Comparative Evaluation of StyleGAN3-Based Augmentation Strategies for Enhanced Medical Image Classification

Faycal Touazi[1], Djamel Gaceb[1], Amira Tadrist[1] and Sara Bakiri[1]

[1] *LIMOSE Laboratory, Computer Science Department, University M'hamed Bougara, Independence Avenue, 35000 Boumerdes, Algeria*

Abstract

Deep learning models for medical image classification face significant challenges due to class imbalance and the limited availability of annotated datasets, particularly for rare diseases. Traditional data augmentation techniques, such as rotation, translation, etc., often fail to provide sufficient diversity to perform a good classification for minor classes. To address this issue, various strategies have been explored, including oversampling, undersampling, cost-sensitive learning, and synthetic data generation using generative adversarial networks (GANs). In this study, we evaluate the impact of using a generative AI based approaches and demonstrate that the most effective strategy is to combine synthetic augmentation with traditional methods. Specifically, we employ StyleGAN3 to generate high-fidelity synthetic images that, when integrated with traditional data-augmentation techniques, may improve the performance of deep learning models on medical image classification. We validate our method on datasets, including COVID-19 chest X-rays and HAM10000. Experimental results show that this hybrid approach leads to an improvement in classification accuracy, particularly for minority classes, surpassing standalone augmentation strategies. Our findings highlight the potential of AI-driven synthetic data generation as a complementary solution to traditional augmentation, offering a more balanced and diverse dataset for medical image analysis.

Keywords

Medical imaging, Data augmentation, Generative Adversarial Networks, StyleGAN, Class imbalance

1. Introduction

In the field of deep learning for medical imaging, one of the significant challenges is class imbalance coupled with small annotated data. It is especially challenging when working with rare diseases, where the low occurrence and brief duration of the appearance of symptoms can make the collect of data difficult, which may affect the quality of the model training.

As a result, such a deficiency affects classification model performance, particularly in classifying complicated pathologies that are important even their rarity in the datasets. This asymmetry degrades the performance of the model in favor of the majority classes, thus decreasing the precision and reliability of the predictions compared to the minority classes.

The imbalance datasets, coupled with the limited availability of annotated datasets, presents obstacles to the development of efficient and high-quality models in the field of medical imaging. Traditional classification models become overfitted to the dominant classes, leading to a significant loss of accuracy for the minority classes. While traditional data augmentation techniques such as rotation, resizing, and cropping—are often applied to alleviate this issue, they often fail to generate the necessary diversity and do not notably enhance the generalization capacity of the models.

Classical data augmentation techniques [1], such as rotation, flipping, scaling, and cropping, are widely used to artificially increase the size of training datasets and improve model generalization. These methods help in introducing minor variations to the images, making the model more robust to small transformations. However, they have significant limitations, especially in the medical imaging domain. Since medical images often contain complex and subtle patterns that are crucial for diagnosis, simple transformations may not sufficiently capture the variability needed to enhance model performance. Furthermore, these techniques do not create new pathological patterns but merely

[1]CMIS-2025: Eighth International Workshop on Computer Modeling and Intelligent Systems, May 5, 2025, Zaporizhzhia, Ukraine

f.touazi@univ-boumerdes.dz (F. Touazi); d.gaceb@univ-boumerdes.dz (D. Gaceb); a.tadrist@univ-boumerdes.dz (A. Tadrist); s.bakiri@univ-boumerdes.dz (S. Bakiri)

0000-0001-5949-5421 (F. Touazi); 0000-0002-6178-0608 (D. Gaceb)

modify existing ones, limiting their effectiveness in addressing class imbalance. As a result, they may not significantly improve the classification of rare diseases, which require more sophisticated augmentation strategies capable of generating realistic and diverse samples.

To address class imbalance in deep learning, various strategies can be employed (see [2] for an exhaustive review).

