

3-2021

WG²AN: Synthetic Wound Image Generation Using Generative Adversarial Network

Salih Sarp
Old Dominion University, ssarp@odu.edu

Murat Kuzlu
Old Dominion University, mkuzlu@odu.edu

Emmanuel Wilson
eKare, Inc., Fairfax, Virginia, USA

Ozgur Guler
eKare, Inc., Fairfax, Virginia, USA

Follow this and additional works at: https://digitalcommons.odu.edu/engtech_fac_pubs



Part of the [Artificial Intelligence and Robotics Commons](#), [Computer Engineering Commons](#), and the [Theory and Algorithms Commons](#)

Original Publication Citation

Sarp, S., Kuzlu, M., Wilson, E., & Guler, O. (2021). WG²AN: Synthetic wound image generation using generative adversarial network. *The Journal of Engineering*, 2021(5), 286-294. <https://doi.org/10.1049/tje2.12033>

This Article is brought to you for free and open access by the Engineering Technology at ODU Digital Commons. It has been accepted for inclusion in Engineering Technology Faculty Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

WG²AN: Synthetic wound image generation using generative adversarial network

Salih Sarp¹ | Murat Kuzlu¹ | Emmanuel Wilson² | Ozgur Guler²

¹ Batten College of Engineering & Technology, Old Dominion University Norfolk, Virginia, USA

² eKare, Inc Fairfax, Virginia, USA

Correspondence

Salih Sarp, Batten College of Engineering & Technology, Old Dominion University Norfolk, VA, USA.

Email: ssarp001@odu.edu

Funding information

Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development in Virginia.

Abstract

In part due to its ability to mimic any data distribution, Generative Adversarial Network (GAN) algorithms have been successfully applied to many applications, such as data augmentation, text-to-image translation, image-to-image translation, and image inpainting. Learning from data without crafting loss functions for each application provides broader applicability of the GAN algorithm. Medical image synthesis is also another field that the GAN algorithm has great potential to assist clinician training. This paper proposes a synthetic wound image generation model based on GAN architecture to increase the quality of clinical training. The proposed model is trained on chronic wound datasets with various sizes taken from real hospital environments. Hyperparameters such as epoch count and dataset size for training tasks are studied to find optimum training conditions as well. The performance of the developed model was evaluated through a mean squared error (MSE) metric to determine the similarity between generated and actual wounds. Visual inspection is performed to examine generated wound images. The results show that the proposed synthetic wound image generation (WG²AN) model has great potential to be used in medical training and performs well in producing synthetic wound images with a 1000-image training dataset and 200 epochs of training.

1 | INTRODUCTION

Computer–human interaction has an essential impact on facilitating knowledge generation, dissemination, access, and utilization [1]. The study of this interaction has evolved by broader terms into the active research domain of machine learning (ML) or Artificial Intelligence (AI) over the past 60 years [2].

A key element of successful deep learning depends on the availability of massive amounts of data [3]. Deep learning (DL) applications in healthcare are found to be lagging, in large part, due to the high cost of accurate and tagged data collecting. The total cost of data management in healthcare could rise to millions of dollars [4], which affects medical training as well. Medical students' education costs are high, with many instances of not enough data to showcase the pathologies, wounds, and diseases. In addition, class imbalances hinder the performance of AI models, such as decision trees (DT), neural networks (NN),

and support vector machines (SVM) [5]. Limited generalization ability and overfitting problems are too common without reliable training datasets utilizing neural networks with millions of parameters [6]. One could say that data augmentation techniques could be used to prevent overfitting and improve imbalanced class problems [7]. However, augmented data still resembles the original image intuitively [8]. This is why the quantity of training data is crucial to train and validate a model utilizing machine learning techniques [9].

Determining optimal hyperparameters is critical to AI models. However, the process is largely considered a black art [10]. Training dataset and the number of epochs play an important role while building new types of AI models, that is data-hungry, deep learning models. The number of epochs to train a model could be determined through loss functions in traditional machine learning models, whereas a Generative Adversarial Network (GAN) algorithm provides complicated loss functions. Hence, finding the optimum

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *The Journal of Engineering* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

number of epochs is also a critical parameter that requires labour, time, and expertise in the healthcare field [11, 12]. In the scope of this study, the performance of the synthetic wound generation was examined extensively regarding the number of training epochs and the training dataset size. Six different wound datasets were created to measure the effect of dataset size, and the model was trained for 2000 epochs. Results of every five epochs till the 2000th epoch was compared to analyse the effect of epoch count in detail.

In the literature, there are limited studies using the GAN algorithm in the healthcare domain. A conditional Wasserstein GAN framework was introduced for Electroencephalography (EEG) data augmentation to improve emotion recognition [13]. Magnetic Resonance (MR) image generation using Wasserstein GAN was presented for data augmentation and physician training [14]. In the study by Antoniou et al. [15], a data augmentation GAN model was developed to generate broader data augmentation techniques such as mirroring and rotation. Synthetic data was generated through GAN models to augment the dataset with synthetic images to maximize the performance of the classifier [16] and to improve the classification task on blood cells [17]. Authors in [18], propose balancing GAN (BAGAN) to balance class distribution while training the GAN with all available images and generate synthetic underrepresented class images utilizing conditioning in the latent space. A study by Wang et al., introduces the generation of belittled class plankton images while training the GAN with belittled class images and classify plankton types using a shared convolutional neural network (CNN) layers with a discriminator network [19]. The authors in [20], propose a training scheme that implements classical data augmentation techniques to enlarge the Computed Tomography (CT) images of the liver and then enlarge the dataset a second time through synthetic image generation with GAN using classically augmented data for training. The study in [21], investigates different GAN architectures, such as Super Resolution GAN (SRGAN) and DualGAN, to generate CT images from real Magnetic Resonance Imaging (MRI). The Authors in [22] propose an autoencoder combined generative adversarial network to synthesize jellyfish using a small dataset compared to other GANs. The study in [23] investigates the generation of synthetic fundus images of age-related macular degeneration (AMD), which are indistinguishable from real ones using progressively grown generative adversarial networks (ProGANs).

