Augmenting Data with DCGANs to Improve Skin Lessions Classification

No Author Given

No Institute Given

Abstract. One of the main problems in computer vision applied to medical field is the limited amount of data. This represents a limitation in making an early diagnosis of diseases. Therefore, synthetic data has proven significantly crucial in minimizing the problems of acquiring medical images. This paper proposes using Generative Adversarial Networks (GANs) to generate synthetic skin lesion data. These data are helpful in the process of classifying benign and malignant skin lesions such as melanoma. Hence, the Deep Convolutional Generative Adversarial Network (DCGAN) was used to produce malignant synthetic images. An Inception V3 classification network was used to evaluate the impact of adding this synthesized data to a binary classification task. With the increase of synthetic data to the malignant class, an accuracy score of 0.88 was obtained. These synthetic images were used in the training set to improve the final performance of the classification network for skin lesions. The achieved results were compared with the classification of the original data. They showed that the augmentation of images generated by the DCGAN improves the network's performance for the classification task.

Keywords: GAN · DCGAN · Melanoma · Image classification · Data augmentation.

1 Introduction

Detection of medical pathologies is one of the significant challenges in Artificial Intelligence for healthcare. One of the topics with a meaningful presence in research is the early detection of cancer. In 2020, the World Health Organization (WHO) [1] reported more than 10 million deaths in 2020 related to different types of cancer. In particular, they report around 1.20 million cases of skin cancer. Additionally, in 2023 the American Cancer Society [2] estimates that only in the United States about 97,610 new melanomas will be diagnosed, and about 7,990 people may die because of this disease.

Melanoma is one of the most common skin cancer pathologies, with a death rate of 75% incidence [3]. In medical terms, melanoma is a malignant neoplasm originating from melanocyte cells. Those are skin cells responsible for the production of melanin. Melanin is the pigment responsible for skin, hair and eye color. Melanoma occurs when melanocytes begin to grow and multiply abnormally and become malignant [4].

The challenge with this pathology is classifying benign and malignant skin lesions because there are different clinical subtypes, and it isn't easy to differentiate between them. In addition, if the pathology is detected early, it should be treaty only with rapid resection, making fast diagnosis extremely important. To look for a quick and automatic melanoma classification tool, the International Skin Imaging Collaboration (ISIC) has released a large-scale publicly accessible dataset of dermoscopic images [5]. The main objective of ISIC is to promote collaboration and knowledge sharing among experts in dermatology, research and technology to advance the detection and diagnosis of skin diseases. They even provide competencies in skin lesion analysis, including segmentation, feature extraction and lesion classification [6].

Although some organizations continue to release datasets, the amount of images for specific malignant pathologies is still limited [7]. Because of this, synthetic images seem to be an alternative to increase the number of samples for datasets. However, generating synthetic images related to medicine is difficult for some reasons: the anatomical structure features a large amount of detail, individual patient characteristics, accuracy, realism, and the protection of patient privacy. Some Deep Learning and Computer Vision techniques currently focus on generating synthetic images through different approaches [8]. One of them are the Generative Adversarial Networks (GANs) [9]. GANs can be helpful in medicine where limited data is available and make the task hard to identify pathologies. This tool has proven to help generate realistic images that contribute to creating relevant data for various medical studies. Therefore, synthetic images help improve performance in tasks like classification and segmentation of pathologies.

In this work, a study about the use of GANs to generate synthetic images for skin cancer classification is presented. The ISIC dataset is used, and the number of images in the training phase is increased through synthetic images generated by GANs. To evaluate the impact of the GANs in the performance of skin cancer classification, an Inception V3 architecture is used. The evaluation consists of two main stages: The first one is related to training the Inception model only with the original and unbalanced data retrieved from the ISIC dataset. The second stage is training the Inception model using the dataset augmented with synthetic images generated using GANs. This work's core objective is to generate images of skin lesions in the most realistic way possible.

2 Related Works

2.1 Image Classification on Skin Lessions

Image classification is a Computer Vision (CV) task that assigns a specific label or category to an image according to the information retrieved from an input image. In the medical field, recognizing and distinguishing different pathologies can be crucial.

