Journal of Research High School

Vol. 2025 (9) 767 – 774
ISSN 2688-3651 [online]

A Study of Data Augmentation Techniques to Overcome Data Scarcity in Wound Classification Using Deep Learning

Harini Narayanan[1 *]

[1] BASIS Independent Fremont, Fremont, CA, USA
*Corresponding Author: harini.sowmi.bb@gmail.com

Advisor: Sindhu Ghanta, sindhu@aiclub.world

Received February 8, 2025; Revised June 11, 2025; Accepted July 2, 2025

Abstract

Chronic wounds are a significant burden on individuals and the healthcare system, affecting millions of people and incurring high costs. Wound classification using machine learning (ML) techniques is a promising approach for faster diagnosis and treatment initiation. However, lack of high quality data to train the ML models is a major challenge to realize the potential of ML in wound care. In fact, data limitations are the biggest challenge in studies using medical or forensic imaging today. This study includes data augmentation techniques that can be used to overcome the data scarcity limitations and unlock the potential of deep learning based solutions. It also explores a range of data augmentation techniques from geometric transformations of wound images to advanced Generative Adversarial Networks (GANs), to enrich and expand datasets. Using the Keras, Tensorflow, and Pandas libraries, data augmentation techniques that can generate realistic wound images were implemented. The results show that geometric data augmentation can improve classification performance, F1 scores, by up to 11% on top of state-of-the-art models, across several key classes of wounds. The experiments with GAN based augmentation prove the viability of using Decoder-Encoder Generative Adversarial Networks (DE-GANs) to generate wound images with richer variations. The study and results show that data augmentation is a valuable privacy-preserving tool with huge potential to overcome the data scarcity limitations and be part of any real-world ML-based wound care system.

Keywords: Wounds classification, data augmentation, deep learning, Generative Adversarial Networks (GANs)

1. Introduction

Wounds, the modern day focus of extensive medical and forensic research, play an important role in quality of life, causing an immense amount of complications, ranging from small to large-scale. A chronic wound is defined as a wound that has been open for more than a month and that has not healed normally (Sen, 2021). Chronic wounds are a significant burden on individuals and the healthcare system, affecting millions of people and incurring high costs. Most common in elderly people, they affect around 8.2 million medicare beneficiaries in the United States, which shows the astounding ubiquity of wounds (Sen, 2021).

Chronic wound classes include, but are not limited to, diabetic, pressure, surgical, and venous ulcers. Each of these types of wounds are created by a multitude of factors, such as diabetes, surgery, and venous insufficiencies. More than being painful and inconvenient, chronic wounds are a financial burden as well. Costs of individual patient care alone for pressure ulcers is $20,900 minimum for each patient, with additional costs amounting to $43,180 per year (Sen, 2021). Given the substantial impact a single wound can have on an individual's quality of life, the cumulative effect on the larger population diagnosed with chronic wounds is significant. Therefore, developing methods and systems for efficient, effective and accurate wound care are of tremendous importance to society.

The length and complexity of the medical care process often lead to individuals avoiding medical consultation, exacerbating their wound's condition. Automated techniques based on deep learning algorithms can be used to classify

Vol. 2025 (9) 767 – 774
ISSN 2688-3651 [online]

wounds, allowing for quicker diagnosis and treatment initiation. They can assist doctors in saving time and focusing on more critical steps of wound care. Machine Learning (ML)-based wound classification can improve the quality of life in many people around the world.

Many studies have concluded that application of automated techniques in wound classification and care is essential for the forensic and medical industry's improvements (Chan et al., 2022; Dabas et al., 2023; Reifs et al., 2023). There were also many studies that demonstrated basic wound classification with different ML models. Rostami et. al (2021) used a Convolutional Neural Network (CNN) that classified between surgical, diabetic, and venous ulcers with high accuracies. Similar research conducted by Sarp et al. (Sarp et al., 2021), used transfer learning and XAI models that classified chronic wounds, similar to the research performed by Anisuzzaman et al. (2022). Chairat et al. (2023) worked on increasing the accessibility of wound care by using transfer learning models such as U-Net with EfficientNet and U-Net with MobileNetV2 (Sandler et al., 2019) to assess a wound and determine the rate of wound healing and made wound care more accessible to the general public by training their models on images taken by an iPhone at different locations, times of day, and brightness. Taking this one step further, subsequent studies went beyond just classifying wounds and looked at wound analysis such as wound severity and healing rate (Anisuzzaman et al., 2022; Hüsers et al., 2022; Gupta et al., 2023).