The aim of this paper is to utilize a GAN architecture to create a novel model that generates synthetic wound images, which was not done previously. This can help the training of medical students and clinicians on wound type and wound stage prediction more accurately. This study also presents a criterion for the optimum number of training images and epochs for generating wound images using GAN algorithms.

The paper is organized into five sections. Section 1 introduces the proposed algorithm details and states the problem. Section 2 presents the methodology employed. The workflow for data collection, dataset preparation for processing, environment, and validation steps are presented in Section 3. Experiments using the various training datasets and epoch counts are

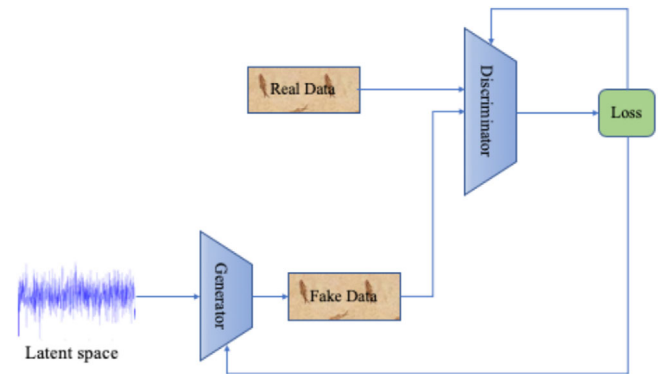


FIGURE 1 The basic structure of the GAN model.

examined in Section 4. Results and related discussions are summarized in Section 5. Objectives of this study are:

- Development of a model using the GAN algorithm in a conditional setting to generate synthetic wound images.
- Use of Mean Squared Error (MSE) metric to compare similarity between generated and actual wound images.
- Validate and evaluate model performance visually.
- Develop a hyperparameter selection guideline that can be utilized by healthcare researchers while training medical AI applications.

2 | METHODOLOGY

GAN has the ability to mimic any data distribution with the help of adversarial generator and discriminator networks. The min-max optimization between the adversarial generator (G) and discriminator networks (D) is at the heart of the GAN architecture. Both networks are trained simultaneously with real data to learn the data distribution evenly. The generator network gets uniform or Gaussian noise (z) as an input and produces fake images (y), which are fed to the discriminator network so as to be classified as real or fake, $D: \{y\} \rightarrow [0,1]$. The generator and the discriminator networks are trained once with the training set. Following this, the generator network is used to generate fake images that are not differentiable from the training set. Figure 1 shows the basic structure of the GAN model.

There are various versions of the GAN algorithm for different applications. For our study, Vanilla GAN, DCGAN, CycleGAN, and cGAN based versions were examined. The Vanilla GAN and DCGAN based models for synthetic wound generation were not suitable for our study as generated wound images were not compatible with the tissue segmentations. On the other hand, CycleGAN based model was promising due to its ability of unpaired image-to-image translation. The initial trials with the CycleGAN model revealed that the generated synthetic wound images experience a mode collapse problem, a well-known complication in the GAN field, which causes the synthesized wound images to look similar despite different tissue segmentation inputs [24]. As these models failed to

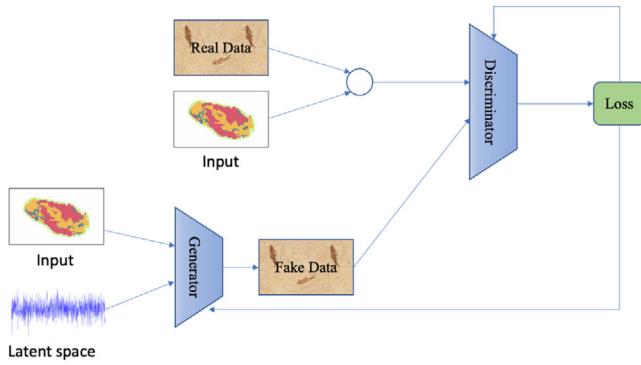


FIGURE 2 The general framework of the model architecture.

generate proper wound images coherent to the given wound tissue segmentation images, these different GAN approaches, namely Vanilla, DCGAN, and CycleGAN based models, are discarded from this study due to mentioned issues above.

A cGAN based architecture is used for conditioning the output (generated image), for specific data distributions, with an input image (x) and the addition of a noise (z) factor. In order to condition the output of the generator, the input is given to the generator. The input image is the segmented image ground truth that is gathered from eKare LLC. The discriminator is also fed with the input segmented image and the real data to distinguish whether the generator performed meaningful image synthesis or not.