Li *et al.* [10] proposed a FCRN-50 and FCRN-101 architectures for classifying melanoma, seborrheic keratosis and nevus images. Also, they constructed a

Lesion Indexing Network (LIN) for skin lesion image analysis. Similarly, Yilmaz *et al.* [11] took data from ISIC 2017 challenge and applied MobileNet, MobileNetV2, and NASNetMobile with transfer learning to perform the classification task. They obtained the best performance with the NASNetMobile with a batch size of 16 and got an accuracy of 82%.

The authors in [12] presented an experiment with different neural networks for classifying benign and malignant skin lesions. They implement a PNASNet-5-Large, InceptionResNetV2, SENet154, and InceptionV4. Data pre-processing and data augmentation was performed. The pre-processing consists of normalization to convert the pixel values into 0 and 1. For data augmentation, transformations were applied to reduce possible performance loss due to the dataset imbalance. The transformations applied to the images were rotation, random crop, brightness and contrast adjustment, pixel jitter, aspect ratio, random shear, zoom, and vertical and horizontal shift and flip. Better results were obtained by the PNASNet-5-Large model, which has a 0.76 in validation score. Furthermore, Devries *et al.* [13] used a multi-scale convolutional network to classify data retrieved from ISIC 2017 skin lesion classification challenge. They performed a fine-tuning using a pre-trained Inception-v3 model trained in the ImageNet dataset to get better results. Also, it was important to add images from the ISIC_MSK2_1 dataset to improve the results of melanoma prediction. As a result of the training, they obtained an accuracy of 0.893 for melanoma, 0.913 for seborrheic keratosis, and an average of 0.903 for the general classification task.

2.2 ISIC Data Augmentation with GANs

In recent years, Deep Learning techniques have proven to help offer efficient alternatives for data augmentation. Mainly, the use of Generative Adversarial Networks (GANs) helps generate synthetic images similar to the real data. Pollastri *et al.* [14] proposed using GANs to augment data in the skin lesion segmentation task. They use the 2017 ISIC dataset in this implementation. For this purpose, the authors implemented a Deep Convolutional Generative Adversarial Network (DCGAN) and Laplacian GAN (LAPGAN) to generate the synthetic data. Ultimately, they used a Convolutional-Deconvolutional Neural Network (CDNN) to measure the accuracy improvement by adding the synthetic data into the training process. The DCGAN was better for one experiment, while the LAPGAN performed best in four experiments. Both achieve an improvement near 1%. In addition, using a U-Net with original and synthetic data showed better results (a gain of about 1%

Additionally, Baur *et al.* [15] proposed generating realistic and high-resolution images of skin lesions with GANs. They use the 2017 ISIC dataset with three classes: benign lesions, seborrheic keratosis samples and melanoma. They implemented a DCGAN and LAPGAN to generate the synthetic data. Those images were used to train a variety of classifiers for skin lesions. For the classification task, a ResNet-50 was used with different variations of the original data and adding the synthetic images generated by GANs. Finally, they achieved a near

1% more accuracy score with the images from LAPGAN in training (99.29%) and validation (74.00%) sets.

Another work was presented by Rashid *et al.* [7]. They proposed using GANs to solve the problem of the limited amount of data in medical imaging. They use a Vanilla GAN to generate realistic images using the 2018 ISIC dataset for this task. Also, an interesting feature of this GAN is that it can also be used for classification. Therefore, for the classification task, they used fine-tuned DenseNet and ResNet architectures together with the Vanilla GAN as baseline models. The metric used to compare the experiments' performance is the balance accuracy score due to the unbalance data presented. Consequently, they achieve a balance accuracy score of 0.815 with DenseNet, 0.792 with ResNet, and 0.861 with GANs.

There are also more modern applications of GANs that offer more significant opportunities for generating images with great detail and style. Limeros *et al.* [16] implemented a StyleGAN2-ADA using an original implementation from the NVIDIA Research group. This was performed using ISIC 2020 and ISIC 2019 datasets. The classification task was performed using the EfficientNet-B2 model pre-trained on ImageNet. They balance the malignant class by adding 22.000 synthetic images. After this, they performed different classification scenarios to compare the results with baseline and augmented data. In all cases, they tested on the same real image validation set. As a result, the authors achieved better performance, with the augmented data getting 0.979 accuracy.

