Application of U-Net with Inceptionv4 Encoder for Segmenting and Counting Monkeypox Lesions in Patient Photographs

Andrew J. McNeil[1,2,3] ANDREW.J.MCNEIL@VANDERBILT.EDU
[1] *Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, USA*
[2] *Dermatology Service and Research Service, Department of Veterans Affairs, Tennessee Valley Healthcare System, Nashville, TN, USA*
[3] *Vanderbilt Department of Dermatology, Vanderbilt University Medical Center, Nashville, TN, USA*

David W. House[3] DAVID.W.HOUSE@VANDERBILT.EDU

Placide Mbala[4] MBALAPLACIDE@GMAIL.COM
[4] *National Institute of Biomedical Research, Kinshasa, Democratic Republic of the Congo*

Olivier Tshiani Mbaya[4,6] OLIVIER.TSHIANIMBAYA@NIH.GOV
[6] *Clinical Monitoring Research Program Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, USA*

Lori E. Dodd[5] DODDL@NIAID.NIH.GOV
[5] *Clinical Trials Research Section, Division of Clinical Research, National Institute of Allergy and Infectious Disease, Bethesda, MD, USA*

Edward W. Cowen[7] COWENE@MAIL.NIH.GOV
[7] *Dermatology Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), Bethesda, MD, USA*

Ziche Chen[3] ZICHE.CHEN@VANDERBILT.EDU

Madeline Marks[3] MADELINE.MARKS@VUMC.ORG

Inga Saknite[3,8] INGA.SAKNITE@VUMC.ORG
[8] *Biophotonics Laboratory, Institute of Atomic Physics and Spectroscopy, University of Latvia, Riga, Latvia*

Eric R. Tkaczyk[2,3,9] ERIC.TKACZYK@VUMC.ORG
[9] *Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, USA*

Benoit M. Dawant[1] BENOIT.DAWANT@VANDERBILT.EDU

Editors: Under Review for MIDL 2022

Abstract

Monkeypox is a disease caused by infection with monkeypox virus that causes significant morbidity in several central and western African countries. There are currently no proven, safe treatments for monkeypox virus infection. Lesions pass through several stages before resolution. Evaluating the clinical efficacy of possible treatments involves manually tracking changes in lesion counts over time until lesion resolution (scabbed or desquamated), which is both labor intensive and prone to human error. To support a randomized controlled trial evaluating putative therapeutics we developed a deep learning method for monkeypox lesion segmentation and counting in patient photos. In 20 photos from 12 patients with monkeypox, we manually annotated all visible lesions and trained a U-Net network with Inceptionv4 encoder to segment lesions. In a leave-one-out evaluation our method shows promising results for lesion segmentation, with a median Dice of 0.74 (interquartile range: 0.72, 0.79) on the unseen photos. Automated lesion counting was evaluated on a second held-out set of 20 photographs. Automatic lesion counting performs similarly to human raters when compared to ground truth, with median lesion count difference of 2.00 (-8.00, 10.25) for the algorithm compared to 5.50 (2.75, 12.00) and -2.00 (-7.00, 1.25) for two other human raters.

Keywords: Monkeypox, segmentation, deep learning, U-Net, lesion counting

1. Introduction

Monkeypox is a re-emerging disease which causes significant morbidity, with human monkeypox cases have been increasing in sub-Saharan Africa since 2000 (Rimoin et al., 2010), and is considered a high threat pathogen causing a disease of public health importance (Sklenovska and Van Ranst, 2018). The first human case of recorded in 1970 in the Democratic Republic of the Congo (DRC). It has since spread to several other central and western African countries, often in remote areas where communication infrastructure and availability of medical expertise is limited. A monkeypox outbreak was declared on 9 December 2021 in Maniema, Democratic Republic of the Congo, with a total of 196 cases and 24 deaths as of 12 December 2021 (WHO, 2021).

Symptoms of monkeypox are similar to smallpox. Illness begins with fever, muscle aches, swollen lymph nodes, chills, and/or exhaustion. Within 1–3 days of the onset of initial symptoms, the patients develop lesions starting on the face then spreading across the body. Lesions progress through several stages, from macules to papules to vesicles to pustules, and finally to a scabbing and desquamating end stage. The illness typically lasts 2 to 4 weeks causing significant morbidity, and a mortality rate as high as 1 in 10.