1. Oversampling Methods: Oversampling techniques aim to increase the representation of minority- class samples to balance the dataset. Synthetic Minority Over-sampling Technique (SMOTE) and its variants generate synthetic data points to improve class distribution [3, 4]. Additionally, data augmentation techniques introduce transformed versions of existing images to improve model generalization. While oversampling has been shown to be one of the most effective techniques for CNN-based classification [5], its effectiveness can be limited on images in general and for medical imaging.
2. Undersampling Methods: In contrast, undersampling reduces the number of majority-class samples to achieve a more balanced dataset, thereby improving class proportions and reducing computational cost. Random undersampling removes a subset of majority samples, while more advanced techniques, such as cluster-based undersampling, aim to retain the most informative samples. Although undersampling is effective in extreme imbalance scenarios, it may lead to information loss, particularly in complex medical datasets [5].
3. Other Learning Approaches: Beyond sampling strategies, cost-sensitive learning modifies the loss function to assign higher penalties for misclassifications in the minority class, with focal loss being a notable example that prioritizes hard-to-classify instances [6, 7]. Ensemble learning improves prediction accuracy by combining multiple classifiers, but its high computational cost can be prohibitive [8]. Hybrid approaches integrate data-level and algorithm-level solutions, such as clustering with sampling techniques or cost-sensitive learning with neural networks. Semi-supervised and self-supervised learning leverage unlabeled data to enhance feature representation and generalization, while deep metric learning and contrastive learning focus on learning more discriminative representations without altering class distribution. Each of these methods has trade-offs in efficiency, robustness, and complexity, so a choice can be made based on the nature of the dataset and the nature of application demands.

To overcome the limitations of traditional imbalance-handling techniques, AI-based image-generation methods [9, 10, 11, 12, 13], and particularly StyleGAN [14, 15, 16], have emerged as a groundbreaking solution for generating realistic synthetic medical images. StyleGAN's ability to produce high-fidelity images enables dataset augmentation without compromising the valuable pathological characteristics essential for medical imaging. By generating samples for underrepresented classes, StyleGAN helps mitigate class imbalances and improves the stability of classification models.

In this work, we introduce a StyleGAN3-based data augmentation approach that combines state-of-the- art generative modeling with classical balancing techniques to enhance the diversity and representation of minority-class samples. StyleGAN3, with its improved spatial coherence, is particularly well-suited for medical image synthesis, preserving intricate morphological features of pathological conditions. Unlike conventional oversampling methods that risk overfitting, our approach generates diverse and realistic synthetic samples, enriching the dataset and improving the generalization of classification models. By training StyleGAN3 on the minority class, we aim to restore class balance, enhance dataset variability, and ultimately improve the robustness of medical image classification systems.

This paper is structured as follows: Section 2 is a literature review of data augmentation and GANs in medical imaging. Section 3 is a description of the methodology of our work, including data preprocessing, model architecture, and performance metrics. Section 4 is the experimental results and their discussion. Section 5 concludes the paper and gives future directions of research

2. Related Works

2.1. Based on the Covid_19 Radiography Dataset

Abdul Waheed et al. [17] introduced CovidGAN, an ACGAN-based model generating synthetic chest X-ray (CXR) images to address data scarcity in medical imaging. Trained on three datasets (IEEE

Covid Chest X-ray, COVID-19 Radiography Database, COVID-19 Chest X-ray Dataset), CovidGAN improved CNN classification accuracy from 85% to 95% with augmented data.

Sharmila V J et al. [18] proposed a DCGAN-CNN hybrid for classifying CXR images (normal, pneumonia, COVID-19). The DCGAN generates 64×64 synthetic images, later resized for classification. The CNN, comprising eight convolutional layers, achieved accuracy between 94.8% and 98.6%, surpassing AlexNet and GoogLeNet.

2.2. Based on the HAM10000 Dataset

Bilal Ahmad et al. [19] developed TED-GAN, a hybrid VAE-GAN approach for skin lesion image generation. Using a dual-GAN framework, their model significantly improved melanoma classification, increasing sensitivity from 53% to 82% and specificity from 75% to 94%.

Qinchen Su et al. [20] introduced STGAN, a GAN-based augmentation method for multi-class imbalanced skin lesion classification. Trained on the HAM10000 dataset, it improved FID, Inception Score, Precision, and Recall over StyleGAN2 and achieved an accuracy of 98.23% with a ResNet50 classifier.

2.3. Based on Other Datasets

Bilal Ahmad et al. [21] proposed VAE-GAN, leveraging informative noise instead of Gaussian noise for brain tumor image generation. Applied to 3,064 CE-MR images, their approach boosted classification accuracy from 72.63% to 96.25%.

Guilherme C et al. [22] implemented StyleGAN2-ADA, enhancing image quality for fundus imaging via adaptive discriminator augmentation to mitigate data scarcity and imbalance.