The generator network learns a mapping from the input domain to the output domain, $G: \{x, z\} \rightarrow y$. The discriminator network learns a loss function to train this mapping [25] and tries to differentiate the fake image (y) from the real one, $D: \{x, y\} \rightarrow [0,1]$. This architecture is used to generate outputs that are similar to the input. This architecture gets paired and aligned images as an input to generate look-alike images. Both networks, the discriminator and generator, observe the input. Figure 2 shows the general architecture of the model that is fed with noise and input data. Loss coming from discriminator is fed to generator.

The discriminator network has two losses, namely D_{real} and D_{fake} , which indicate the ability of the discriminator network to differentiate the real and the fake images. Another loss used for the generator loss is G_{L1} , which compares the generated fake image and the real image to generate more plausible fake images. Adversarial loss (\mathcal{L}_{Adv}) coming from the discriminator network together with $L1$ loss (\mathcal{L}_{L1}) of the generator network (G_{L1}) used for generator training causes complicated objective functions for the generator network. This is why the process of training a GAN architecture is hard to interpret.

$$\begin{aligned} \mathcal{L}_{Adv}(G, D) = & \mathbb{E}_{x,y} [\log D(x,y)] \\ & + \mathbb{E}_{x,z} [\log(1 - D(x, G(x, z)))]. \end{aligned} \quad (1)$$

$L1$ loss behaves as a regularizing term with a hyperparameter, lambda (λ), which is chosen as 100. The final objective function is an analogy of a minimax game where the generator tries to

minimize, and the discriminator maximizes an adversarial objective. This relationship is depicted as:

$$G^* = \operatorname{argmin}_G \max_D \mathcal{L}_{Adv}(G, D) + \lambda \mathcal{L}_{L1}(G). \quad (2)$$

The hyperparameter problem was examined through various experiments. To measure the effect of dataset size, six (6) training datasets were created. For each of these datasets, a model was trained for 2000 epochs. The performance of the model was measured every five epochs through MSE metric for extensive analyses of the hyperparameter selection. The significance of the hyperparameter is examined via visual inspection as well.

3 | DATA COLLECTION, PRE-PROCESSING, ENVIRONMENT, AND VALIDATION

This section discusses data collection, data pre-processing, the simulation environment, and the validation methods.

3.1 | Data collection

Chronic wound images were provided by eKare, Inc (Fairfax, VA). Chronic wounds of various types, that is diabetic, burn, lymphovascular, pressure injury, and surgical, are anonymized and semi-automatically labelled for training and testing [26]. The diversity of chronic wound types increases the applicability of the model. This dataset includes around 4100 wound images ranging in size from 1224×1224 to 2160×2160 pixel dimensions (depending on the camera used to acquire measurement). Wound images were acquired under normal ambient room lighting by clinical users with commercially available 3D wound cameras.

3.2 | Pre-processing

Chronic wound images are rescaled to 512×512 , with the original wound and its labelled pair concatenated to form a 512×1024 pixel image. Before concatenation, the background of the labelled image is cleaned and made white. One hundred wound images out of the 4100 are chosen for the test dataset. The training dataset constitutes the rest of the dataset. To study the effectiveness of the dataset size, training datasets with sample sizes of 100, 250, 500, 1000, 2000, and 4000 were created. Normalization and formatting were done to ensure high performance during training.

3.3 | Environment

The proposed synthetic wound image generation using the GAN (WG²AN) model is implemented using Pytorch deep learning framework with Python version 3.7. The model runs on an Intel® Core™ i7 -8700X CPU @3.20 GHz with 32 GB

memory and NVIDIA GeForce GTX 1080 GPU with 8 GB dedicated and 16 GB shared memory. The model training for 2000 epochs took around 4, 6, 10, 16, 29, and 58 hours for 100, 250, 500, 1000, 2000, and 4000 images, respectively. The batch size is 64 for maximum GPU performance.

The Adam optimizer with a learning rate of 0.0002 is utilized for the first half of the training, while the rest is done with a linearly decaying learning rate to zero. The Adam optimizer is used as it converges faster, and also lower loss values are generated for the generator network in comparison to other optimizers in GAN applications [27]. The learning rate setting is chosen in line with the previous works done in image synthesis using GAN algorithms [28].

3.4 | Validation

Validation of the proposed method is done with the mean squared error (MSE) metric, which is a pixel-wise error measurement. Three colour channels (RGB) are considered while calculating the MSE scores where Y_R, Y_G, Y_B are used for the colour channels of real images, and Y'_R, Y'_G, Y'_B are used for the colour channels of generated images. Additionally, n denotes the number of pixels.

$$\text{MSE} = 1/n \sum_{i=1}^n \left[\left(Y_R - Y'_R \right)^2 + \left(Y_G - Y'_G \right)^2 + \left(Y_B - Y'_B \right)^2 \right]. \quad (4)$$

The MSE metric is applied to each pixel in the fake and real images. Each pixel has three channels, namely red, green, and blue. Each corresponding pixel in fake and real images are compared, and the MSE score is calculated. The MSE metric provides better quality scores [29]. Converging and meaningful results for the synthetic wound generation task is obtained using the MSE metric. The synthetic wound results are expected to be similar to actual wound images where the MSE metric could evaluate such similarity properly.

4 | EXPERIMENTS

Model and loss function outputs, the effect of dataset size, and epoch count on synthetic wound generation are discussed with MSE score and visual inspection in this section. Additionally, the optimum training conditions are determined with hyperparameter tuning.

4.1 | Model output

The proposed model gets wound tissue segmentation and produces synthetic wound images. Wound segmentation is done with respect to wound tissue types, namely necrotic (blue colour), granulation (red colour), and slough (yellow colour).

