Augmenting data with DCGANs to improve skin lessions classification

No Author Given

No Institute Given

Abstract. In the medical field, one of the biggest challenges is the limited amount of data. This represents a limitation to make an early diagnosis of diseases. Therefore, the use of synthetic data has proven to be significantly important to minimize the problems related to the acquisition of medical images. In this paper we propose the use of Generative Adversarial Networks (GANs) to generate synthetic skin lesion data. These data are useful 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 these 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 and showed that the augmentation of images generated by the DCGAN improves the performance of the network for the classification task.

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

1 Introduction

Detection of medical pathologies is one of the major challenges in Artificial Intelligence for healthcare. One of the topics with meaningful presence in research is related to the early detection of cancer. In 2020, the World Health Organization (WHO) [14] reported that there are more than 10 million deaths in 2020 related with different types of cancer. In particular, they report around 1.20 million cases for skin cancer. Additionally, in 2023 the American Cancer Society [15] 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.

In fact, melanoma is one of the most common skin cancer pathologies, with a death rate of 75% incidence [10]. In medical terms, melanoma is a malignant neoplasm originating from melanocytes 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 [11].

The challenge with this pathology is to classify benign and malignant skin lesions because there are different clinical subtypes and it is difficult differentiate between them. In addition, if the pathology is detected in early stage, it should be treaty only with rapid resection making fast diagnosis extremely important. In order to look for a fast an automatic melanoma classification tool, the International Skin Imaging Collaboration (ISIC) has released a large-scale publicly accessible dataset of dermoscopic images [12]. 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 on skin lesion analysis including segmentation, feature extraction and lesion classification [13].

Although some organizations continue to relase datasets, the amount of images for certain malignant pathologies is still limited [3]. 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. Currently. there are some techniques in Deep Learning and Computer Vision that focus in the generation of synthetic images thought different approaches [5]. One of them are the Generative Adversarial Networks (GANs) [6]. GANs can be help full in medicine where limited data is available and makes hard the task to identify pathologies. This tool has proven to be useful in the generation of realistic images that contribute to the creation of relevant data for various studies in medicine. Therefore, synthetic images are useful to improve performance in tasks like classification and segmentation of patholoigies.

In this work an study about the using 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 thought 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 in two main stages: The first one is related with the training of the Inception model only with the original and unbalance data retrieved from ISIC dataset. The second stage is about the training of the Inception model using the dataset augmented with synthetic images generated using a GANs. The core objective in this work is the generation of images of skin lesions in the most possible realistic way.

2 Related Works

2.1 Image Classification on Skin Lessions

The image classification is a Computer Vision (CV) task which consists in assign an specific label or category to an image according to the information retrieved from an input image. In the medicine field, this can be important to recognize and distinguish different pathologies.

Li *et al.* [16] proposed a FCRN-50 and FCRN-101 architectures for classification of melanoma, seborrheic keratosis and nevus images. Also, they constructed a Lesion Indexing Network (LIN) for skin lesion image analysis. Similarly, Yilmaz *et al.* [17] took data from ISIC 2017 challenge and apply 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 get an accuracy of 82%.

The authors in [18] presented an experiment with different neural networks for the classification of benign and malignant skin lessions. They implement a PNASNet-5-Large, InceptionResNetV2, SENet154, and InceptionV4. Data preprocesing and data augmentation was performed. The pre-processing consist in a normalization to convert the pixel values into 0 and 1. For data augmentation, a set of transformations were applied in order to reduce possible loss of performance 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 gotten by PNASNet-5-Large model which has 0.76 in validation score. Futhermore, Devries *et al.* [19] used a multi-scale convolutional network to classify data retrieved from ISIC 2017 skin lesion classification challenge. In order to get better results, they performed a fine-tuning using a pretrained Inception-v3 model trained in ImageNet dataset. Also, it was important to add images from ISIC_MSK2_1 dataset to improve the results of melanoma prediction. As a result of the training, they obtained a accuracy of 0.893 for melanoma, 0.913 for seborrheic keratosis, and an average 0.903 for the general classification task.

2.2 ISIC Data Augmentation with GANs

In recent years, the Deep Learning techniques have proven to be quite useful in offering efficient alternatives for the task of data augmentation. Particularly, the use of Generative Adversarial Networks (GANs) helps with the generation of synthetic images which are similar to the real data. Pollastri *et al.* [1], proposed the use of 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. At the end, they used a Convolutional-Deconvolutional Neural Network (CDNN) to measure the improvement of the accuracy by adding the synthetic data into the training process. With the DCGAN was beter for one experiment, while the LAPGAN had the best performance in four experiments. Both achieve an improvement near of 1%. In addition, the use of a U-Net with original and synthetic data showed better results (a gain about 1% of accuracy with the DCGAN).