Table 1
Summary of Related Works

Study	Methodology	Dataset	Accuracy	Year
Waheed et al. [17]	CovidGAN (ACGAN)	COVID-19 Radiography	95%	2020
Sharmila et al. [18]	DCGAN-CNN	COVID-19 Radiography	94.8%-98.6%	2021
Ahmad et al. [19]	TED-GAN (VAE-GAN)	HAM10000	Sensitivity: 82%	2022
Su et al. [20]	STGAN	HAM10000	98.23%	2021
Ahmad et al. [21]	VAE-GAN (Brain Tumors)	CE-MR Brain Tumor	96.25%	2023
Guilherme et al. [22]	StyleGAN2-ADA	Fundus Imaging	85%	2021

3. Proposed Approach

Our approach stands out by leveraging GANs not only for data augmentation but also for dataset balancing. To ensure fair evaluation and prevent data leakage, our dataset was initially divided into 80% for training and 20% for testing. This split remains consistent across all experiments, and transformations are applied only to the test set. We conducted our study on two medical imaging datasets: HAM10000 [23] and COVID-19 Radiography [24]. We propose four augmentation strategies:

1. Approach 1 - Traditional Data Augmentation Classical transformations (rotation, flipping, resizing) are applied to enhance training data diversity.
2. Approach 2 - Targeted Augmentation for Balancing Augmentation is applied specifically to minority classes to balance the dataset.
3. Approach 3 - StyleGAN3-Based Augmentation with Batch Injection Synthetic images generated by StyleGAN3 are injected into training batches to improve diversity.
4. Approach 4 - Hybrid StyleGAN3 Augmentation and Traditional Balancing A percentage of StyleGAN3-generated images is added, followed by traditional balancing techniques.

3.1. Approach 1: Traditional Data Augmentation

We apply standard transformations such as rotation, horizontal/vertical flipping, and color jittering before feeding images into ResNet50 and InceptionV3. Table 2 summarizes the transformations.

Table 2
Transformations applied for traditional augmentation

Transformation	Value
RandomHorizontalFlip	0.5
RandomVerticalFlip	0.5
RandomRotation	30°
ColorJitter	brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1 RandomAffine
RandomHorizontalFlip	0.5
RandomAffine	translate=(0.1, 0.1)

3.2. Approach 2: Targeted Augmentation for Balancing

This approach is inspired by Random Over-Sampling (ROS), a common technique for handling imbalanced datasets by duplicating samples from minority classes to match the distribution of majority classes. However, instead of simply duplicating existing images, we apply targeted data augmentation techniques (e.g., rotation, scaling, contrast adjustments) to generate new synthetic samples. The gener- ated images are saved and used to balance the dataset, ensuring that minority classes have the same number of images as the majority classes.

3.3. Approach 3: StyleGAN3-Based Augmentation with Batch Injection

StyleGAN3 is used to generate high-quality synthetic medical images that are directly injected into training batches during model training. Unlike traditional augmentation, which applies transformations to existing images, StyleGAN3 synthesizes new samples that mimic the distribution of real medical images.

In this approach, synthetic images are generated before training and dynamically included in mini-batches alongside real images. This ensures that the model learns robust representations by exposing it to a more diverse dataset. The dataset split remains unchanged, with synthetic images used only during training, preventing bias in the test evaluation.

3.4. Approach 4: Hybrid StyleGAN3 Augmentation and Traditional Balancing

This approach combines StyleGAN3-generated images with traditional dataset balancing techniques to optimize model performance. The augmentation process consists of two steps:

1. Generation of Synthetic Images: StyleGAN3 is used to generate additional images. We test four strategies by adding synthetic samples to the original training dataset, increasing the minority classes by 10%, 20%, 30%, and 40%, respectively.
2. Traditional Balancing Techniques: Once the synthetic images are added, classical balancing methods are applied. This includes oversampling the minority class and targeted augmentations (rotation, flipping, and intensity scaling) to equalize class representation.

This hybrid strategy ensures that the dataset remains well-balanced while introducing new variations through GAN-generated samples. The classifier is trained on the augmented dataset using ResNet50 and InceptionV3, and performance is evaluated based on classification metrics.

4. Results of data-augmentation

4.1. Results of StyleGAN3

The generated images show appreciable diversity in the characteristics of each class. This diversity is crucial to avoid overfitting and to improve the generalization of classification models by including realistic variations in the data.

Below, you will find samples of the images generated by StyleGAN3 for each minority class.