There are currently no proven, safe treatments for monkeypox virus infection. Phase I clinical trials have demonstrated a reasonable safety profile of tecovirimat for treating orthopox virus infections in humans (Jordan et al., 2010; Mucker et al., 2013). However, randomized controlled clinical trials are needed to evaluate the efficacy of tecovirimat and similar compounds for treatment of human monkeypox infection. The primary measurement of disease progression is manual lesion counting from onset until resolution. WHO guidelines classify Monkeypox cases based on the number of skin lesions: mild (<25 skin lesions), moderate (25—99 skin lesions), severe (100—250 skin lesions), or grave (>250 skin lesions) (Jezek and Fenner, 1988). However, such manual assessments are both labor intensive and prone to human error, and no studies have been conducted to characterize the variability between observers or establish acceptable limits of agreement. To support future monkeypox studies, we propose a deep learning method for lesion segmentation and counting in patient photos.

No previous studies have been reported for monkeypox lesion segmentation or counting from photographs. However, a similar metric of lesion counts and endpoint of lesion resolution is used in the staging and tracking of acne. Lesion counting and classification are the primary targets, with segmentation performance a secondary target. Lesion localization and segmentation has been proposed by template matching (Humayun et al., 2012), multi-level thresholding and GLCM feature extraction (Hanifa Setianingrum et al., 2020), k-means clustering (Lucut and Smith, 2016; Malik et al., 2014), support vector machines (Hanifa Setianingrum et al., 2020; Alamdari et al., 2016), and semi-automated region growing with manual seed points (Budhi et al., 2017). Lesion classification has been assessed using shallow neural networks (Junayed et al., 2019; Malgina and Kurochkina, 2021), and classification models based on hand-crafted features (Hanifa Setianingrum et al., 2020; Malik et al., 2014; Lucut and Smith, 2016). Other deep learning, instance segmentation, and thresholding methods have been proposed for other skin diseases such as herpes zoster (Arias and Mejia, 2020; Mejía Lara and Arias Velasquez, 2022), chicken pox (Oyola et al., 2012), and histology images (Chen et al., 2017; Moen et al., 2019).

For a planned prospective study of monkeypox photographs which will be collected in 2022, we propose using a deep learning U-Net architecture to segment lesions then estimate lesion counts from a connected component analysis of the segmentation mask. We present this proof-of-concept study for as a baseline method for automating the segmentation and measurement of monkeypox lesions, developed and evaluated using a set of 40 historical photographs from 17 patients.

2. Materials & Methods

2.1. Data

The set of patient images provided for this study was gathered in the Democratic Republic of Congo between 2007 – 2011, using two consumer-grade cameras (Samsung Digimax L70 and Canon Powershot A630). These were acquired to document cases, particularly in mothers and young infants (Mbala et al., 2017). No imaging protocol was followed as they were not acquired for the purpose of lesion counting, leading to a variety of lighting conditions, backgrounds, fields of view, imaging distances, and body sites. Photographs have also been downsampled from their original resolution to 1024x768 pixels for archival.

A data curation process was followed for this historic set in order to gather all photographs of sufficient quality for our study. From the set of 381 photographs across 30 patients, we first removed those which were unrelated to skin lesions, such as those showing only a single large wound on the scalp or closeup photos of the eye. From the remaining 190 photos we removed those which were too blurred due to either motion or camera focus (58 photos) and repeated photos of the same skin area at the same time point (66 photos) retaining only the best example for annotation. Finally, no expert consensus guidelines have been established for counting large, coalesced lesions or wounds, so photographs with high numbers of these were also removed (26 photos).

The remaining 40 photographs were used in this study, covering 17 patients. We split this into two equal sets. Image set 1 was used for algorithm development and training, which we will refer to as the "development set". Image set 2 was held-out and only used to evaluate our final model on the task of lesion counting, and will be referred to as the "test set". Due to the paucity of data, we selected the photographs in the development set to cover representative examples of lesion appearance, body sites, and lighting conditions. The development set contained 20 photographs from 12 patients. The test set contained 20 photographs from 13 patients, with 11 images from 5 patients not present in the development set and the remaining 9 photographs coming from 8 patients represented in the development set. The photographs of patients present in both sets were of different body sites and time points, with no overlap of skin areas represented in the development and test sets. We recognize that the overlap of previously seen patients may lead to an overestimation of performance on the test set, so we also examine the subgroup of unseen patients in our analysis.