Additionally, Baur *et al.* [2] proposed generate realistically and high resolution images of skin lesions with GANs. For this, 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 also, adding the synthetic images generated by GANs. Finally, they achieve near of 1% more accuracy score with the images from LAPGAN both in training (99.29%) and validation (74.00%) sets.

Another work was presented by Rashid *et al.* [3]. They proposed the use of GANs to solve the problem of the limited amount of data in medical imaging. For this task, they use a Vanilla GAN to generate realistic images using the 2018 ISIC dataset. Also, an interesting feature of this GAN is that it can be also used for classification. Therefore, for the classification task, they used a fine-tuned DenseNet and ResNet architectures together with the Vanilla GAN as baseline models. The metric used to make the performance comparison of the experiments is the balance accuracy score due to the unbalance data presented. Consequently, they achieve a balance accuracy score of 0.815 ith DenseNet, 0.792 with the ResNet , and 0.861 with GANs.

There are also more modern applications of GANs that offer greater opportunities for generating images with a great deal of detail and style. Limeros *et al.* [4] implemented a StyleGAN2-ADA using an original implementation from NVIDIA Research group. This was performed using ISIC 2020 and ISIC 2019 datasets. The classification task was performed using EfficientNet-B2 model pre-trained on ImageNet. They balance the malignant class adding 22.000 synthetic images. After this, they performed different classification scenarios to compare the results with baseline data and augmented data. In all cases, they tested on the same real images validation set. As a result, authors achieve a 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) [23]. It was designed to release digital skin images to produce tools that helps to reduce skin cancer mortality. They provide a large image gallery and filters to search through the dataset letting us to browse inside a lot of public images. For the purpose of this work, we only focused on the first filter of the gallery that is diagnostic attributes. With this filter, we retrieved 8833 benign and 7361 malignant images. Each image has a size of 1769x1769. Some examples of this dataset is given by Figure 1.

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 with the addition of convolution layers to replace the multilayer perceptron. The convolution layer part is used to discriminate between the images received by the discriminative network, meanwhile the

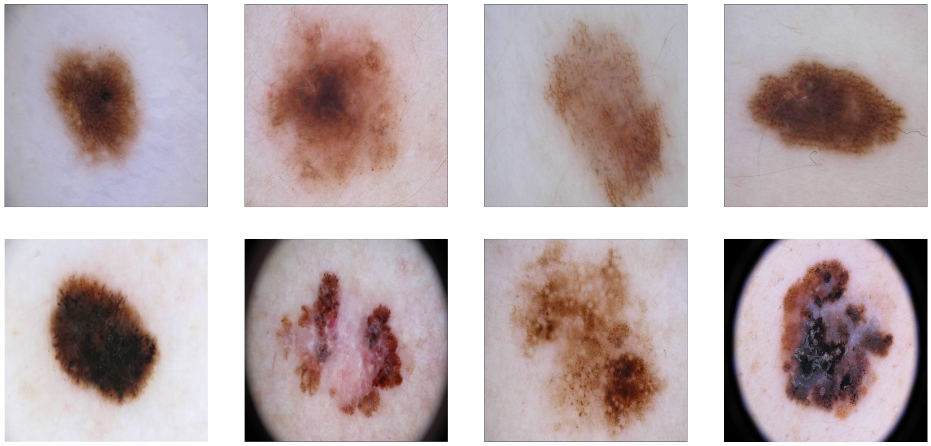

Fig. 1: Examples of ISIC dataset. First row: bening images. Second row: malignant images.

deconvolution layer part is used to generate the images in the generative network [7]. DCGANs use various techniques to improve the training phase. These techniques are the following: 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; the using of 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 is a derivation of the original Generative Adversarial Network (GAN) proposed by [?], 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 into the network. Next, transposed convolution layers (unsampling layers) are used to increase the dimension of the input tensor up to the desired dimension. This increase occurs for the height and width, whereas there is a reduction in channels dimension. These transposed convolution layers help to decompress the noise to be transformed 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 in order to stabilize the training process. A Rectified Linear Unit (ReLu) activation function is applied to introduce nonlinearity into the network. This

process of applying 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 with the purpose 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) [7]. The general architecture of this network is shown in the Figure 2.

2. DCGAN Discriminator: The discriminative network takes as input either a synthetic image generated by the generator or a real image extracted from the training dataset. 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 patterns of the image. After each convolution layer, a Batch Normalization layer is applied join to a LeakyReLy activation function. Then, a final probability is generated as output by the Sigmoid activation function. 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) [7].