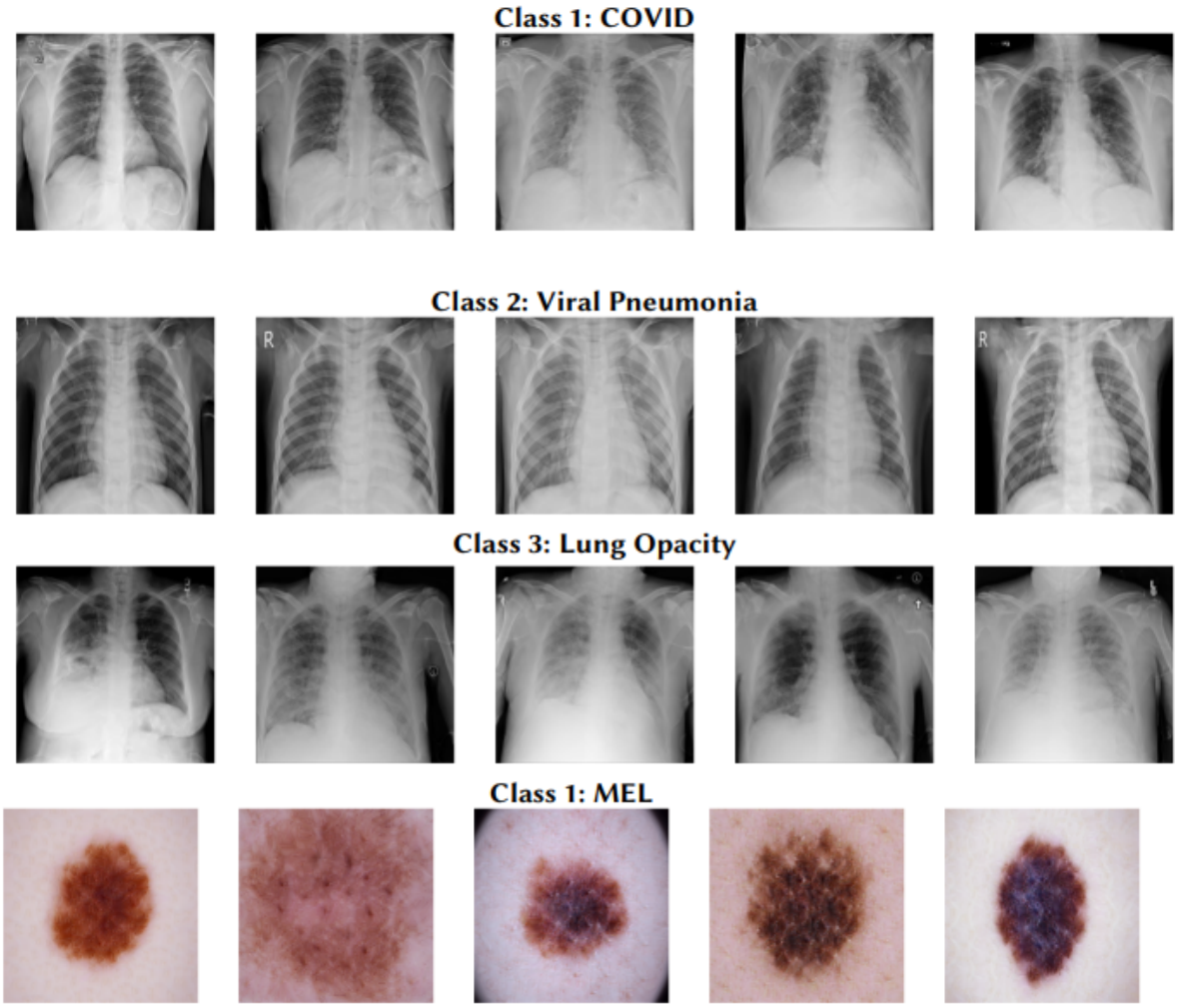


Figure 1: Images generated by StyleGAN3 for the COVID-19 related classes of the Covid-19 dataset and for HAM10000 MEL class.