All these studies demonstrate the huge promise of deep learning models to transform wound care. However, data scarcity is a huge limitation in practical use of such techniques in wound care. This is explicitly stated in the conclusions of studies (Sarp et al., 2021; Rostami et al., 2021; Anisuzzaman et al., 2022), and in the discussions written by studies (Anisuzzaman et al., 2022; Hüsers et al., 2022). For example, both (Anisuzzaman et al., 2022; Hüsers et al., 2022) described in their future work that their model should be validated on a larger dataset. An efficient and scalable way to solve this problem is essential for continued research and successful application of deep learning techniques in the field of wound care. Data augmentation provides a scalable and privacy preserving technique to overcome this data scarcity and this study provides practical solutions to it.

This study takes a pragmatic view point and evaluates data augmentation techniques that can be used to build real-world deep learning based wound care systems. This study also demonstrates the viability of using GANs to generate wound images with richer variations.

2. Materials and Methods

This section describes the three major phases in the experiments: transfer learning (Bozinovski, 2020) from foundational CV models, geometric data augmentation, and GAN based data augmentation. Figure 1 visually shows the key steps and the three phases. To begin, the composition of the dataset is described, as well as how the dataset is split. This is followed by an outline of the training process, and detailed descriptions of the base model and data augmentation techniques implemented.

2.1 Dataset

The original dataset, obtained from another study (Anisuzzaman, 2022), had a very imbalanced dataset –some classes having as many as 2 times more images than others, as shown in Table 1. These classes included, background (BG), diabetic ulcers (D), not an ulcer (N), pressure ulcers (P), surgical ulcers (S), and venous ulcers (V). Initially, the category with the most images was the venous ulcers class, and the not an ulcer class had the least amount of images. Selective sampling was used to create a balanced dataset. This was accomplished by randomly choosing about 75 images from each category. Each one of these wound types have a very distinct appearance, as shown in Figure 2. Train and test sets were created by splitting the original dataset with an 80-20 split, respectively.

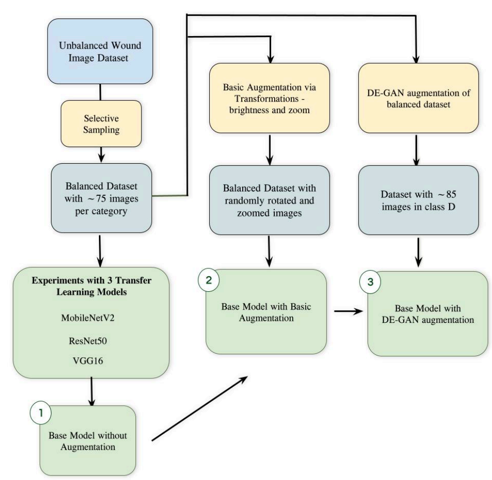


Figure 1. A flow chart representing the entire methodology employed in the study

Vol. 2025 (9) 767 – 774
ISSN 2688-3651 [online]

Table 1. Original Dataset with Number of Data Points

Wound Type	Number of Images
Venous (class V)	185 images
Background (class BG)	75 images
Diabetic (class D)	139 images
Not an Ulcer (class N)	75 images
Pressure (class P)	100 images
Surgical (class S)	122 images