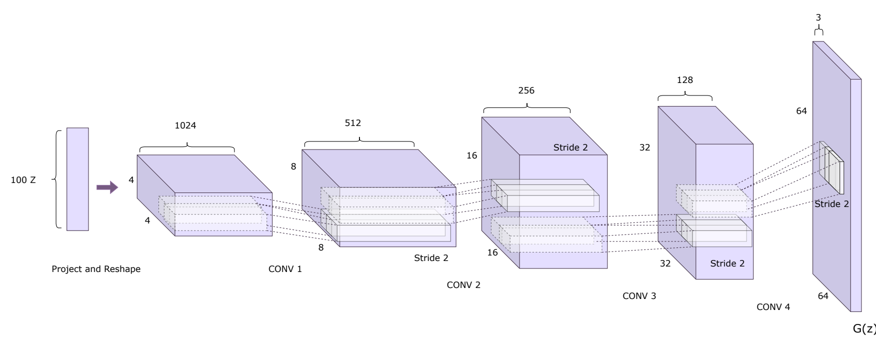

Fig. 2: DCGAN Generator Architecture. A 100 dimensional uniform distribution Z is projected to a small spatial convolutional representation with some feature maps. This is follow by four fractionally-strided convolutions. Finally, this converts to a representation of 64x64 pixel image [7].

3.3 Evaluation Metrics

In this section are described metrics to evaluate both the quality of the synthetic generated images and the performance of the classification tasks. 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 for the evaluation of generative models, including the FID. It is used to calculate the distance between feature vectors calculated for real and synthetic generated images [9] i.e., 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 very 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 [8]. After generating a set of synthesized images, each of them is sent to the InceptionV3 network. This evaluates the probability of each generated image to belong to the set of real images. Additionally, the entropy of the probability distributions is calculated to measure the diversity of the generated images [22]. 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 important to first know how the classification task performs without any increase in data and then to have a point of comparison respect to the training performance with the augmented dataset with synthetic images. For the classification task, we get an InceptionV3 model which was trained during 20 epochs, with a batch size of 128, learning rate of 0.001, Adagrad as optimizer, and Cross-Entropy as loss function. Additionally, data augmentation techniques like random horizontal and vertical flip, and random rotation. All the input data was normalization before enter to the network. With this parameters we obtained the results shown in Table 1 which are computed from the values of the confusion matrix 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 with the balancing of 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 the Figure 4.

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

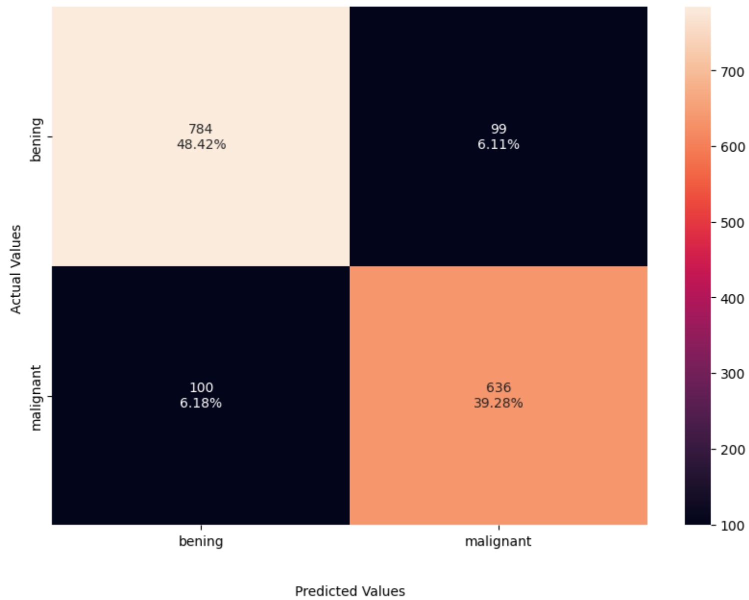

Fig. 3: Confusion matrix for the ISIC dataset classification.