2.2. Ground Truth

Lesion Segmentation Manual segmentation masks were created for all 20 images in the development set using the open-source Gnu Image Manipulation Program (GIMP). A single annotator, Rater A, followed a predefined protocol whereby all visible lesions were traced on a transparent annotation layer using the pencil tool. Lesions from all stages were demarcated in the same manner, with the lesion boundaries drawn along to the edge of affected and normal appearing skin for each lesion. This required approximately 25 hours to annotate images in the development set. Once completed, the annotation layer of each image was exported to create a binary mask of lesion pixels.

Lesion Counts Ground truth lesion counts were also collected manually for the 20 images in the test set. GIMP was also used for this task, with the pencil tool used to mark the center of each visible lesion on a transparent annotation layer. Touching and coalesced lesions were marked separately if defined structures could still be discerned. A pencil diameter of 3 pixels was used to ensure that the markings never overlapped, even for small adjacent lesions. Each image was assessed by the same annotator (rater A) who had provided segmentation ground truth, in addition to two other human raters (B and C). Given the much higher level of experience and familiarity with the dataset, we consider the lesion counts by Rater A as the ground truth for all subsequent analyses.

2.3. Segmentation Algorithm

We selected the ubiquitous U-Net architecture first proposed by Ronneberger et al. (Ronneberger et al., 2015). This network uses a symmetric encoder-decoder structure with a contracting path which captures context and expanding path which enables precise localization. Long skip connections are used to concatenate the upsampled feature map in the expansive path with the corresponding feature map from the contracting path. Our algorithm uses an InceptionV4 network (Längkvist et al., 2014) as the encoder, initialized with ImageNet weights, following the implementation in “Segmentation Models Pytorch” (Yakubovskiy, Accessed December 12th, 2021).

An image-level leave-one-out experiment was conducted to assess the segmentation performance of the algorithm using the development set. For each fold a single photo was held out for testing. The training and validation sets were constructed from patches extracted from the remaining 19 photos. Each model was trained for 50 epochs with binary cross entropy loss using the Adam optimizer (Kingma and Ba, 2014), with an initial learning rate of 0.0001, reduced to 0.00001 after 25 epochs. The best weights for each model were selected from the highest Dice value on the validation set.

Data preparation consisted of extracting 300 128x128 pixel patches for training and validation. These were generated from random pixel locations in the image, with the constraint that the corresponding ground truth mask contained at least 1 lesion pixel.

During training, data augmentation was applied at each iteration via a random combination of elastic transformations, left/right flipping, gaussian blur, affine scaling and translation, perspective transforms, color temperature adjustments, and gamma adjustments.

Model testing was carried out with a sliding window approach, with a window size of 128x128 and stride of 32. Each pixel therefore receives 16 predications, with the final segmentation mask calculated via majority vote.

2.4. Performance Assessment

We evaluated the segmentation performance on the network using the Dice Similarity Coefficient (DSC), given by,

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

where $|X|$ is the number of all lesion pixels in the predicted segmentation mask and $|Y|$ is the number of all the lesion pixels in the ground truth mask.

Lesion count performance was evaluated by calculating the number of predicted lesions through connected component analysis of the segmentation mask for a given image, and comparing to the ground truth number of lesions counted by rater A. We also show correlation and Bland-Altman plots for comparisons between the network and ground truth.