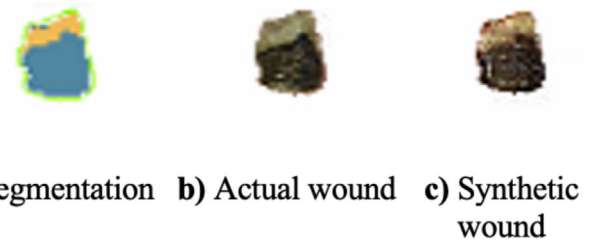


FIGURE 3 WG²AN model results trained with 4000 images and at 2000 epochs.

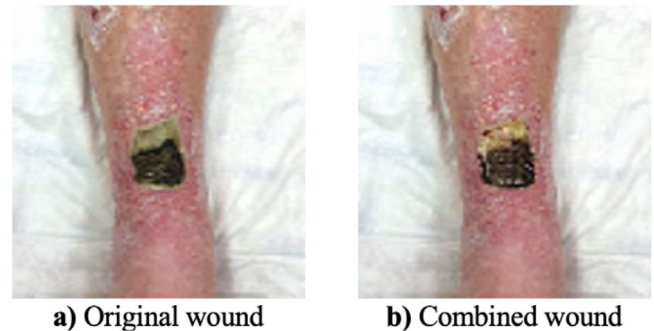


FIGURE 4 WG²AN model results applied on a real limb that is trained with 4000 images and at 2000 epochs.

This classification is of foremost importance to the caregiver while assessing a wound and deciding on a treatment plan. A successful result from the WG²AN model is depicted in Figure 3, which indicates proper synthetic wound generation by training with 4000 images and 2000 epochs. Actual wound, Figure 3B, and synthetic wound image, Figure 3C, were presented without background for a better comparison with the segmentation input, Figure 3A.

In addition to the original wound image, a combination of the synthetic wound and original wound background is shown in Figure 4A-B. The generated image has characteristics of a real wound.

4.2 | Effect of number of images and epochs on model loss

The chronic wound dataset is pre-processed and arranged as six (6) subgroups, which have 100, 250, 500, 1000, 2000, and 4000 images. The proposed model with WG²AN architecture to synthesize wound images are trained for 2000 epochs. The effects of dataset size and epoch count are examined to find the optimum training dataset.

The G_L1 loss curves of the model with different training dataset sizes are depicted in Figure 5. The adversarial and the discriminator network losses, that is D_fake, D_real, do not provide useful information, and they are not included in Figure 5. The proposed model could not settle down in the first 75th epochs as the losses increase, which is a common problem because of the randomized weights in the neural networks. The

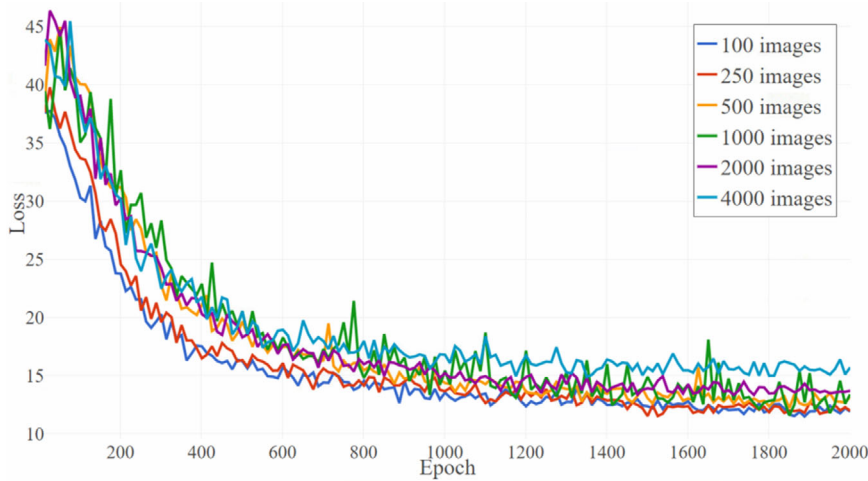


FIGURE 5 Model's loss graphs using a different number of images.

loss curves also share a similar fluctuating but decreasing pattern. The general zigzag behaviour of the loss curves is a result of the alternating wound image samples during training. The proposed model moves along the different training samples without reaching an equilibrium, which causes this oscillation [30].

The loss curve of the model with 100-images has the lowest loss. The model with 250-images has the second-lowest loss values. Other curves follow this characteristic as well. It can be deduced that models with fewer training images produce lower loss values. The loss graphs are compared to evaluate the learning ability of the model in Figure 5. The model with 100-images goes below the loss of 25 at the 180th epoch and stabilizes under the loss of 15 at the 600th epoch. 250-image model's loss decreases to 20 at the 200th epoch and becomes stable under 15 at the 700th epoch. The models with 100 and 250 images move closer curves as their training dataset size difference is small, that is 150.

The loss of the model with 500 images reduces to 25 at the 300th epoch and balances under 15 at the 950th epoch. The 1000-image model drops to the 25 at the 350th epoch and flattens under 15 at the 1100th epoch. The 2000-image model decreases to the 275th epoch and stays stable under 15 at the 950th epoch. The model with 4000 images goes down to 25 at the 250th epoch and balanced around 15 at the 1000th epoch.

The 4000-image model has trained for 4000 epochs; however, it does not stabilize under 15 and oscillate around it. The drop rate of the G_L1 loss indicates that our proposed model learns the data distribution faster with a smaller dataset. The models with a larger training set have more complicated data distributions that take more epoch to mimic.