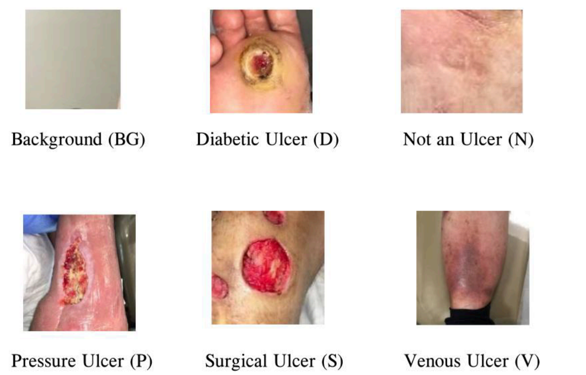


Figure 2. Example of All Categories of Wounds Addressed in this Study

Wound classification models using transfer learning

Over the years, a class of foundational computer vision models have emerged. These foundational models such as ResNet50 (He et al., 2015), VGG16 (Simonyan and Zisserman, 2015), MobileNetV2 (Sandler et al., 2019), can be adapted to domain specific tasks such as wound classification, by applying transfer learning. This study tests three of these foundational models with a vast range of hyperparameters to select the one that is most accurate (see Figure 1). Different combinations between 10, 30, and 50 epochs and learning rates of 0.01, 0.001, 0.0001, 0.05, and 0.0005 were also used. Though these models have high accuracy and are highly adaptable to specific tasks, due to the small dataset size, even the best models prove to have overfitting on the dataset.

Geometric Data Augmentation

An effective method to combat overfitting is data augmentation. In this study, geometric data augmentation was implemented by using multiple python libraries. Tensorflow was used to implement transformations on the train dataset that rotate images by a degree between 0 and 30, and change the brightness of an image to a value ranging from 0 to 0.2 randomly. The balanced dataset underwent these transformations to yield a new dataset with the same number of images, but with each image rotated or dimmed/brightened randomly. The most accurate model from phase 1 of experiments was then trained with this new, transformed dataset. These variations force the model to learn the features of the dataset, which helps combat overfitting. Although geometric data augmentation is effective in increasing the accuracy of the model to some extent, it is limited in variations it can produce. Generative modeling, such as GAN, can be used to overcome this limitation to create new, varied images for the dataset.

GAN-based data augmentation

Generative Adversarial Networks (GAN) with Decoder-Encoder Output Noise (DE-GAN) introduced by (Zhong et al., 2020) were used with the built-in loss function. The DE-GAN was trained on the training data set to create synthetic diabetic ulcer images to expand the training dataset (see Figure 3 and Equation 1). The new loss function shown in equation 1 used by the DE-GAN adds an additional component, L_{hid}, in addition to the loss function that all GANs use, L_{adv}. The DE-GANs use this loss function to gauge the quality of the image produced, and to consequently improve it. The DE-GANs were used to introduce variation and combat overfitting and confusion between classes, and they have also been proven to create higher quality images in less time than the naive GAN. Many hyperparameter

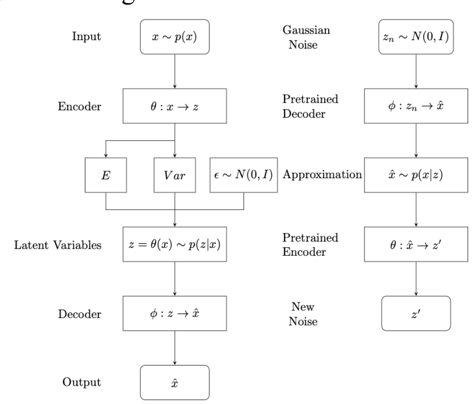


Figure 3. *A Flow Chart Representing the Parts of a DE-GAN (Zhong et al., 2020)*
Note. The right diagram represents the structure of the variational autoencoder in the DE-GAN, and the left diagram represents the structure of the naive GAN component of the DE-GAN.

Vol. 2025 (9) 767 – 774
ISSN 2688-3651 [online]

combinations were tested, from 3000-6000 epochs and many learning rates ranging from 0.0005 to 0.1. The images generated by the DE-GANs were then used to inflate the original training dataset for class D by 14 images. This new dataset was used to train the most accurate foundation model from the initial experiments.