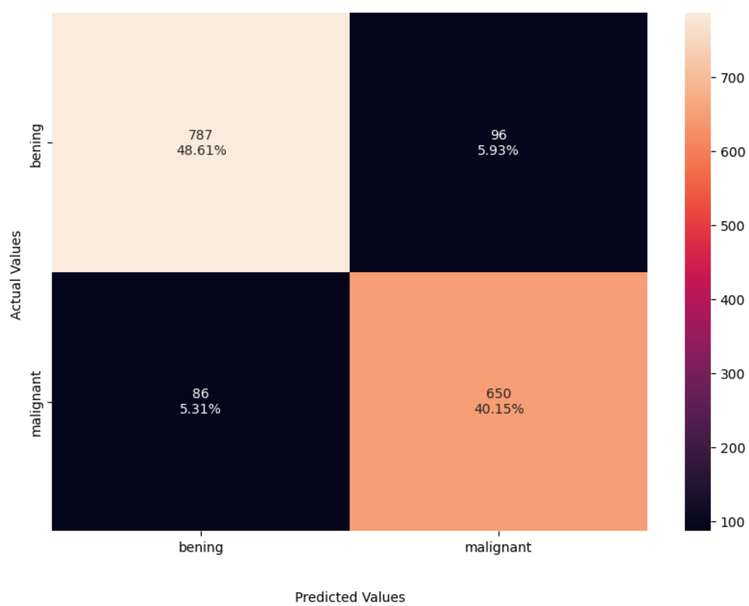

Fig. 4: Confusion matrix for the ISIC dataset classification augmented with DCGAN synthetic images.

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 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 training during 300 epochs, with an image size and a batch size of 64. The evolution of the loss fuction both for the discriminator and generator networks are shown in Figure 5. In one hand, the value of the loss function began to increase until 2500 iterations approximately. After that, this value begin 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 by means of FID and IS is presented in Figure 6. The FID value until the 50th epoch is high. After this epoch, this value begin to decrease. On the contrary, the IS value before the 50th epoch begin small. After this epoch, this value begin to increase. The evolution of both values indicate us that the synthetic images produced by the generator are good enough as we can see in Figure 7 were is shown a comparison between real and synthetics images generated with the DCGAN.

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 in order to balance the dataset. In particular, for the skin lesion classification, if we use the original unbalanced data from the ISIC dataset, 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, there are other metrics which show a better performance of the 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

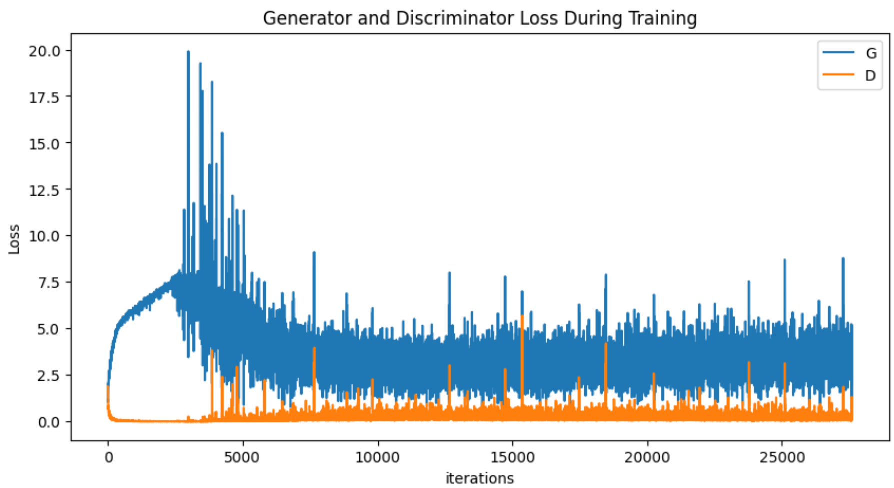

Fig. 5: Evolution of the loss function for the generator and discriminator during training phase of the DCGAN for the ISIC dataset into malignant class.

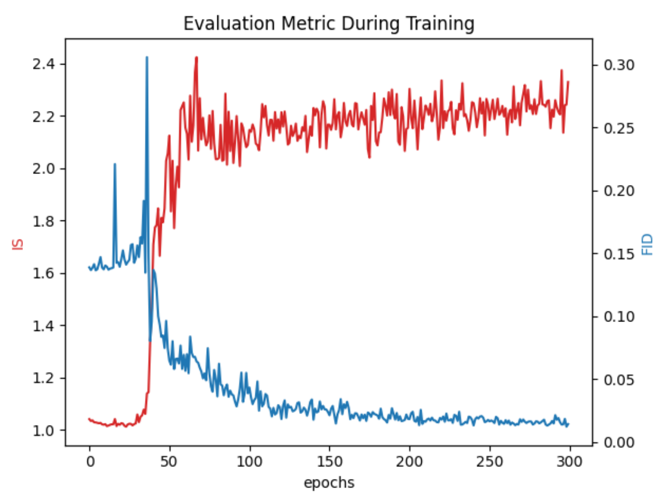

Fig. 6: Evolution of FID and IS metrics during the DCGAN training phase.