3 System Model and Methodology

3.1 Dataset

The dataset is provided by the International Skin Imagin Collaboration (ISIC) [17]. It was designed to release digital skin images to produce tools that help to reduce skin cancer mortality. They provide an extensive image gallery and filters to search through the dataset, letting us browse inside a lot of public images. For the purpose of this work, we only focused on the first filter of the gallery, which is diagnostic attributes. With this filter, we retrieved 8833 benign and 7361 malignant images. Each image has a size of 1769x1769. Figure 1 shows some examples of this dataset.

3.2 Proposed Model

Deep Convolutional Generative Adversarial Network (DCGAN) Deep Convolutional Generative Adversarial Networks (DCGANs) are based on the operation of traditional GANs but by adding convolution layers to replace the multilayer perceptron. The convolution layer part discriminates between the images received by the discriminative network. Meanwhile, the deconvolution layer part generates the images in the generative network [18]. DCGANs use various techniques to improve the training phase. These techniques are the following:

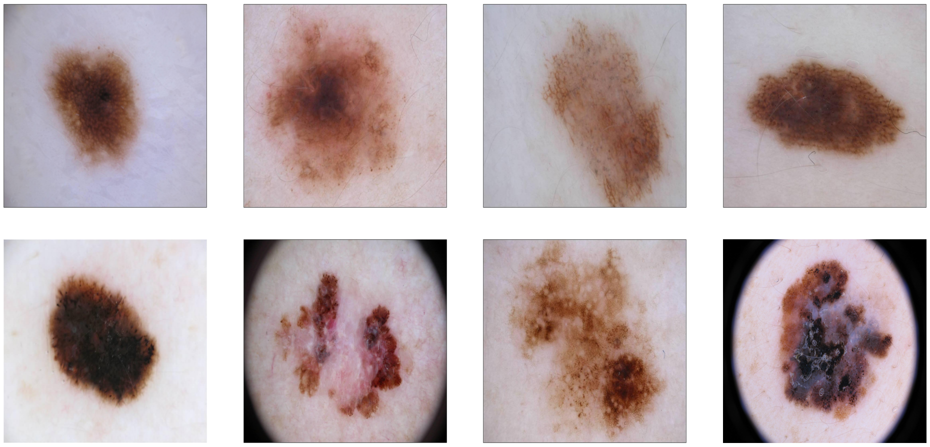

Fig. 1: Examples of ISIC dataset. First row: bening images. Second row: malignant images.

convert max-pooling layers to strides convolution layers for the discriminator and convert fractional-strides convolutions for the generator; convert fully connected layers to global average pooling layers in the discriminator; using batch normalization layers either in the generator and in the discriminator (except the output layer for generator and input layer for discriminator); the using of rectified linear unit (ReLu) in the generator (except for the output which uses hyperbolic tangent function (Tanh); and the using of leaky ReLU activation functions in the discriminator.

This generative model derives from the original Generative Adversarial Network (GAN) proposed by [9], which has a generative and discriminative network. The details of the implementation of botch networks are described below:

1. DCGAN Generator: The generator network takes a 100-dimensional uniform distribution Z random noise input to feed a fully connected layer that resizes it and converts it to a suitable size for the next layer in the network. Next, transposed convolution layers (unsampling layers) are used to increase the dimension of the input tensor up to the desired size. This increase occurs for the height and width, whereas there is a reduction in channel dimensions. These transposed convolution layers help to decompress the noise and transform it into a more detailed representation of the information. After each transposed convolution stage, a Batch Normalization is applied, which helps to normalize the output tensor by adjusting the mean and variance to stabilize the training process. A Rectified Linear Unit (ReLu) activation function is applied to introduce nonlinearity into the network. This process of using transposed convolutions, normalization and activation is repeated several times on the generator network to adjust and refine the representation of the generated images into the network to be as similar as possible to

the real data. Finally, for the output layer, a transposed convolution with the hyperbolic tangent function (Tanh). In the case of using color images, the output is a 3-dimensional tensor representing the color channels (red, green and blue) [18]. The general architecture of this network is shown in Figure 2.