As stated in (2), the generator and the discriminator networks compete over a min-max game, which results in a non-converging problem [25]. Although G_L1 loss has a meaningful curve that could be used to determine if the model is learning the training data distribution, it lacks the contribution of the adversarial loss. That's why an additional evaluation metric is needed, that is MSE. MSE score provides a better understanding of the generation of synthetic wound images by the proposed WG²AN model if they look like real wounds or not.

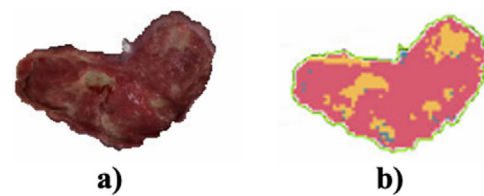


FIGURE 6 (a) Original wound and (b) wound segmentation.

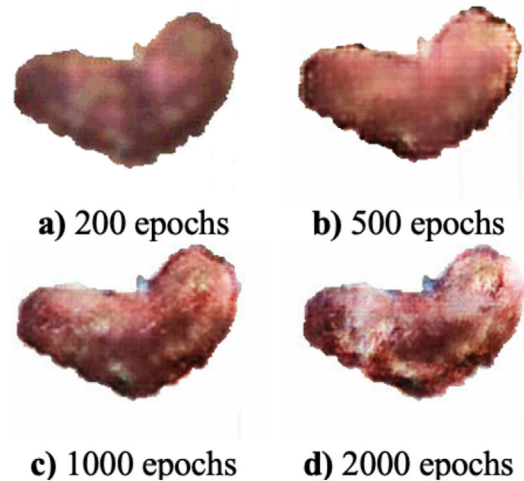


FIGURE 7 WG²AN output using 100 images for different epochs.

4.3 | Effect of number of images and epochs on synthetic wound generation

The effect of training parameters, that is dataset size and epoch count, are discussed in this section. The models with different training dataset sizes (100, 250, 500, 1000, 2000, and 4000) are evaluated visually and using the MSE score. Figure 6 indicates the actual wound and the segmented wound, which is the input of the model. The outputs of four different models are compared in Figures 7–12 with respect to epoch counts (200, 500, 1000, and 2000).

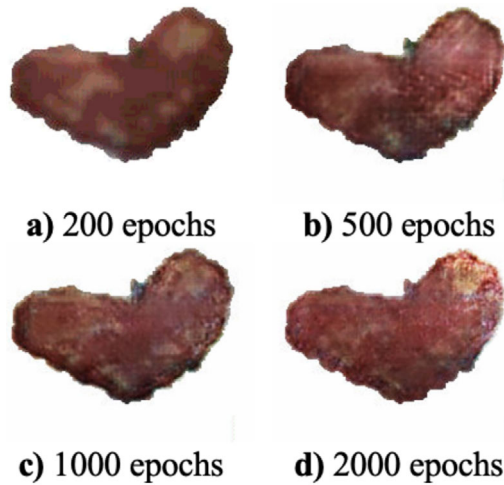


FIGURE 8 WG²AN output using 250 images for different epochs.

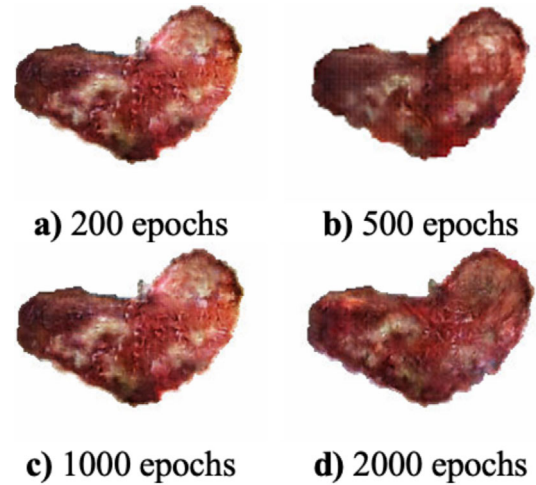


FIGURE 10 WG²AN output using 1000 images for different epochs.

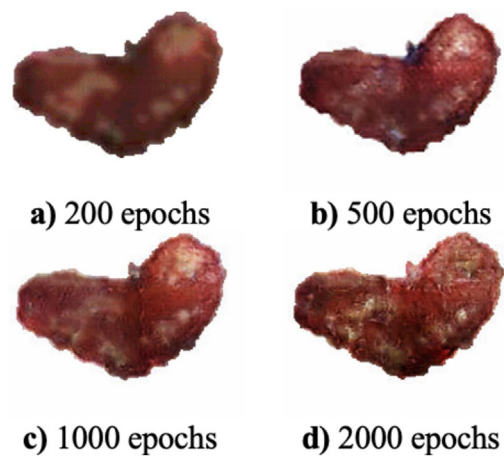


FIGURE 9 WG²AN output using 500 images for different epochs.