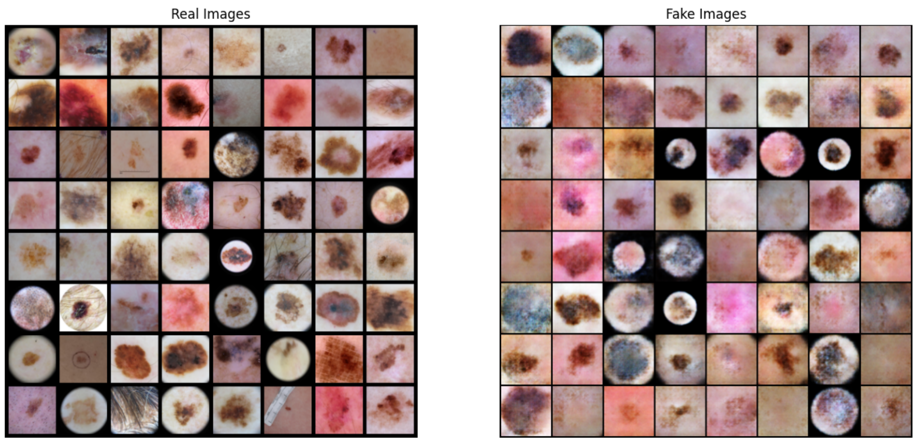

Fig. 7: Comparison between real images and synthetic images generated by the DCGAN. Left batch: real images from the ISIC dataset in the malignant class. Right batch: synthetic images for the malignant class.

improvements, however small they may seem, help to make correct diagnoses, reduce false positives and negatives, and improve disease management with early diagnosis.

In additit-ion, into the confussion matrices can be observed the reduction of false positives and the increase for true positives values. Similarly, can be observed the reduction of false negatives and the increase for true negatives values. For the malignant class, the improvement in value classification is more significant. Finally, with respect to the quantification of image quality by means of FID and IS, favorable results were obtained. 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 classify 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. The use of data augmentation by means of GANs has been shown to increase the accuracy rate for the skin lesion classification task. This is demonstrated by the comparison made between the results gotten in the training of skin lesion classification model 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 implementing the generation of skin lession synthetic images with other existing GAN models. A comparison between the presented model and other existing models for the generation of 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 with respect to the classification of skin lesions.

References

1. Pollastri, Federico, et al. "Augmenting data with GANs to segment melanoma skin lesions." Multimedia Tools and Applications 79 (2020): 15575-15592.
2. Baur, Christoph, Shadi Albarqouni, and Nassir Navab. "MelanoGANs: high resolution skin lesion synthesis with GANs." arXiv preprint arXiv:1804.04338 (2018).
3. 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). IEEE, 2019.
4. Limeros, Sandra Carrasco, et al. "GAN-based generative modelling for dermatological applications–comparative study." arXiv preprint arXiv:2208.11702 (2022).
5. 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 (2018): 2572-2581.
6. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
7. Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).
8. Barratt, Shane, and Rishi Sharma. "A note on the inception score." arXiv preprint arXiv:1801.01973 (2018).
9. 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).
10. Saginala, Kalyan, et al. "Epidemiology of melanoma." Medical sciences 9.4 (2021): 63.
11. Adegun, Adekanmi A., and Serestina Viriri. "Deep learning-based system for automatic melanoma detection." IEEE Access 8 (2019): 7160-7172.
12. 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 (2022): e330-e339.
13. Wen, David, et al. "Characteristics of publicly available skin cancer image datasets: a systematic review." The Lancet Digital Health 4.1 (2022): e64-e74.

14. Cancer, https://www.who.int/news-room/fact-sheets/detail/cancer,2022, accessed on:21-JUN-2023
15. American Cancer Society. Key statistics for melanoma skin cancer, 2023. Accessed: 21-JUN-2023.
16. Li, Yuexiang, and Linlin Shen. "Skin lesion analysis towards melanoma detection using deep learning network." Sensors 18.2 (2018): 556.
17. Yilmaz, Abdurrahim, et al. "Benchmarking of Lightweight Deep Learning Architectures for Skin Cancer Classification using ISIC 2017 Dataset." arXiv preprint arXiv:2110.12270 (2021).
18. 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).
19. DeVries, Terrance, and Dhanesh Ramachandram. "Skin lesion classification using deep multi-scale convolutional neural networks." arXiv preprint arXiv:1703.01402 (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. Song, Yang, and Stefano Ermon. "Improved techniques for training score-based generative models." Advances in neural information processing systems 33 (2020): 12438-12448.
23. 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].
24. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). `https://doi.org/10.10007/1234567890`
25. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
26. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
27. LNCS Homepage, `http://www.springer.com/lncs`. Last accessed 4 Oct 2017