3. Results & Discussion

3.1. Lesion Segmentation

An image-level leave-one-out experiment was conducted using the development set to estimate the segmentation performance of the network. For each fold, the model is trained on all patches from 19 photographs then tested on the held-out photo. Example lesion predictions are shown in Figure 1. Predicted lesions are shown as colored contours, with green showing lesions present in both the predicted mask and ground truth, blue showing areas only in the predicted mask (false positives), and magenta showing areas only in the ground truth (false negative). Figure 1a shows a close-up view of the back of the hands and lower right leg, with the majority of lesions in the umbilicated stage (i.e. belly-button-like in appearance). Figure 1b shows the predictions for hypopigmented papules on the back. For both lesion stages, the localization performance of the network is very good with a low number of false positive and false negative regions, despite the differences in lesion appearance, size, and lighting conditions. Figure 1c shows a more complex case, with multiple coalesced necrotic lesions on the hands and arms of an infant. Despite the good overall performance of the network for most lesions, the complex structures of the coalesced and necrotic regions results in a large discrepancy between predicted and ground truth lesion estimates when calculated from the segmentation masks (290 for the network and 219 for the ground truth). In this case, large regions of affected skin were marked as a single lesion by the annotator and the network marked many smaller areas.

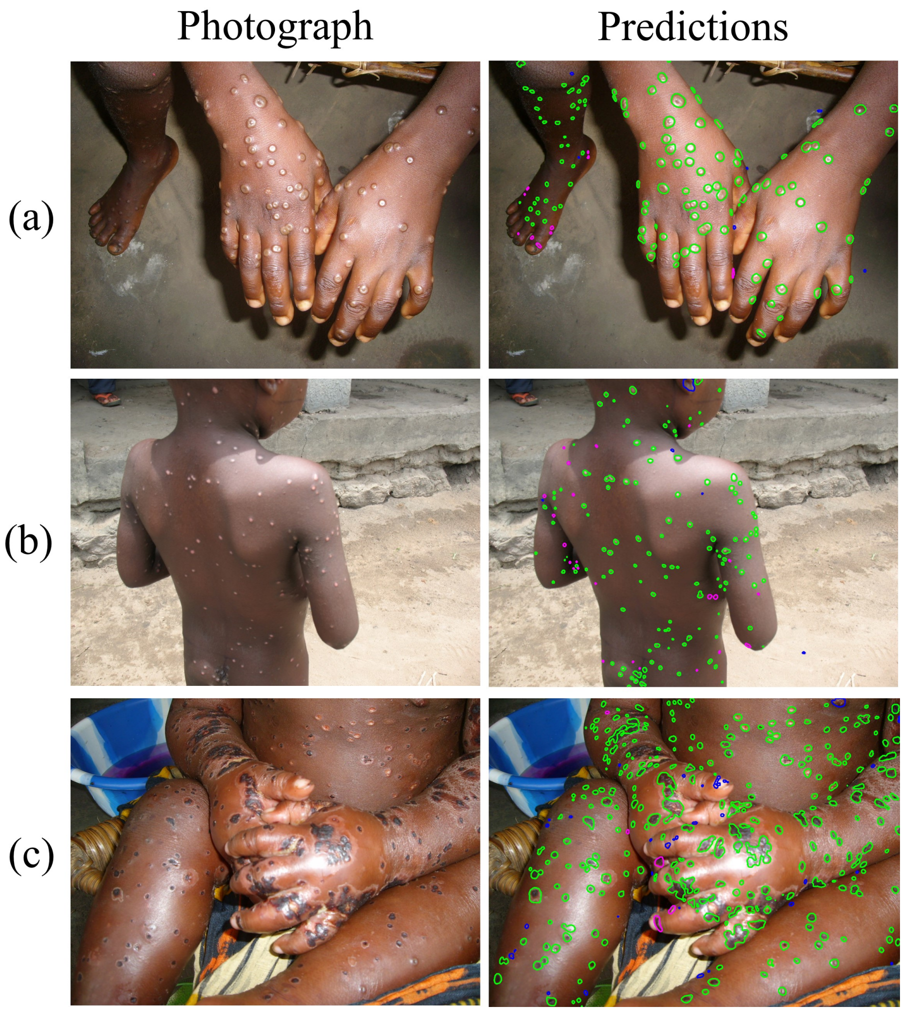

Figure 1: Example predicted lesions for images in the development set. Correctly identified lesions are shown as a green contour, false positives shown in blue, and false negatives shown in magenta. (a) Close view of umbilicated lesions taken indoors with flash. (b) Distant view of hypopigmented papules taken outdoors in natural light. (c) Close view of coalesced necrotic lesions.