2. DCGAN Discriminator: The discriminative network takes either a synthetic image generated by the generator or a real image extracted from the training dataset as input. This image goes through several convolution layers that aim to extract relevant features from the image. These features can be edges, textures or shapes that are specific image patterns. After each convolution layer, a Batch Normalization layer is applied to join to a LeakyReLy activation function. Then, the Sigmoid activation function generates a final probability as output. For this step, fully connected layers are used to classify the features and produce the final output, which is the probability that an image be real or fake using a binary classification (0 indicates that the image is fake, 1 indicates that the image is real) [18].

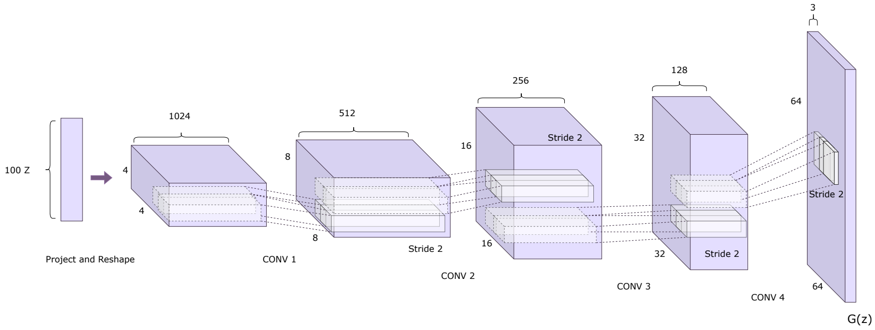

Fig. 2: DCGAN Generator Architecture. A 100-dimensional uniform distribution (Z) is projected to a small spatial convolutional representation with some feature maps. This is followed by four fractionally-strided convolutions. Finally, this converts to a representation of 64x64 pixel image [18].

3.3 Evaluation Metrics

This section describes metrics to evaluate the synthetically generated images' quality and the classification task performance. First, we mentioned the metrics for the evaluation of the quality of the synthetic generated images from the DCGAN:

1. Fréchet inception distance (FID): Several metrics are used to evaluate generative models, including the FID. It is used to calculate the distance between

feature vectors calculated for real and synthetic generated images [19] i.e., and it compares the features extracted from the generated synthetic images with the features of the real image set. A pre-trained InceptionV3 neural network is used to perform the calculation of the FID [20]. With the extracted feature vectors given by the InceptionV3, the distribution of generated feature vectors versus the distribution of the feature vectors from the original images is calculated and compared using the Fréchet distance [21]. A low FID value represents a better quality of the generated images since the characteristics are similar to those of the original images.

2. Inception Score (IS): Another metric used to measure the quality of generative model images is the Inception Score. The IS works with an InceptionV3 pre-trained on ImageNet to calculate some statistics of the network's outputs [22]. After generating a set of synthesized images, each of them is sent to the InceptionV3 network. This evaluates the probability of each generated image belonging to the set of real images. Additionally, the entropy of the probability distributions is calculated to measure the diversity of the generated images [23]. Hence, a high value of IS indicates a higher quality and diversity of images.

Second, we mentioned the metrics used for the evaluation of the performance of the InceptionV3 network for the classification task:

1. Accuracy: It represents the proportion of correct predictions out of the total number of samples evaluated.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

2. Precision: It represents the number of positive predictions out of the total number of true positive samples.

$$precision = \frac{TP}{TP + FP} \tag{2}$$

3. Recall: It represents the proportion of positive samples correctly identified.

$$precision = \frac{TP}{TP + FN} \tag{3}$$

4. F1-Score: It is a measure that takes into account precision and recall. It helps to evaluate the correct prediction of true positive samples. This is useful to identify classification errors by class, especially when there is an imbalance between classes.

$$F1\,Score = 2\ \cdot\ \frac{Precision\ \cdot\ Recall}{Precision\ +\ Recall} \tag{4}$$

3.4 Hardware Acceleration

The training process of the DCGAN was performed using an NVIDIA GeForce RTX 4090 graphical processing unit (GPU) with 24 GB of VRAM and 21 Gbps, together with a 5.40GHz Intel ®Core i7-13700K CPU and 32 GB of RAM DDR5.