Equation 1. *The Combined Loss Function Used by the DE-GAN as Described in (Zhong et al., 2020*

$$L = \lambda_1 L_{adv} + \lambda_2 L_{hid}$$

3. Results

The goals of the experiments are:

- Show transfer learning can be applied to adapt foundation CV models to wound classification
- Show that geometric data augmentation techniques improve accuracy
- Investigate the viability of DE-GAN techniques to successfully augment wound datasets and improve accuracy.

The results from the experiments run using the two types of data augmentation are reported and analyzed. A set of models were trained using the balanced dataset with 75 images per category and transfer learning from the base models. Three different sizes (small, medium, large) of base models (MobileNetV2, ResNet50, and VGG16, respectively) were chosen in order to provide for a good coverage of the sizes of models used. After experimenting with more than 15 different combinations of learning rates and number of epochs, a few top scores listed in table 2 emerged. The two models with top classification accuracies were VGG16 and ResNet50, with about 82% and 84% accuracies, respectively.

Table 2. Top Accuracies after Transfer Learning and Without Data Augmentation

Model	Epochs	Learning Rate	Accuracy
MobileNetV2	30	0.0005	0.64
VGG16	30	0.001	0.82
ResNet50	**37**	**0.0005**	**0.84**

3.1 Results with Data Augmentation

Although these models performed quite well with transfer learning on the original dataset with balancing, their confusion matrices shown in Figure 4 show that they have very low accuracy for class D and class P. This demonstrates one of the key limitations of applying transfer learning, which is that foundation models still need a good amount and variety of data to adapt to a new domain such as wound classification. To address this core problem, data augmentation was explored to combat overfitting and increase accuracy.

Geometric data augmentation provides significant improvement for classes D, P and V as shown in columns Resnet50 with transfer learning and Geometric augmentation in Table 3. Geometric augmentation provides 10%, 11%, and 7% improvements in F1 scores for classes D, P, and V, respectively (see Table 3) and these are precisely the classes that had low F1 scores with just transfer learning. This shows the power of Geometric data augmentation to improve upon transfer learning on state-of-the art CV models.

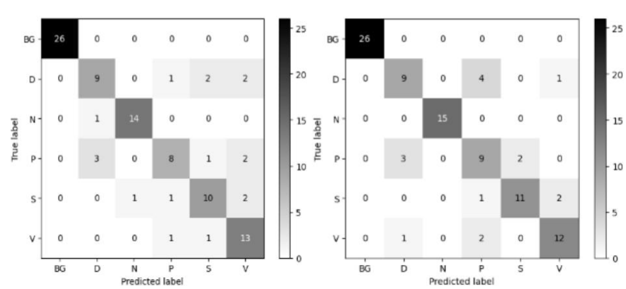


Figure 4. Confusion Matrix of Top Two Models. Note. The right image shows the confusion matrix of the base model VGG16 with 30 epochs and 0.0005 learning rate. The left image shows the confusion matrix of the base model ResNet50 with 37 epochs and 0.0005 learning rate.

Due to the high computation (GPU) requirements to train DE-GANs, one class was selected for experimentation to begin. DE-GAN based augmentation for class D was implemented to increase variation in the training dataset. Good variation in the DE-GAN generated images was observed, as shown in Figure 5 for class D. However, it was also observed that higher quality images did not directly translate into higher F1 scores (see Figure 6 for confusion matrices). Increasing the number of DE-GAN based augmented images and measuring the classification accuracy is an important future work for this model.

Vol. 2025 (9) 767 – 774
ISSN 2688-3651 [online]

Table 3. Precision, Recall and F1-scores for Foundation model (ResNet50) with Transfer Learning, Geometric Augmentation and DE-GAN Augmentation

Wound Classes	Scores	ResNet50 with xfer learning	Geometric aug	DE-GAN aug
BG	Precision	1.0	1.0	1.0
	Recall	1.0	0.93	0.93
	F1	1.0	0.97	0.97
D	Precision	0.69	0.83	0.71
	Recall	0.64	0.71	0.71
	F1	0.67	0.77	0.71
N	Precision	1.0	0.94	1.0
	Recall	1.0	1.0	1.0
	F1	1.0	0.91	1.0
P	Precision	0.56	0.67	0.63
	Recall	0.64	0.77	0.36
	F1	0.60	0.71	0.45
S	Precision	0.85	0.85	0.67
	Recall	0.79	0.79	0.71
	F1	0.81	0.81	0.69
V	Precision	0.8	0.82	0.67
	Recall	0.8	0.93	0.93
	F1	0.8	0.87	0.78