The Dice value was calculated between the predicted lesion mask and the ground truth for each held-out image, with a median Dice of 0.74 (interquartile range: 0.72, 0.79) across the 20 images in the development set (Figure 2a).

3.2. Lesion Counting

To assess the performance of our network at estimating lesion counts a single model was trained using all 20 images in the development set. The model was then tested on the 20 unseen images in the test set, and lesion counts estimated by connected component analysis of

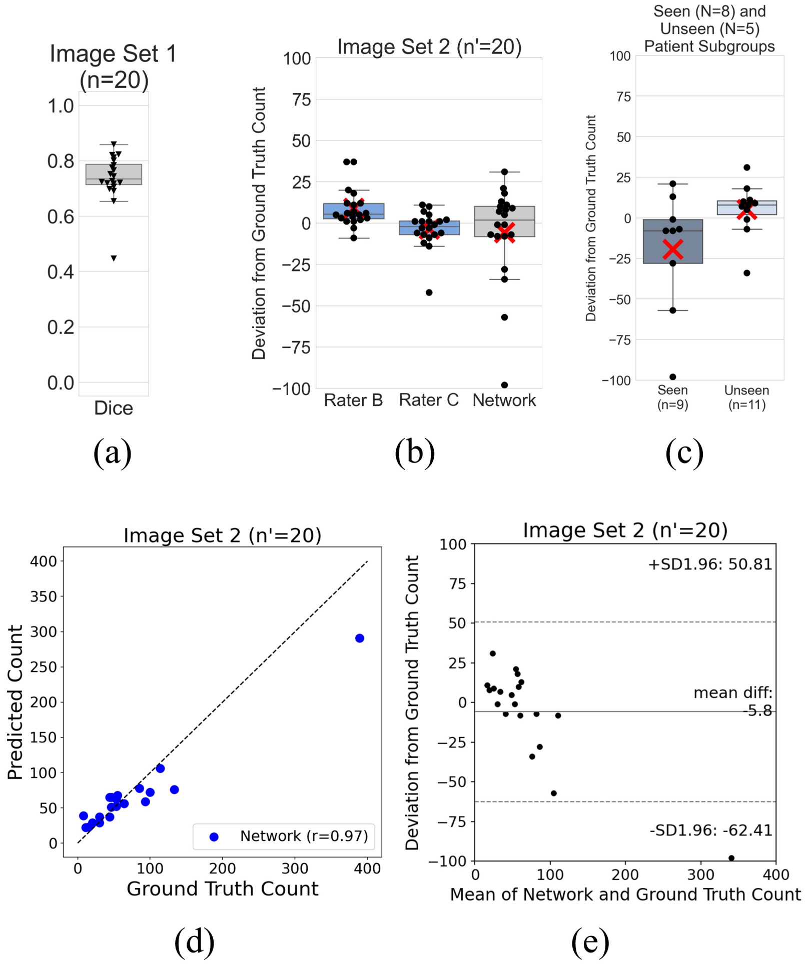

Figure 2: (a) Segmentation performance in leave-one-out experiment using Dice metric. Each point is a single image from the development set (set 1). (b) Difference in lesion counts (median, mean) from ground truth by human rater B (+5.5, +8.9), human rater C (-2.0, -3.6), and network (+2.0, -5.8). Mean value shown as red cross. (c) Lesion count performance of the network by patient sub-group; patients represented in the training set (seen), and entirely held-out patients (unseen) (d) Correlation plot of network vs ground truth lesion counts. Each point is a single image from the held-out test set (set 2). (e) Bland-Altman plot of the difference from network to ground truth, against the mean of network and ground truth count per image.

the segmentation masks. The predicted lesion count was then compared against the ground truth number of lesions (Figure 2d). We observe a strong correlation between predicted and ground truth counts (r = 0.97), with a tendency for the algorithm to underestimate as the density of lesions increases.

Comparing against two other human raters (Figure 2b) we find that the algorithm performs similarly, with differences from the ground truth counts (median, mean) by human rater B of (+5.5, +8.9), human rater C (-2.0, -3.6), and network (+2.0, -5.8). The lesion count performance of the network by patient sub-group (seen and unseen) is shown in Figure 2c. While we cannot draw firm conclusions due to the small sample size (n=9 and n=11 in the seen and unseen groups respectively), there does not appear to be a substantial difference in performance when the model has or has not seen skin from a different body region from the same patient.