4 Results and Discussion

4.1 First approach: Classification without DCGAN augmentation

This section describes the characteristics of the training phase for the classification model with the original data obtained from the ISIC dataset. This step is essential first to know how the classification task performs without any increase in data and then to have a point of comparison for the training performance with the augmented dataset with synthetic images. For the classification task, we get an InceptionV3 model, trained during 20 epochs, with a batch size of 128, a learning rate of 0.001, Adagrad as optimizer, and Cross-Entropy as loss function. Data augmentation techniques include random horizontal and vertical flips and random rotation. All the input data was normalized before entering the network. With these parameters, we obtained the results shown in Table 1, computed from the confusion matrix values presented in Figure 3.

Table 1: Precision, recall and f1-score values for the classification of ISIC dataset.

Classes	precision	recall	f1-score
Bening	0.89	0.89	0.89
Malignant	0.87	0.86	0.86

4.2 Second approach: Classification with DCGAN augmentation

In this part, we present the results obtained by balancing the dataset by adding synthesized images generated by DCGAN. First, 1176 images were added to the malignant class of the original dataset. This balances both classes to 8833 images each. Second, the same parameters as in section 4.1 were used for the classification task with the InceptionV3 network. It is important to highlight this point to demonstrate that the improvement was obtained following the same hyperparameters. The results by class obtained for the classification with synthetic data augmentation are shown in Table 2. In addition, the confusion matrix is displayed to determine the amount of correctly classified new augmented dataset in Figure 4.

With the values obtained for each dataset class, the accuracy score for each approach can be obtained. This metric allows us to measure the overall performance of the network. The results are shown in Table 3.

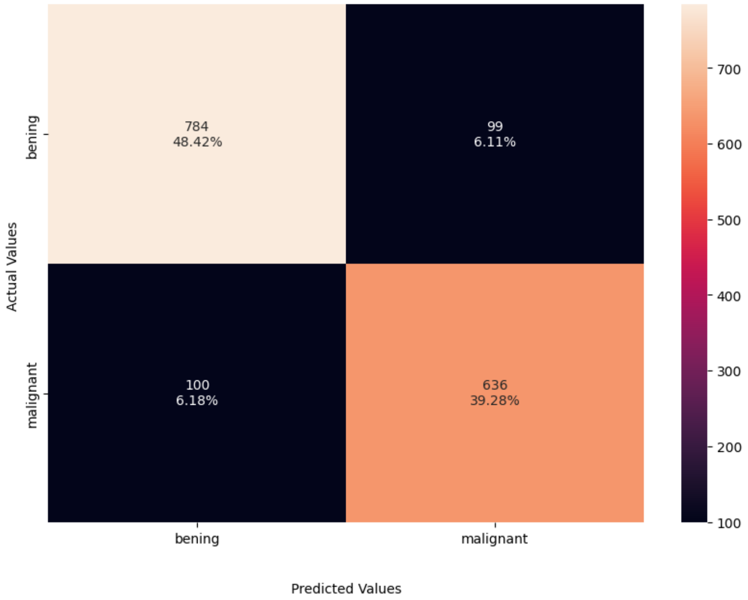

Fig. 3: Confusion matrix for the ISIC dataset classification.