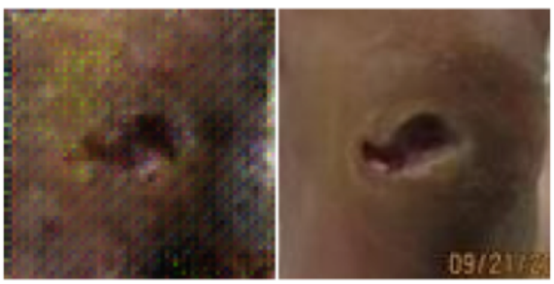


Figure 5. Sample Images Generated by DE-GAN Compared to the Original. *Note.* Left image: generated image of good quality. Right image: real wound image from original dataset.

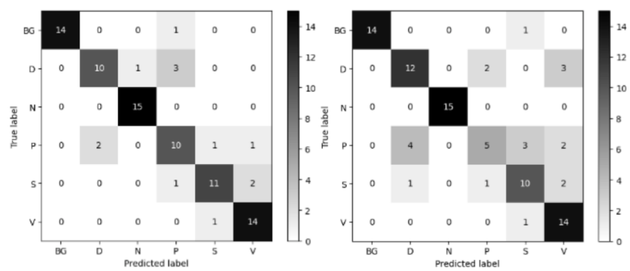


Figure 6. Results with geometric augmentation and GAN Augmentation. *Note.* The leftmost image shows the confusion matrix of the training using geometric augmentation, and the rightmost image shows the confusion matrix of the training using the DE-GAN augmented dataset.

4. Discussion

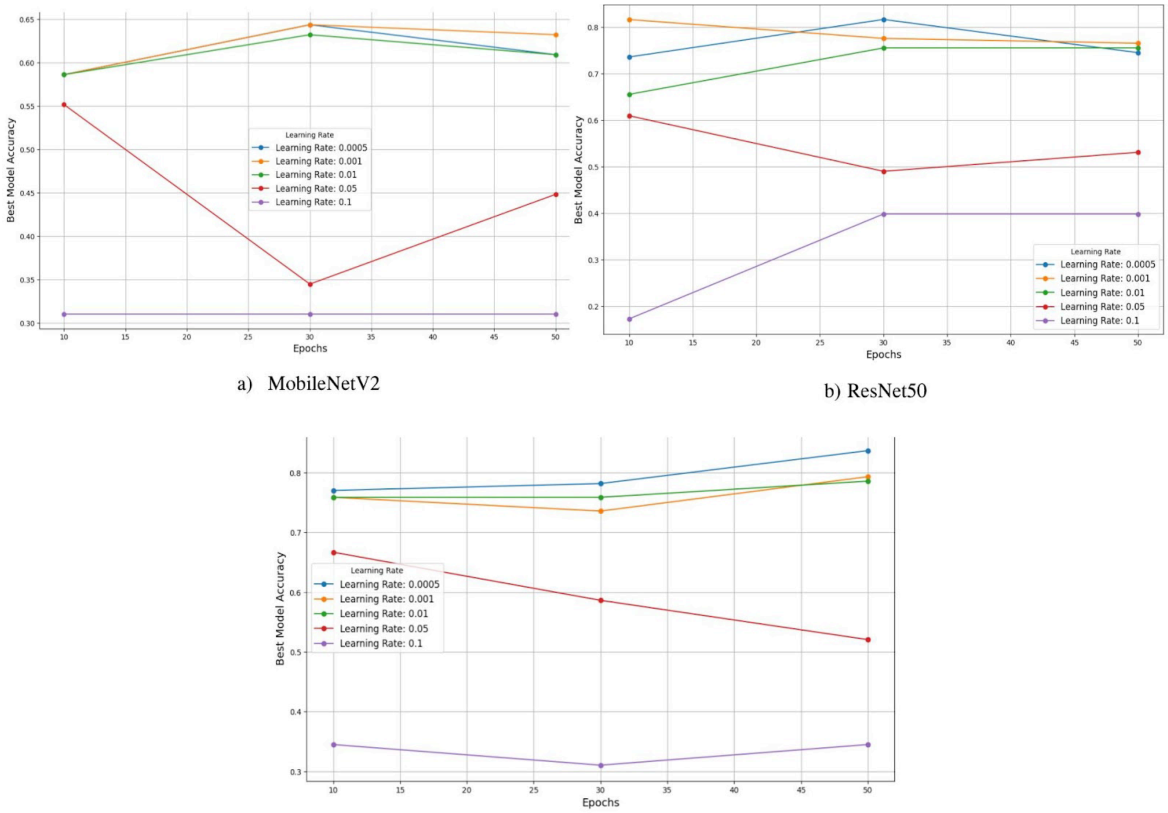
In this section, a few key observations from the experiments are discussed. Experiments were conducted using three foundational CV models, viz., MobileNetV2, VGG16, ResNet50. Hyper-parameter tuning of these models showed interesting trends in model performance. When trained for 50 epochs, there was evidence of early stopping during the transfer learning of all three foundation CV models, though these experiments provided the best results. This led to stopping the training of the model past 50 epochs. All three models were least accurate with 0.01 and 0.05 learning rate, the reason being that the learning rates were much too high and therefore too aggressive to use to train accurate models, as shown in Figure 7. On average, an increase in epochs also induced an increase in accuracy for all three models.