We observe significant heteroscedascity in Figure 2e. This effect was also observed when comparing rater A manual counts to the counts derived from the ground truth segmentation masks in the development set. Coalescing and touching lesions are likely a significant source of this heteroscedascity due to lesion estimation deriving from connected components of the predicted segmentation masks.

4. Conclusions

We have presented the first application of image analysis techniques to the task of monkeypox lesion counting from photographs. The U-Net architecture with Inceptionv4 encoder, trained with a small set of ground truth segmentation images, may provide a reliable method of lesion segmentation in photographs of affected patients. Our method produces comparable performance to human raters for lesion counting, and may offer a more scalable solution to lesion tracking for future clinical studies.

WHO guidelines classify monkeypox cases based on the number of skin lesions. Accurate lesion numbers would therefore help in the initial staging of the disease for a patient. The evolution of lesion numbers at different stages may give a more accurate representation of disease progression over time, which would be a key clinical target for future clinical trials of potential treatments. We present our segmentation and counting method as a first step towards this longer-term goal. We anticipate the real-world application of our approach to a future clinical trial would involve acquiring patient photographs of a standardized set of body sites, manual lesion counting in the field, and automatic lesions counts from analysis of the images. This would serve as both a validation of the image analysis technique, and as a second read of the patient severity to help with quality assurance during the study.

Future work will aim to improve segmentation performance with more annotated patient photographs, and compare other techniques used in similar skin image processing applications against the established baseline method presented here. Lesion count estimation could also be improved through more sophisticated analysis of the predicted segmentation masks. A further extension to classifying lesion types would require further careful labelling of segmented lesion types by trained clinicians, but could also improve the tracking and staging of the disease in larger studies.

Acknowledgments

This work was supported by a Career Development Award from the United States Department of Veterans Affairs Clinical Sciences R&D Service (IK2 CX001785), a Discovery Research Grant from Vanderbilt University, the Vanderbilt University Medical Center Departments of Medicine and Dermatology, the National Institutes of Health (K12 CA090625, R21 AR074589), and the European Regional Development Fund (1.1.1.2/VIAA/4/20/665).

References

Nasim Alamdari, Kouhyar Tavakolian, Minhal Alhashim, and Reza Fazel-Rezai. Detection and classification of acne lesions in acne patients: A mobile application. In *2016 IEEE International Conference on Electro Information Technology (EIT)*, pages 0739–0743, 2016. doi: 10.1109/EIT.2016.7535331.

R. Arias and J. Mejia. Varicella zoster early detection with deep learning. In *2020 IEEE Engineering International Research Conference (EIRCON)*, pages 1–4, 2020.

Gregorius Satia Budhi, Rudy Adipranata, and Ari Gunawan. Acne segmentation and classification using region growing and self-organizing map. In *2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT)*, pages 78–83, 2017. doi: 10.1109/ICSIIT.2017.62.

Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical image analysis*, 36:135–146, 2017.

Anif Hanifa Setianingrum, Siti Ummi Masruroh, and Syifa Fitratul M. Performance of acne type identification using glcm and svm. In *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–4, 2020. doi: 10.1109/CITSM50537.2020.9268797.

Jawad Humayun, Aamir Saeed Malik, Samir Brahim Belhaouari, Nidal Kamel, and Felix Boon Bin Yap. Localization of acne lesion through template matching. volume 1, pages 91–94, 2012. ISBN 9781457719677. doi: 10.1109/ICIAS.2012.6306166.

Zdenek Jezek and Frank Fenner. Human monkeypox. *Monographs in Virology*, 17:1 – 140, 1988.

Robert Jordan, Janet M. Leeds, Shanthakumar Tyavanagimatt, and Dennis E. Hruby. Development of st-246® for treatment of poxvirus infections, 11 2010. ISSN 19994915.

