Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

© 2023 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI230311

Can Synthetic Images Improve CNN Performance in Wound Image Classification?

Leila MALIHI[a], Ursula HÜBNER[b], Mats L. RICHTER[a], Maurice MOELLEKEN[c], Mareike PRZYSUCHA[b], Dorothee BUSCH[d], Jan HEGGEMANN[e], Guido HAFER[e], Stefan WIEMEYER[e], Gunther HEIDEMANN[a], Joachim DISSEMOND[c], Cornelia ERFURT-BERGE[d], Carlotta BARKHAU[f], Achim HENDRIKS[f] and Jens HÜSERS[b,1]
[a]*Institute of Cognitive Science, Osnabrück University, Germany*
[b]*Health Informatics Research Group, Osnabrück University of AS, Germany*
[c]*Department of Dermatology, Venerology and Allergology, University Hospital of Essen, Germany*
[d]*Department of Dermatology, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nürnberg, Germany*
[e]*Christian Hospital Melle, Niels Stensen Hospitals, Germany*
[f]*Symbic GmbH, Osnabrück, Germany*
ORCiD ID: Jens Hüsers https://orcid.org/0000-0003-3324-9155

Abstract. For artificial intelligence (AI) based systems to become clinically relevant, they must perform well. Machine Learning (ML) based AI systems require a large amount of labelled training data to achieve this level. In cases of a shortage of such large amounts, Generative Adversarial Networks (GAN) are a standard tool for synthesising artificial training images that can be used to augment the data set. We investigated the quality of synthetic wound images regarding two aspects: (i) improvement of wound-type classification by a Convolutional Neural Network (CNN) and (ii) how realistic such images look to clinical experts (n = 217). Concerning (i), results show a slight classification improvement. However, the connection between classification performance and the size of the artificial data set is still unclear. Regarding (ii), although the GAN could produce highly realistic images, the clinical experts took them for real in only 31% of the cases. It can be concluded that image quality may play a more significant role than data size in improving the CNN-based classification result.

Keywords. wound imaging, data augmentation, convolutional neural network, classification, artificial intelligence, generative adversarial networks, synthetic images

1. Introduction

Artificial intelligence (AI) systems can support health professionals in wound care by automatically recognising wound characteristics such as maceration [1] and infection [2] in wound images, thereby helping to standardise documentation and to curtail record-

[1] Corresponding Author: Jens Hüsers, Osnabrück University of AS, Health Informatics Research Group, PO Box 1940, 49009 Osnabrück, Germany; E-Mail: j.huesers@hs-osnabrueck.de.

keeping efforts. However, AI-based systems must perform at a very high level to become clinically relevant. Typically, Machine Learning (ML) based AI systems require a large amount of labelled training data to achieve this level, in particular when the complexity of the domain requires dense coverage. Such amounts of data are sometimes difficult to obtain for secondary use in healthcare, where data access and processing depend on the patient's consent. To use ML for rather sparse data, augmentation techniques can artificially inflate the data basis. Basic augmentation of image data sets can be achieved simply by standard transforms like randomly shifting, rotating, and mirroring the raw image. Beyond these simple methods, ML-based computer vision systems can learn how to generate authentic images of medical entities that do not exist in reality [3]. These systems have seen substantial development in recent years [4], e.g., in a study on ophthalmic images, an AI image generator [5,6] could provide synthetic training images that improved ML-based classification.

We transferred this idea to the domain of chronic wounds, where taking wound images is a standard procedure, but their availability for secondary use can be problematic. To augment training data for a classification task by artificial wound images, generative adversarial networks (GAN) can be employed. However, it is unclear if the desired effect on the classification task materialises and if the artificial images resemble real wound images. We, therefore, investigated the quality of such images regarding two questions: Do they improve the training of a convolutional neural network (CNN)? And second: Do they look realistic to human experts?

2. Methods

Two specialised wound care facilities in Germany, the Christian Hospital Melle and the Department of Dermatology, Venerology and Allergology of the University Hospital Essen, provided wound images taken in routine wound care showing two distinct wound types: diabetic foot ulcers and venous leg ulcers. The information on the wound type and the images were retrieved from the patient records. In total, 987 images were curated to build the dataset - 480 images of diabetic foot ulcers and 507 of venous leg ulcers. The average raw image resolution is 2705 x 3374 pixels. A clinician located the wound in the image with a bounding box used to crop the wound with an additional margin of 75 pixels. These cropped images were scaled to 256 by 256 pixels and finally checked for any errors by a second clinician. Next, this dataset was randomly split aiming for a ratio of 9:1 into a training set containing 864 images (88 %) and a hold-out test set containing 123 images (12 %).