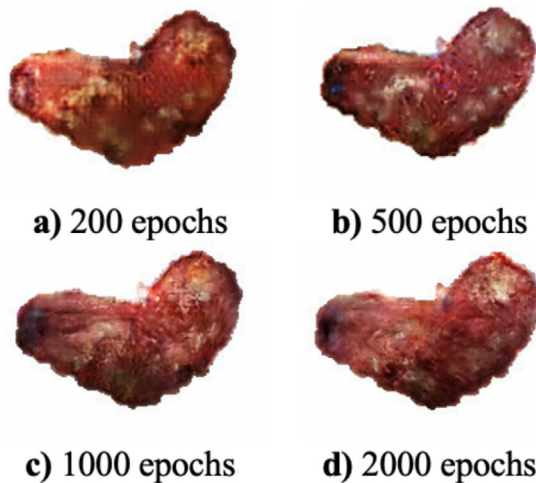


FIGURE 11 WG²AN output using 2000 images for different epochs

The model with the 4000-image dataset is trained for 4000 epochs, but the effect is negligible. That's why we cut training at the 2000th epoch for each model. The generated images before the 75th epoch lack the details and texture of a wound in general and suffer from the checkerboard effect significantly. The up-sampling layer in the generation pipeline of a GAN model, which produces high-resolution images from low-resolution ones, causes checkerboard artefacts [31]. The checkerboard pattern emerges when deconvolution has uneven overlap [32].

The models with smaller training datasets tend to be biased towards a certain data distribution, which is the case for 250 and 500-image models, which produced darker wounds. After further training of the same model, dataset limitation recovered at the 500-image model at 2000th epochs, as seen in Figure 9D. Training the models longer provides detailed texture and balance in the distribution of the tissue (also visible in Figure 9D).

The 100-image model generated insufficient wounds for the 500 epochs and the lower epoch counts. However, increasing the training to 1000 epochs improved the generated images significantly. Results for the 100-image model in Figures 7C

and 7D have close to real composition but generated samples in Figures 7A and 7B are far from a real wound nature.

The 250-image model generated primitive wounds at 200 epochs in Figure 8A. Generated wounds have more balanced wound tissue characteristics with some limitations, such as overall darker colours at 500, 1000, and 2000 epochs in Figure 8B-D. The 200 epochs of training is not enough for a life-like wound image generation, as seen in Figure 8A.

The model with the 500 images produced more life-like wounds at 200 and more epochs of training, shown in Figure 9C and D. The result of 500 images at 200 epochs in Figure 9A does not represent a real wound.

The model with 1000 images exhibits better performance than previous models. At 200 epochs of training in Figure 10A, it shows a reasonable output where previous models just produce an elementary output from fed segmentation data. The results of 500, 1000, and 2000 epochs of training produce better-wound images, as seen in Figure 10B-D.

The 2000 image model produced life-like images at 500, 1000, and 2000 epochs in Figure 11B-D. The 200 epochs of training

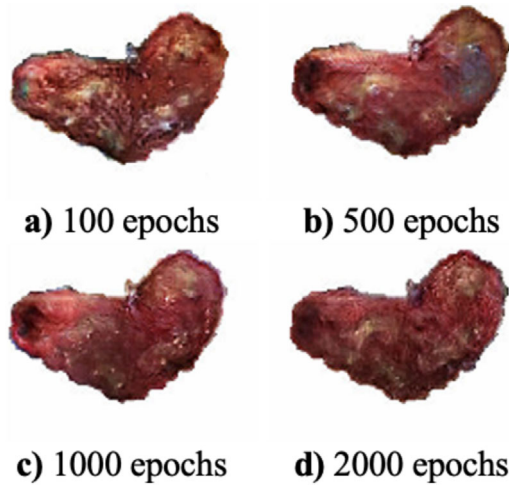


FIGURE 12 WG²AN output using 4000 images for different epochs.

TABLE 1 MSE scores for the WG²AN model

Images	Epochs				
	100	200	500	1000	2000
100	745	732	747	759	785
250	727	711	720	721	721
500	627	707	707	701	691
1000	736	703	703	704	709
2000	729	709	709	706	706
4000	721	697	697	705	710

in Figure 11A also form a wound with close to wound tissue characteristics.

The model with 4000 images generates life-like wounds at 200 epochs to 2000 epochs in Figure 12A-D. Every generated sample carry characteristics of the wound and well-balanced tissue distribution.

By varying the number of epochs and dataset size, generated image quality has improved significantly. It can be inferred that the lack of data could be substituted by further training of the model.

4.4 | Evaluation of the model with MSE metric

The model performance is evaluated by the MSE metric, as summarized in Table 1, which indicated the MSE scores of the models with different training images and epochs. A comparison of the MSE scores can be examined in Figure 13. MSE scores are calculated on individual images, after which the median average is calculated. A comparison of the MSE scores provides guidance for the generated image quality. A lower MSE score means better-generated image quality. The overall trend of the MSE scores in Table 1 is decreasing with expanding dataset

size and further training of the models. The first 75–100 epochs of training are necessary for the model to settle. The models' MSE scores for 100 epochs are also included in Table 1. The MSE score of the 100-image model was the highest for every epoch. It starts with the MSE score of 745, decreases to 732, increases to 747, and then keeps increasing with additional epochs of training.

The 250-image model has an MSE score of 727 at the 100th epoch and drops to 711 and has higher MSE scores with the increase of training. The model with 500 images does not settle at the 100th epoch, where the MSE score is lower than any other MSE score in the table. It becomes 707 and stays stable at that MSE score. There is a drop in the 2000th epoch of the 500-image model, which is a result of an overfitting problem.