For the smaller MobileNetV2 model, accuracy remained relatively low across different learning rates when the number of epochs was a low number such as 10 (see Figure 7). Increasing the number of epochs to 30 showed a general improvement in model accuracy, with the best accuracy observed being 0.64 for the learning rate 0.0005.

A few different variations of the DE-GAN based approach were experimented with. These variations include using colored vs. black and white wound images, and colored images performed better. Adding more layers to the

Vol. 2025 (9) 767 – 774
ISSN 2688-3651 [online]

discriminator of the DE-GAN to increase the depth so that it can provide better signal to the generator was also explored, but this did not lead to better quality images. Finally, the architecture explained in section 3 was converged upon as the best model.



a) MobileNetV2

b) ResNet50

c) VGG16

Figure 7. Test Dataset Results with Different Hyperparameter Combinations

There were some challenges faced with model overfitting and DE-GAN structure. Overfitting was encountered with even some of the best transfer learning models without the augmentation. DE-GANs were very sensitive to large variations in the patterns in the input images and this affected the overall quality of generated images. This learning and the large computational needs of DE-GAN training, led to focusing all the efforts on one class (Class D) at a time. Another learning is that DE-GAN requires extensive tuning. Even after focusing on the Class D images, it took extensive hyperparameter search to create realistic wound images. There could be two reasons for this. The first is that the original dataset used to train the GANs had many different patterns (wound on toe versus wound on skin, etc.), which could have inhibited the generator's learning of the dataset. Mode collapse was also very commonly observed during training and led to generation of images that were all very similar to each other.

5. Conclusion

This study aimed at investigating data augmentation techniques that can be used for building real-world ML-based wound care systems. To ground the study on a real-world setting, state-of-the-art CV foundation models were used as a baseline. This study shows that foundational CV models such as MobileNetV2, ResNet50, and VGG16 can be successfully adapted for wound classification via transfer learning. Building on this, this study shows that geometric data augmentation techniques can provide significant (up to 11%) improvement on key classes of wounds (D, P and V). This study shows the viability of using DE-GAN based techniques to generate wound images with richer variations as compared to applying only geometric augmentations. Further study is needed to quantify the classification improvements of DE-GANs.