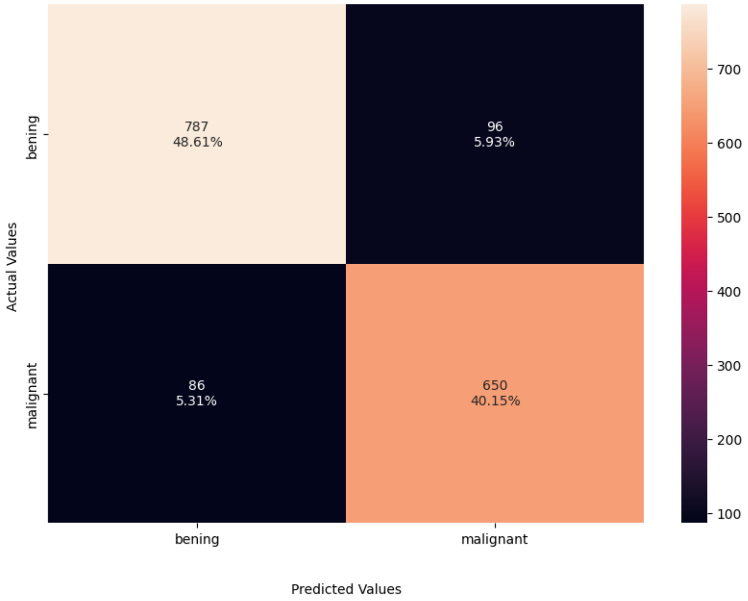

Fig. 4: Confusion matrix for the ISIC dataset classification augmented with DCGAN synthetic images.

Table 2: Precision, recall and f1-score values for the classification of ISIC dataset augmented with DCGAN synthetic images.

Classes	precision	recall	f1-score
Bening	0.90	0.89	0.90
Malignant	0.87	0.88	0.88

Table 3: Performance comparison using InceptionV3 for the original ISIC dataset and DCGANs augmented dataset.

Approach	Accuracy Score
Original Dataset	0.8770
DCGANs Augmented Dataset	**0.8875**

4.3 DCGAN training and Synthetic image quality evaluation

The DCGAN network was trained during 300 epochs, with an image size and a batch size of 64. The evolution of the loss function both for the discriminator and generator networks is shown in Figure 5. On the one hand, the value of the loss function began to increase until approximately 2500 iterations. After that, this value begins to decrease. This indicates that the network began to produce better synthetic images and that the discriminator began to classify synthetic generated images as real. On the other hand, the values of the discriminator loss function began to have a minimal increase due to the incorrect classification images performed by the discriminator.

Finally, the quantitative and objective evaluation of image quality using FID and IS is presented in Figure 6. The FID value until the 50th epoch is high. After this epoch, this value begins to decrease. On the contrary, the IS value before the 50th epoch begins small. After this epoch, this value begins to increase. The evolution of both values indicates that the synthetic images produced by the generator are good enough, as we can see in Figure 7 where a comparison between real and synthetic images generated with the DCGAN is shown.

4.4 Discussion

It is shown how the imbalance in a dataset can hurt performance in a classification model. In the same way, it is shown how we can fix this problem using GANs to generate synthetic images to balance the dataset. In particular, if we use the original unbalanced data from the ISIC dataset for the skin lesion classification, we got an accuracy score of 0.8770. Then, if we add 1176 images generated by DCGAN to the malignant class, an accuracy score of 0.8875 was obtained, which is higher. In addition, other metrics show a better performance of the model. For the benign class, there was an increase in precision (0.89 vs. 0.90) and f1-score (0.89 vs. 0.90). For the malignant class, there was an increase in recall (0.86 vs. 0.88) and f1-score (0.86 vs. 0.88). These improvements, however small they seem,

