft-Nestle algorithm, our results show stronger performance for the severe (IGA 3-4) group, which could have important implications for patients needing urgent treatment.

As the algorithm developed can be easily downloadable, this software can be readily portable and have a wide number of clinical applications. Hence, this approach has a distinct advantage over the current clinical approach as it takes guesswork out from clinicians, and there is no learning curve involved behind the use of the algorithm. We have also begun development of an Android mobile application that uses a more compact neural network architecture to perform classification on photographs
obtained via the mobile phone camera, which may facilitate patient access to clinical consults, and have future applications in tele-dermatology.

This same approach can also be further developed and extrapolated for other more complicated clinical scores done by dermatologists like the Psoriasis Area and Severity Index (PASI) score and SCOring Atopic Dermatitis (SCORAD). These other clinical scores will be more challenging as they take into account a greater number of variable features such as lichenification, erythema, and excoriation.

**Conclusion**

An automated algorithm based on deep learning for the scoring of acne vulgaris has been developed. Future larger patient enrolments will continue to improve performance. The development of a simple, easy to use digital image analysis using validated standardized scoring system would serve as a common language for patients, their physicians, and even amongst researchers around the world.
Table 1: Classification accuracy (1.0 is max score) of models and image sizes on validation and test (shaded) data. Best test accuracy in bold.

<table>
<thead>
<tr>
<th>Image size</th>
<th>DenseNet</th>
<th></th>
<th>Inception v4</th>
<th></th>
<th>Resnet18</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Validation</td>
<td>Test</td>
<td>Validation</td>
<td>Test</td>
<td>Validation</td>
<td>Test</td>
</tr>
<tr>
<td>600x800</td>
<td>0.66</td>
<td>0.58</td>
<td>0.54</td>
<td>0.63</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>750x1000</td>
<td>0.59</td>
<td>0.57</td>
<td>0.56</td>
<td><strong>0.67</strong></td>
<td>0.66</td>
<td>0.61</td>
</tr>
<tr>
<td>1200x1600</td>
<td>0.58</td>
<td>0.60</td>
<td>0.62</td>
<td><strong>0.67</strong></td>
<td>0.67</td>
<td>0.58</td>
</tr>
</tbody>
</table>
Figures

Figure 1: Flowchart showing data sorting and pre-processing.

Figure 2: Workflow of image pre-processing and model training.

Figure 3: Confusion matrices for DenseNet 1200x1600, Inception v4 1200x1600, and ResNet18 600x800. Inception was the best-performing model in our study, and compared favorably to the Microsoft-Nestle method particularly in the severe (IGA 3-4) group.
Figure 4: Pearson correlations between machine-predicted score and human labels (clinical scoring and researcher scoring) for each model and various image input sizes (1.0 is perfect correlation). Correlation between clinical scoring and researcher scoring was 0.77.

Figure 5: Three cases in the test set with ground truth IGA Grades 0-1 and predicted IGA Grades 3-4 by the Inception v4 model.
References

1;17(9):1006-1009.
**Supporting Information**

This document summarizes technical details of the deep learning models used in the study, including architecture and training parameters (Table S1).

**Inception v4**

was one of the models that we have trained and tested on our acne vulgaris dataset for the multiclass classification problem. The Inception v4 model is a complex convolutional neural network (CNN) based model that extracts image features from a range of feature scales through a very deep network. Our network architectures were more compact versions (reduced depth) of the original models. The model was trained to classify frontal face images of acne vulgaris into one of three classes: IGA 0-1, IGA 2 and IGA 3-4.

The Inception v4 model was originally designed for images of size 299x299x3 and trained on the ImageNet dataset for multi-class (1000 classes) classification. In acne vulgaris images, different types of acne have different sizes. Resizing these images to 299x299x3 for use with the pre-trained Inception v4 model may lose features from small lesions. In this work, we trained the Inception v4 model from scratch for three different image sizes of 600x800x3, 750x100x3 and 1200x1600x3. Figure S1 shows the parameters for the model using image size 600x800x3, where $\alpha$, $\beta$ and $\gamma$ are shown in Table S1. Parameters for larger image sizes are also shown in the table. In our more compact version of the Inception v4 model, the number of parameters was reduced by four times (ratio of 4) at the convolutional and fully connected layers.

**ResNet18** (short for Residual Network version 18) is another model that was used in our study for acne classification. The defining characteristic of ResNet is the 'shortcut paths' between layers that improve model training stability. Similarly to Inception v4, pre-trained models are limited to image size of 224x224x3 and are thus not appropriate for the current study. Models were trained from scratch on image sizes of images sizes of 600x800x3, 750x100x3 and 1200x1600x3. Figure 2 shows the architecture of the ResNet18 model. The parameters used are shown in Table S1, where $\alpha$, $\beta$, $\gamma$ and $\lambda$ are set to 2 for ResNet18. As shown in Figure S2, block A, B, C and D contain two layers each. When $\alpha$, $\beta$, $\gamma$ and $\lambda$ are set to 2, each block will contain 4 layers, and the network will have total of 18 layers (ResNet18) with one 7x7 convolutional layer at the start and one fully connected (softmax) layer at the end. In this work, for ResNet18 model, the reduction ratio is set to 2, dividing the number of filters at each layer by 2 thereby reducing the number of parameters by half.
DenseNet is the third model used in our study. DenseNet uses shortcut paths between each convolutional layer and subsequent layers within each block, preserving the features learned at each layer. This results in substantial learning power with a relatively small number of parameters. The model was similarly trained on the large image sizes stated above. Training parameters unique to DenseNet (e.g. growth rate, compression rate, repeat block) are shown in Table S1.
References


### Table S1: Selected parameters for Inception v4, DenseNet and ResNet18 used in our work for acne vulgaris classification.

<table>
<thead>
<tr>
<th>Model</th>
<th>Image size</th>
<th>Batch size</th>
<th>Optimizer</th>
<th>Learning rate</th>
<th>Dropout rate</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>λ</th>
<th>Growth rate</th>
<th>Compression rate</th>
<th>Repeat block</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception v4</td>
<td>600x800</td>
<td>40</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.7</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>750x1000</td>
<td>24</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.7</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1200x1600</td>
<td>8</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.7</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DenseNet</td>
<td>600x800</td>
<td>32</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
<td>0.5</td>
<td>[3,3]</td>
</tr>
<tr>
<td></td>
<td>750x1000</td>
<td>32</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
<td>0.5</td>
<td>[3,3]</td>
</tr>
<tr>
<td></td>
<td>1200x1600</td>
<td>12</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
<td>0.5</td>
<td>[3,3]</td>
</tr>
<tr>
<td>ResNet18</td>
<td>600x800</td>
<td>100</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>750x1000</td>
<td>80</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1200x1600</td>
<td>24</td>
<td>Adam</td>
<td>$1 \times 10^{-4}$</td>
<td>0.8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure S1: Schematic of the Inception v4 network for the classification of acne vulgaris images to IGA 0-1, IGA 2 and IGA 3-4. For the details of modules: Stem, Inception-A, Inception-B, Inception-C, Reduction-A, Reduction-B refer to the Inception v4 reference (1).
Figure S2: Schematic of the ResNet18 model used for the classification of acne vulgaris, where $\alpha$, $\beta$, $\gamma$ and $\lambda$ control for how many times each block with two convolution layers would be repeated.
Figure S3: Schematic of the DenseNet model used for the classification of acne vulgaris. Parameters such as the growth rate and compression rate control for the complexity of the model.