The use of DE-GANs to generate a wider variety of wound images and quantifying the classification performance is being actively explored. Another line of future work is experimentation with other types of GAN and training methods, such as the Wasserstein GAN with gradient penalty (Gulrajani et al., 2017), which has more stability during training and helps prevent mode collapse. Layering GAN-based augmentation on top of geometric augmentation is a very promising avenue for higher classification performance, which is also being actively explored.

References

Anisuzzaman, D., et al. (2022). Multi-modal wound classification using wound image and location by deep neural network. *Scientific Reports*, 12(1), 20057.

Anisuzzaman, D., et al. (2022). Wound severity classification using deep neural networks. *arXiv preprint* arXiv:2204.07942.

Bozinovski, S. (2020). Reminder of the First Paper on Transfer Learning in Neural Networks, 1976. *Informatica*, *44*(3). https://doi.org/10.31449/inf.v44i3.2828

Chan, K. S. et al. (2022). Clinical validation of an artificial intelligence-enabled wound imaging mobile application in diabetic foot ulcers. *International Wound Journal*, 19(1), 114-124.

Chairat, S., et al. (2023). AI-assisted assessment of wound tissue with automatic color and measurement calibration on images taken with a smartphone. *Healthcare*, 11, 273.

Dabas, M., et al. (2023). Application of artificial intelligence methodologies to chronic wound care and management: A scoping review. *Advances in Wound Care*, 12(4), 205-240.

Goodfellow, I., et al. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2672-2680).

Gulrajani, I., et al. (2017). Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30.

Gupta, R., et al. (2023). Towards an AI-based objective prognostic model for quantifying wound healing. *IEEE Journal of Biomedical and Health Informatics*, 28(2), 666-677.

Gutbrod, M., Geisler, B., Rauber, D., & Palm, C. (2024). Data augmentation for images of chronic foot wounds. In A. Maier, T. M. Deserno, H. Handels, K. Maier-Hein, C. Palm, & T. Tolxdorff (Eds.), Bildverarbeitung für die Medizin 2024 (pp. 261-266). Springer Fachmedien Wiesbaden.

Häusers, J., et al. (2022). Automatic classification of diabetic foot ulcer images–A transfer-learning approach to detect wound maceration. In Informatics and Technology in Clinical Care and Public Health (pp. 301-304). IOS Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognitio*n. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Kairys, A., & Raudonis, V. (2023). Analysis of training data augmentation for diabetic foot ulcer semantic segmentation. *Electronics*, 12(22).

Liu, Z., et al. (2024). Chronic wound image augmentation and assessment using semi-supervised progressive multi-granularity EfficientNet. *IEEE Open Journal of Engineering in Medicine and Biology*, 5, 404-420.

Lo, Z. et al. (2024). Development of an explainable artificial intelligence model for Asian vascular wound images. *International Wound Journal*, 21(4), e14565.

Vol. 2025 (9) 767 – 774
ISSN 2688-3651 [online]

Niri, R., et al. (2021). Multi-view data augmentation to improve wound segmentation on 3D surface model by deep learning. *IEEE Access*, 9, 157628-157638.

Reifs, D., et al. (2023). Clinical validation of computer vision and artificial intelligence algorithms for wound measurement and tissue classification in wound care. *Informatics in Medicine Unlocked*, 37, 101185.

Rostami, B., et al. (2021). Multiclass wound image classification using an ensemble deep CNN-based classifier. *Computers in Biology and Medicine*, 134, 104536.

Sandler, M., et al. (2019). MobileNetV2: Inverted residuals and linear bottlenecks.

Sarp, S., et al. (2021). The enlightening role of explainable artificial intelligence in chronic wound classification. *Electronics*, 10(12), 1406.

Sen, C. K. (2021). Human wound and its burden: Updated 2020 compendium of estimates. *Advances in Wound Care*, 10(5), 281-292.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1-48.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Yun, S., et al. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features.

Zhao, T., Gu, J., & Zhang, Q. (2023). A data augmentation method for wound infection prediction. In *2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* (pp. 85-91).

Zhong, G., et al. (2020). Generative adversarial networks with decoder–encoder output noises. *Neural Networks*, 127, 19-28.