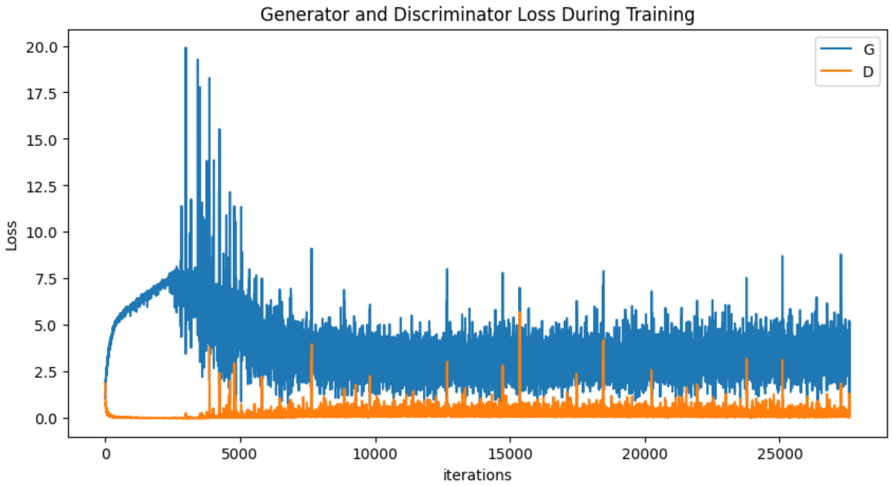

Fig. 5: Evolution of the loss function for the generator and discriminator during training phase of the DCGAN for the ISIC dataset into malignant class.

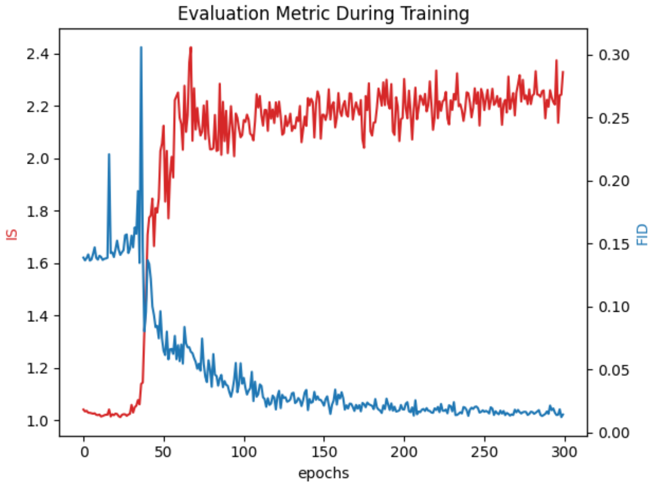

Fig. 6: Evolution of FID and IS metrics during the DCGAN training phase.

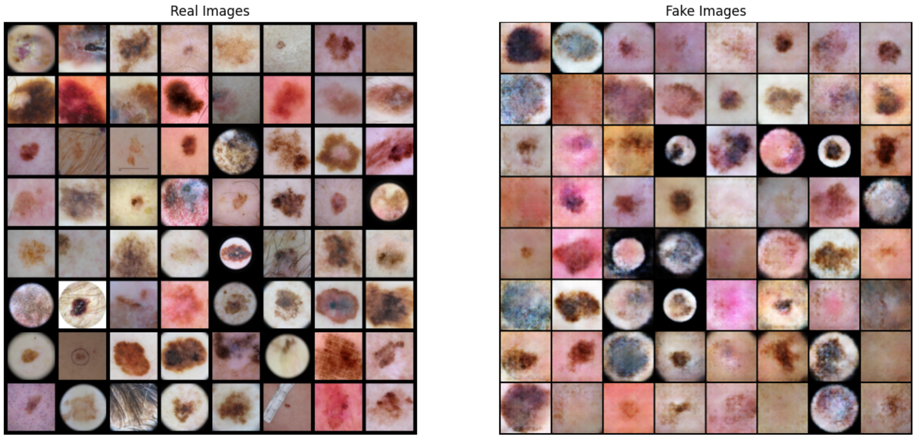

Fig. 7: Comparison between real images and synthetic images generated by the DCGAN. Left batch: real images from the ISIC dataset in the malignant class. Right batch: synthetic images for the malignant class.

help make correct diagnoses, reduce false positives and negatives, and improve disease management with early diagnosis.

In addition, the confusion matrices can be observed the reduction of false positives and the increase in true positives values. Similarly, the reduction of false negatives and the increase in true negatives values can be observed. For the malignant class, the improvement in value classification is more significant. Finally, favorable results were obtained for quantifying image quality employing FID and IS. The distance between the generated synthetic images and the original images was close enough to deduce that the generated images have similar qualities as the original ones. It is represented by a low FID score and a high IS score, which not only represent values for image quality but also for image diversity.