The 1000-image model has an MSE score of 736 at the 100th epoch and reduces to 703 at the 200th epoch. It stays stable till the 1000th epoch. It rises slightly to 704 at the 1000th epoch and 709 at the 2000th epoch. The smaller increase means that there is a deformation in the generated images. This can be observed by comparing Figures 10C and 10D. The slough area gets smaller in the former one, which causes a slight MSE increase. Models with a higher number of training images, that is 2000 and 4000, follow a similar pattern with the 1000-image model.

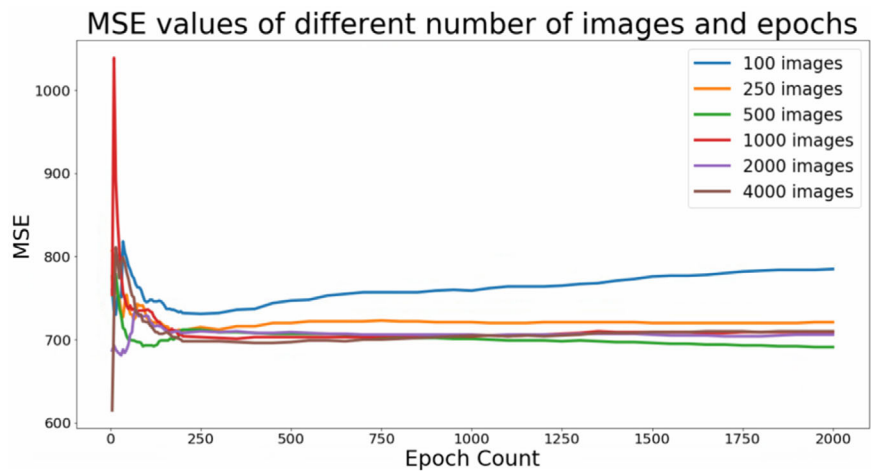
MSE score of the 5th epoch results is negligible as generated images are not good enough to mimic wound texture and detail. Training the model with a smaller dataset for a longer number of epochs results in overfitting, that is an increase in MSE score. This can be seen in Figure 13, especially after the 500th and 200th epochs. It is confirmed that increasing dataset size and epoch count until the 500th epoch gives a better MSE score. For the first 200 epochs of training, the decrease in MSE score is so significant that later increasing of training has minimal effect on the score. Therefore, 200-epoch is the optimum epoch number for training the proposed method. The amount of dataset size to generate plausible images is 1000 because the 1000-image model generated most life-like wounds, as seen in Figure 10A. That's why we conclude that the 1000 images and 200 epochs of training are the optimum training hyperparameters for plausible wound image generation.

5 | DISCUSSION

The proposed method generates synthetic wound images that have an input of segmented tissue outline, shown in Figure 2. The output is compared with the ground truth wound images. Two views of a successful result from the model are presented in Figures 3 and 4, indicating optimal synthetic wound generation from the segmentation of a wound by training with 4000 images and 2000 epochs. The generated wound image is combined with the rest of the limb. The generated wound tissue has a life-like structure, proper colour, and detailed texture. The colours of the wound tissues are well-matched and conformable.

Figure 5 shows the loss graphs of L1 loss of models with the different training datasets. The loss curves have similar

FIGURE 13 MSE scores (Median) of the models with the different training dataset



characteristics. The model with less training images learns the data distribution faster and stabilizes loss earlier than the model with more training images.

The generated image comparison with respect to epoch count and training dataset size is shown in Figures 7–12. An increase in the training dataset size helps to generate more plausible wound images. Blurry images become more realistic with an increase in dataset size given the same conditions, that is epoch count. These results show that a lack of training can be overcome by increasing the training dataset size. The effect of epoch count is also examined in Figures 7–12, where better wound images are produced with increased training. Figure 9A–D prove that the impact of epoch count overcome the deficiency of the training dataset. Further training yields better-wound images.

These extensive investigations suggest that the optimal performance for synthetic wound generation could be achieved using 1000 images and 200 epochs of training.

On the study results, the following observations regarding the application of the WG^2AN model could be made.

- Observation 1: The proposed model can perform synthetic wound generation from provided wound segmentation.
- Observation 2: WG^2AN has a high potential of producing close to real synthetic images.
- Observation 3: The quality of the generated images is in line with the image count. 1000 image count is the threshold for a valid generated image as the results of our study.
- Observation 4: The epoch count has a significant impact on the generated image quality. Yet, after surpassing a 200-epoch threshold, the model reaches its convergence, and additional training has a marginal effect on the quality of the generated image.
- Observation 5: The WG^2AN model can generate detailed tissue texture.
- Observation 6: Lack of training can be solved by increasing the training dataset.
- Observation 7: Scarcity of training images can be mitigated to some extent by further training of the model.

- Observation 8: Generated wound images can be combined with any part of the body to demonstrate the wound characteristics at that body location.

6 | CONCLUSION

This paper compiles a synthetic wound image generation using the proposed WG^2AN architecture. Synthetic wound generation using GAN is implemented for the first time in the literature. The wounds are segmented with respect to tissue types using semi-automatic machine learning techniques. The proposed model is then fed with segmented wound images to generate synthetic wounds.

With respect to patient privacy and lack of enough dataset for wound images in healthcare, the generation of synthetic anonymous wound images could enable further studies in AI, which could improve the adoption of AI in clinician training. L1 loss is also examined in order to understand the impact of training dataset size and epoch count, where the loss curve does not reveal much information other than models' learning behaviour. The generated images are examined and compared visually to evaluate their resemblance to a real wound. It can be concluded that the hardship of finding adequate training images in healthcare can be mitigated by additional model training. The effect of dataset size is also evaluated visually for further analyses. An increase in the training dataset brings in more life-like wound images.