A GAN of the StyleGAN3 architecture was trained to produce colorized synthetic but realistic looking wound images using the curated dataset described above. As with any GAN, StyleGAN3 comprises two neural networks. A first generator network produces natural-looking images, hereby trying to fool a second network, the discriminator, whose task is to distinguish the artificial from the real-world images. The discriminator's decision amplifies the generator's learning process and helps to improve the quality of the generated images. Simultaneously, the discriminator improves its performance on the increasing difficulty of the task as artificial images become ever more realistic [3]. We used a RTX8000 GPU for the training process.

In the next step, we trained deep CNN based on the Xception architecture [7]. As a basic data augmentation method, we randomly manipulated the training images' brightness and shear to account for the unstandardised lighting conditions and viewing

angles in the original wound images. The models were trained on a maximum of 100 epochs with an early stopping callback of 20 epochs of non-improvement. The ones with the lowest validation loss were evaluated on the test set using accuracy, recall, precision, and F1-score as performance indicators. The model training was done in Python using Tensorflow 2.9 with an NVIDIA Tesla T4 GPU. In the following experiment, we used this setup to evaluate the effect of augmentation by GAN-generated images on model performance (first research question): First, we trained a neural network using only the real images (plus the basic augmentations). Then, we doubled and quadrupled the dataset size by including the synthetic images.

To assess the second research question, we recruited 217 clinicians with self-reported wound expertise of 5 (median) on a scale from 1 (no experience) to 7 (maximum experience). They were asked to identify the synthetic images among a random subset of 60 real-world images from the test set and a random subset of 64 generated images in an online survey. The survey data was analysed using contingency tables comparing the ground truth (real vs synthetic) and the expert's predictions.

3. Results

We trained a GAN that produced synthetic wound images for diabetic foot ulcers and venous leg ulcers and performed the experiment that resulted in three classification models (Table 1). All models showed convergence and early stopping triggered before the maximum of 100 epochs in all training runs. All models yielded acceptable performance metrics. With the growing dataset size (that we obtained by augmenting the original training dataset with synthetic images), the accuracy and the F1-score metrics improved steadily with the dataset size (Table 1). However, precision and recall did not follow this trend. The maximum precision value was achieved with the largest dataset; however, the second largest dataset produced a precision value smaller than that of the original dataset. In contrast, the maximum recall value was achieved with the second largest dataset. However, the F1-score increased. Generally, the effect of the dataset's size on performance was inconsistent, depending on the performance indicator.

Table 1. Performance of the deep neural network classifier trained with three different training sets (raw, doubled and quadrupled dataset size, see left column). All three models with different training sets were evaluated on the same hold-out set.

Training Data	Size	Accuracy	Precision	Recall	F1
Original raw images	864	0.851	0.872	0.850	0.844
Double dataset size (50% real, 50% synthetic)	1,728	0.870	0.852	0.881	0.867
Quadrupled dataset size (25% real, 75% synthetic)	3,456	0.878	0.885	0.871	0.878

The GAN's performance on the quality of the images (research question ii) was investigated based on 26,908 decisions made by the 217 clinicians, each evaluating 124 (60 real and 64 synthetic) images. Table 2 shows the 26,908 decisions tabulated against the ground truth. The upper contingency table (table 2) reveals that out of the 13,020 decisions on the synthetic images, the clinicians regarded the images in 4,061 cases as real ones (31%).

Similarly, among the 12,321 decisions of voting for "real", there were 4,061 decisions that were based on synthetic images (33%). These conditional frequencies showed that clinicians regarded synthetic images in one-third or less of their decisions as real ones. Overall, 64% of the decisions (17,219) regarding synthetic and real images

were correct. Conversely, there were also real images regarded as synthetic by the clinicians.

Figure 1 provides epitomic images of synthetic and real images and the majority vote of the clinicians. As figure 1 demonstrates, "bad" synthetic images were identified as such (upper left) by many clinicians. Still, there were also "good" synthetic images, as the left lower corner example reveals. Likewise, true real images, such as the slightly blurred one in the upper right corner of Figure 1, were rated as synthetic.

Table 2 Contingency tables with conditional frequencies of ground truth contrasted by clinical decision

Aggregation		**Ground Truth**		
Per columns		Synthetic	Real	Total
Decision	Synthetic	69% (8,959)	41% (5,628)	54% (14,587)
	Real	**31% (4,061)**	59% (8,260)	46% (12,321)
	Total	100% (13,020)	100% (13,888)	100% (26,908)
Per rows		Synthetic	Real	Total
Decision	Synthetic	61% (8,959)	39% (5,628)	100% (14,587)
	Real	**33% (4,061)**	67% (8,260)	100% (12,321)
	Total	48% (13,020)	52% (13,888)	100% (26,908)

4. Discussion

In this study, a GAN was trained to produce synthetic wound images to augment a dataset for CNN training. The resulting CNN classification models showed good performance (F1-score ≥ 0.844) and tended to improve slightly with growing dataset size concerning F1-score and accuracy, albeit not for recall and precision. This small and inconsistent effect of data size does not follow previous studies, which showed that CNNs generally improve with more data available [8]. A possible reason for this finding could be the quality of the synthetic images. The clinicians' decisions provided valuable information on how well the GAN-generated images match real images. When presented with a synthetic image, the clinicians took them for real only in 31% of the cases. However, the GAN was capable of producing high-quality synthetic wound images like the one in the lower left panel of Figure 1.