5 Conclusion

In this paper, we present the implementation of a DCGAN for data augmentation and demonstrate that such data helps to improve classification tasks. In particular, we address the study of classifying images of skin lesions. However, this technique can be extended to many other case studies in the medical industry. This is beneficial since medicine is a field where data is limited. Data augmentation using GANs has been shown to increase the accuracy rate for the skin lesion classification task. This is demonstrated by comparing the results obtained in the skin lesion classification model training with the original images from the ISIC dataset and the dataset with the synthesized data enhancement produced by GAN for the malignant class. In the same way, the quality of the generated images was proved by evaluating them using the FID and IS metrics.

6 Future Works

Future work could include using skin lesion synthetic images with other existing GAN models. A comparison between the presented model and other existing models for generating synthetic images would be interesting to differentiate the characteristics of the generated images. In addition, an important task is to work with a larger number of images. This could be used to evaluate whether there is a better performance for classifying skin lesions.

References

1. Cancer, https://www.who.int/news-room/fact-sheets/detail/cancer,2022, accessed on:21-JUN-2023
2. American Cancer Society. Key statistics for melanoma skin cancer. 2023. Accessed: 21-JUN-2023
3. Saginala, Kalyan, et al.: Epidemiology of melanoma. Medical sciences 9.4: 63 (2021)
4. Adegun, Adekanmi A., and Serestina Viriri.: Deep learning-based system for automatic melanoma detection. IEEE Access 8, 7160-7172 (2019)
5. Combalia, Marc, et al.: Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. The Lancet Digital Health 4.5, e330-e339 (2022)
6. Wen, David, et al.: Characteristics of publicly available skin cancer image datasets: a systematic review. The Lancet Digital Health 4.1, e64-e74 (2022)
7. Rashid, Haroon, M. Asjid Tanveer, and Hassan Aqeel Khan.: Skin lesion classification using GAN based data augmentation. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), (2019)
8. Mahmood, Faisal, Richard Chen, and Nicholas J. Durr.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE transactions on medical imaging 37.12, 2572-2581 (2018)
9. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
10. Li, Yuexiang, and Linlin Shen.: Skin lesion analysis towards melanoma detection using deep learning network. Sensors 18.2, 556 (2018)
11. Yilmaz, Abdurrahim, et al.: Benchmarking of Lightweight Deep Learning Architectures for Skin Cancer Classification using ISIC 2017 Dataset. arXiv preprint arXiv:2110.12270, (2021)
12. Milton, Md Ashraful Alam.: Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. arXiv preprint arXiv:1901.10802, (2019)
13. DeVries, Terrance, and Dhanesh Ramachandram.: Skin lesion classification using deep multi-scale convolutional neural networks. arXiv preprint arXiv:1703.01402, (2017)
14. Pollastri, Federico, et al.: Augmenting data with GANs to segment melanoma skin lesions. Multimedia Tools and Applications 79, 15575-15592 (2020)
15. Baur, Christoph, Shadi Albarqouni, and Nassir Navab.: MelanoGANs: high resolution skin lesion synthesis with GANs. arXiv preprint arXiv:1804.04338, (2018)
16. Limeros, Sandra Carrasco, et al.: GAN-based generative modelling for dermatological applications–comparative study. arXiv preprint arXiv:2208.11702, (2022)

17. The International Skin Imaging Collaboration, ISIC, [On line]. Available: Available: https://gallery.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D. [Accessed on:1-JUN-2023].
18. Radford, Alec, Luke Metz, and Soumith Chintala.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, (2015)
19. Heusel, Martin, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30, (2017)
20. Jung, Steffen, and Margret Keuper.: Internalized biases in fréchet inception distance. NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, (2021)
21. Mathiasen, Alexander, and Frederik Hvilshøj.: Backpropagating through Fréchet Inception Distance. arXiv preprint arXiv:2009.14075. (2020)
22. Barratt, Shane, and Rishi Sharma.: A note on the inception score. arXiv preprint arXiv:1801.01973, (2018)
23. Song, Yang, and Stefano Ermon.: Improved techniques for training score-based generative models. Advances in neural information processing systems 33, 12438-12448 (2020)