The results of different dataset sizes and epoch count are evaluated through the MSE metric to compare the generated image. The generated images are expected to be very similar to actual wound images. This is why the MSE metric gives sufficient guidelines when comparing actual and synthetic images. The proposed model confirms that the 1000-image model with 200 epochs of training yields optimum results. Increasing these parameters could provide better synthetic images.

As future work, the segmentation of the wound could be done in more detail. In addition, the use of the proposed model in an education package could bolster the performance and practice of medical training.

Balancing underrepresented classes with synthetic image generation could help with the adoption of AI in the healthcare industry, where sourcing context-specific data is expensive. It is expected that this study can provide a handy clinician tool for generating and interacting with live wound models.

ACKNOWLEDGEMENTS

This work was supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development in Virginia. For more information about CCI, visit cyberinitiative.org.

REFERENCES

- Gaines, B.R.: Knowledge acquisition: Past, present and future. *Int. J. Hum. Comput. Stud.* 71(2), 135–156 (2013)
- Brunette, E.S., et al.: A review of artificial intelligence. In: 2009 4th International Conference on Autonomous Robots and Agents, Wellington (2009)
- Jiménez-Sánchez, A., et al.: Capsule networks against medical imaging data challenges. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pp. 150–160. Springer, Cham (2018)
- Pavlović, I., et al.: Comparison of paper-based and electronic data collection process in clinical trials: Costs simulation study. *Contemporary Clinical Trials* 30(4), 300–316 (2009)
- Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. *Intell. Data Anal.* 6(5), 429–449 (2002)
- Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 international interdisciplinary PhD workshop (IIPhDW), Świnoujście (2018)
- Wong, S.C., et al.: Understanding data augmentation for classification: When to warp?, In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast (2016).
- Hosseini, H., et al.: On the limitation of convolutional neural networks in recognizing negative images. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun (2017)
- Grajeda, C., et al.: Availability of datasets for digital forensics—and what is missing. *Digital Invest.* 22, S94–S105 (2017)
- Reimers, N., Gurevych, I.: Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. [arXiv:1707.06799](https://arxiv.org/abs/1707.06799) (2017)
- Hwangbo, S., et al.: Identification of hyperparameters with high effects on performance of deep neural networks: Application to clinicopathological data of ovarian cancer. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego (2019)
- Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay. [arXiv:1803.09820](https://arxiv.org/abs/1803.09820) (2018)
- Luo, Y., Lu, B.: EEG data augmentation for emotion recognition using a conditional wasserstein GAN. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu (2018)
- Han, C., et al.: GAN-based synthetic brain MR image generation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC (2018)
- Antoniou, A., et al.: Data augmentation generative adversarial networks. [arXiv:1711.04340](https://arxiv.org/abs/1711.04340) (2017)
- Nielsen, C.S.: Improving image classification through generative data augmentation. Master's thesis, Schulich School of Engineering (2019)
- Klang, O., Carlberg, M.: Blood cell data augmentation using deep learning methods. Master's Thesis in Mathematical Sciences, Lund University, Lund (2020)
- Mariani, G., et al.: Bagan: Data augmentation with balancing gan. [arXiv:1803.09655](https://arxiv.org/abs/1803.09655) (2018)
- Wang, C., et al.: CGAN-plankton: Towards large-scale imbalanced class generation and fine-grained classification. In: 2017 IEEE International Conference on Image Processing (ICIP), Beijing (2017)
- Frid-Adar, M., et al.: Synthetic data augmentation using GAN for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC (2018).
- Prokopenko, D., et al.: Unpaired synthetic image generation in radiology using gans. In: Workshop on Artificial Intelligence in Radiation Therapy, Cham: Springer (2019)
- Kim, K., Myung, H.: Autoencoder-combined generative adversarial networks for synthetic image data generation and detection of jellyfish swarm. *IEEE Access* 6, 54207–54214 (2018)
- Burlina, P.M., et al.: Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA ophthalmology* 137(3), 258–264 (2019)
- Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, Venice (2017)
- Bakkay, M.C., et al.: BSCGAN: Deep background subtraction with conditional generative adversarial networks. In: 2018 25th IEEE International Conference on Image Processing (ICIP), Athens (2018).
- Bills, J.D., et al.: Pilot study to evaluate a novel three-dimensional wound measurement device. *International Wound Journal* 13(6), 1372–1377 (2016)
- Srivastava, M., et al.: Comparison of optimizers implemented in Generative Adversarial Network (GAN). *International Journal of Pure Appl. Math.* 119(12), 16831–16836 (2018)
- Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu (2017)
- Löhdefink, J., et al.: Scalar and vector quantization for learned image compression: A Study on the Effects of MSE and GAN Loss in various spaces. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes (2020)
- Goodfellow, I.: NIPS 2016 tutorial: Generative adversarial networks. [arXiv:1701.00160](https://arxiv.org/abs/1701.00160) (2016)
- Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), Delft (2019)
- Gauthier, J.: Conditional generative adversarial nets for convolutional face generation. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester 2014(5), 2 (2014)

How to cite this article: Sarp, S., Kuzlu, M., Wilson, E., Guler, O.: WG²AN: Synthetic wound image generation using generative adversarial network. *J. Eng.* 2021;2021:286–294.
<https://doi.org/10.1049/tjc2.12033>