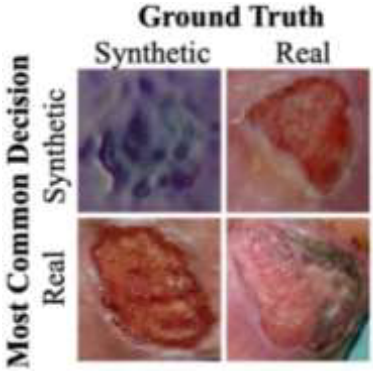

Figure 1. Image examples illustrating the contingency tables (Tab.2)

While it seems possible to improve the GAN to create better synthetic images, we assume that another direction of future research is more promising: If not the complete

set of synthetic images was used for training but only the ones that appeared realistic to clinicians, performance of the CNN is likely to rise with data size. Though this procedure requires manual classification into "good" and "bad" synthetic images, the effort is still lower compared to the acquisition of a larger database of real images.

Additional information obtained from clinicians' comments on the synthetic images might lead to further improvements: They frequently commented on unfavourable lighting conditions and image resolution that impeded identification. These comments were probably provoked by the unstandardised wound images we used in our curated dataset. We thus anticipate that wound images standardised with respect to angle, distance, lighting conditions, and proportion of wound area in the image would improve the image generation by GANs. Furthermore, the health professionals commented, that the wound images lacked peri-wound context information (due to a-priori cropping). We assume that this missing context may have made the task difficult for the clinicians as it differed too much from a real-world wound imaging setting. To close the gap to real world-wound imaging, GANs should aim additionally to synthesise peri-wound characteristics. It remains to be tested if the peri-wound context would provide useful additional features for CNN classification. Considering these options, we are optimistic about bringing GAN-generated images to a level of quality that will significantly improve automated classification from sparse training sets.

Acknowledgement

This study is part of a project funded by the German Ministry of Education and Research (BMBF Grant No. 16SV8616). We thank our project partners of the University Osnabrück New Public Health and apenio GmbH for their support and cooperation. This work was supported by a fellowship within the IFI program of the German Academic Exchange Service (DAAD).

References

[1] Hüsers J, Hafer G, et al. Automatic Classification of Diabetic Foot Ulcer Images: A Transfer-Learning Approach to Detect Wound Maceration. IOS Press; 2022. doi:10.3233/SHTI210919.

[2] Yogapriya J, Chandran V, Sumithra MG, Elakkiya B, Shamila Ebenezer A, Suresh Gnana Dhas C. Automated Detection of Infection in Diabetic Foot Ulcer Images Using Convolutional Neural Network. Journal of Healthcare Engineering. 2022 Apr 6;2022. doi:10.1155/2022/2349849.

[3] Yi X, Walia E, Babyn P. Unsupervised and semi-supervised learning with categorical generative adversarial networks assisted by wasserstein distance for dermoscopy image classification. arXiv preprint arXiv:1804.03700. 2018 Apr 10.. doi:10.48550/ARXIV.1804.03700.

[4] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 10684-10695). doi:10.48550/ARXIV.2112.10752.

[5] Kim M, Kim YN, Jang M, Hwang J, Kim HK, Yoon SC, Kim YJ, Kim N. Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization. Scientific Reports. 2022 Oct 15;12(1):17307. doi:10.1038/s41598-022-20698-3.

[6] Chen JS, Coyner AS, Chan RP, Hartnett ME, Moshfeghi DM, Owen LA, Kalpathy-Cramer J, Chiang MF, Campbell JP. Deepfakes in ophthalmology: applications and realism of synthetic retinal images from generative adversarial networks. Ophthalmology Science. 2021 Dec 1;1(4):100079. doi:10.1016/j.xops.2021.100079.

[7] Chollet F. Xception: Deep learning with depthwise separable convolutions. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 1251-1258). doi:10.1109/CVPR.2017.195.

[8] Luo C, Li X, Wang L, He J, Li D, Zhou J. How does the data set affect cnn-based image classification performance?. In2018 5th international conference on systems and informatics (ICSAI) 2018 Nov 10 (pp. 361-366). IEEE. doi:10.1109/ICSAI.2018.8599448.