Masum Shah Junayed, Afsana Ahsan Jeny, Syeda Tanjila Atik, Nafis Neehal, Asif Karim, Sami Azam, and Bharanidharan Shanmugam. Acnenet - a deep cnn based classification approach for acne classes. In *2019 12th International Conference on Information Communication Technology and System (ICTS)*, pages 203–208, 2019. doi: 10.1109/ICTS.2019.8850935.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 12 2014. URL http://arxiv.org/abs/1412.6980.

Martin Längkvist, Lars Karlsson, and Amy Loutfi. Inception-v4, inception-resnet and the impact of residual connections on learning. *Pattern Recognition Letters*, 42:11–24, 2014. ISSN 01678655. URL http://arxiv.org/abs/1512.00567.

Sergiu Lucut and Michael R. Smith. Dermatological tracking of chronic acne treatment effectiveness. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5421–5426, 2016. doi: 10.1109/EMBC.2016.7591953.

Elina Malgina and Maria-Anastasia Kurochkina. Development of the mobile application for assessing facial acne severity from photos. In *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, pages 1790–1793, 2021. doi: 10.1109/ElConRus51938.2021.9396382.

Aamir Saeed Malik, Roshaslinie Ramli, Ahmad Fadzil M. Hani, Yasir Salih, Felix Boon-Bin Yap, and Humaira Nisar. Digital assessment of facial acne vulgaris. In *2014 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pages 546–550, 2014. doi: 10.1109/I2MTC.2014.6860804.

Placide K. Mbala, John W. Huggins, Therese Riu-Rovira, Steve M. Ahuka, Prime Mulembakani, Anne W. Rimoin, James W. Martin, and Jean Jacques T. Muyembe. Maternal and fetal outcomes among pregnant women with human monkeypox infection in the democratic republic of congo. *Journal of Infectious Diseases*, 216:824–828, 11 2017. ISSN 15376613. doi: 10.1093/infdis/jix260.

Jennifer Vanessa Mejía Lara and Ricardo Manuel Arias Velasquez. Low-cost image analysis with convolutional neural network for herpes zoster. *Biomedical Signal Processing and Control*, 71:103250, 2022.

Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. Deep learning for cellular image analysis. *Nature methods*, 16(12):1233–1246, 2019.

Eric M. Mucker, Arthur J. Goff, Joshua D. Shamblin, Douglas W. Grosenbach, Inger K. Damon, Jason M. Mehal, Robert C. Holman, Darin Carroll, Nadia Gallardo, Victoria A. Olson, Cody J. Clemmons, Paul Hudson, and Dennis E. Hruby. Efficacy of tecovirimat (st-246) in nonhuman primates infected with variola virus (smallpox). *Antimicrobial Agents and Chemotherapy*, 57:6246–6253, 12 2013. ISSN 00664804. doi: 10.1128/AAC.00977-13.

Julián Oyola, Virginia Arroyo, Ana Ruedin, and Daniel Acevedo. Detection of chickenpox vesicles in digital images of skin lesions. In Luis Alvarez, Marta Mejail, Luis Gomez, and Julio Jacobo, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 583–590, 2012.

Anne W. Rimoin, Prime M. Mulembakani, Sara C. Johnston, James O. Lloyd Smith, Neville K. Kisalu, Timothee L. Kinkela, Seth Blumberg, Henri A. Thomassen, Brian L. Pike, Joseph N. Fair, Nathan D. Wolfe, Robert L. Shongo, Barney S. Graham, Pierre Formenty, Emile Okitolonda, Lisa E. Hensley, Hermann Meyer, Linda L. Wright, and Jean-Jacques Muyembe. Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the democratic republic of congo. *Proceedings of the National Academy of Sciences*, 107:16262–16267, 9 2010. ISSN 0027-8424. doi: 10.1073/PNAS.1005769107. URL https://www.pnas.org/content/107/37/16262.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:12–20, 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4.

Nikola Sklenovska and Marc Van Ranst. Emergence of monkeypox as the most important orthopoxvirus infection in humans. *Frontiers in Public Health*, 6, 2018.

WHO. Weekly bulletin on outbreaks and other emergencies. Week 52, 13–19 December 2021.

Pavel Yakubovskiy. Segmentation models pytorch, Accessed December 12th, 2021. URL https://github.com/qubvel/segmentation_models.pytorch.