4.2. Results of Augmentation

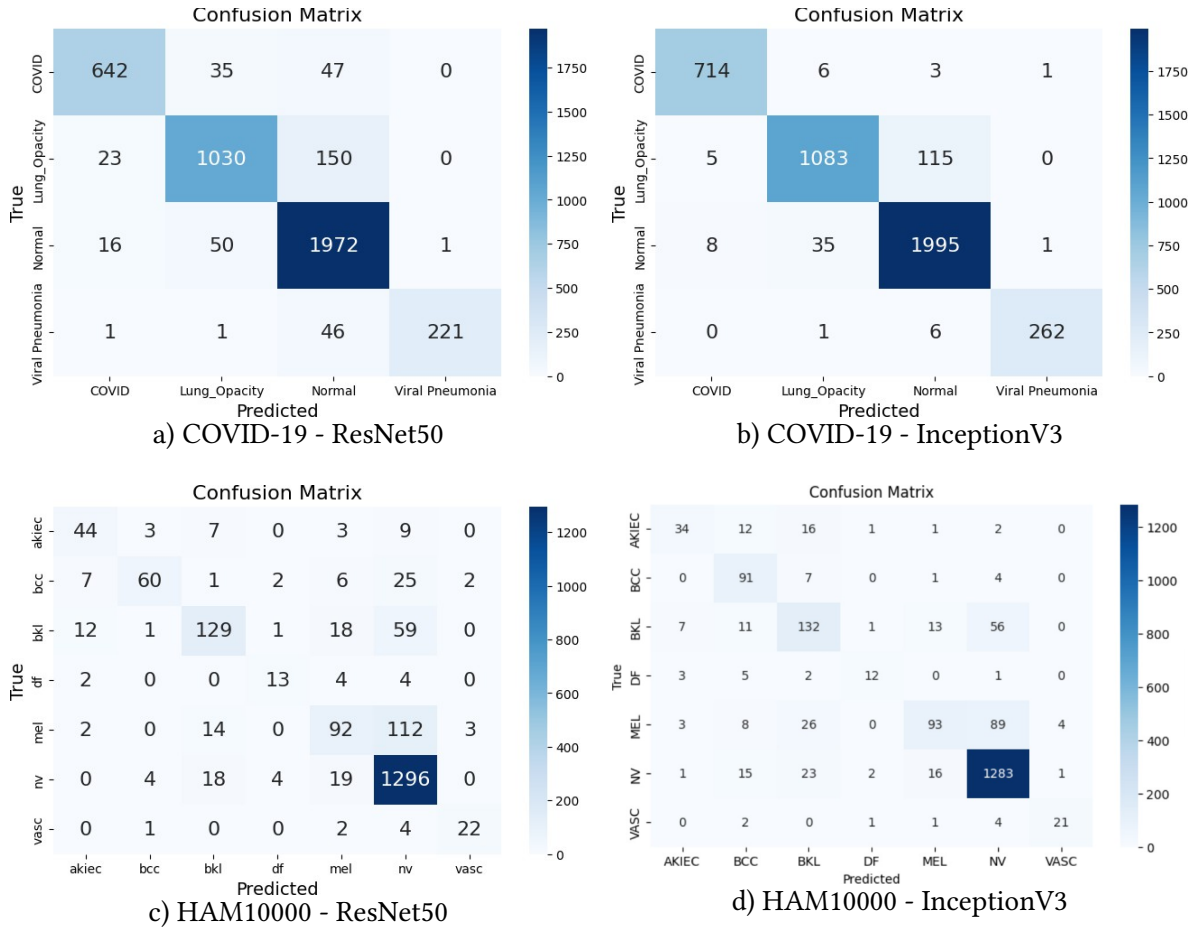
4.2.1. Approach 1: Traditional Augmentation

Table 3 presents the performance of the ResNet50 and InceptionV3 models on the COVID-19 Radiography and HAM10000 datasets using Approach 1, evaluated in terms of accuracy, recall, precision, and F1-score.

It is observed that the InceptionV3 model slightly outperforms the ResNet50 model in terms of accuracy and F1-score, achieving an accuracy of 94.82% compared to 91.26% for ResNet50 on the COVID-19 dataset and 82.49% compared to 81.48% on the HAM10000 dataset

Table 3
Performance results of the models with Approach 1: Traditional augmentation

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	91.26%	91.20%	91.26%	91.50%	82.59%	81.27%	82.59%	81.48%
InceptionV3	94.82%	94.84%	94.83%	94.88%	83.09%	82.02%	83.09%	82.49%



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 2: Confusion matrices for ResNet50 and InceptionV3 on the COVID-19 and HAM10000 datasets with Approach 1.

The confusion matrices in (see Figure 2) visualize the performance of the ResNet50 and InceptionV3 models on the COVID-19 Radiography and HAM10000 datasets using the traditional augmentation approach. They allow for evaluating the quality of each model's predictions on these two datasets in terms of correct and incorrect classifications.

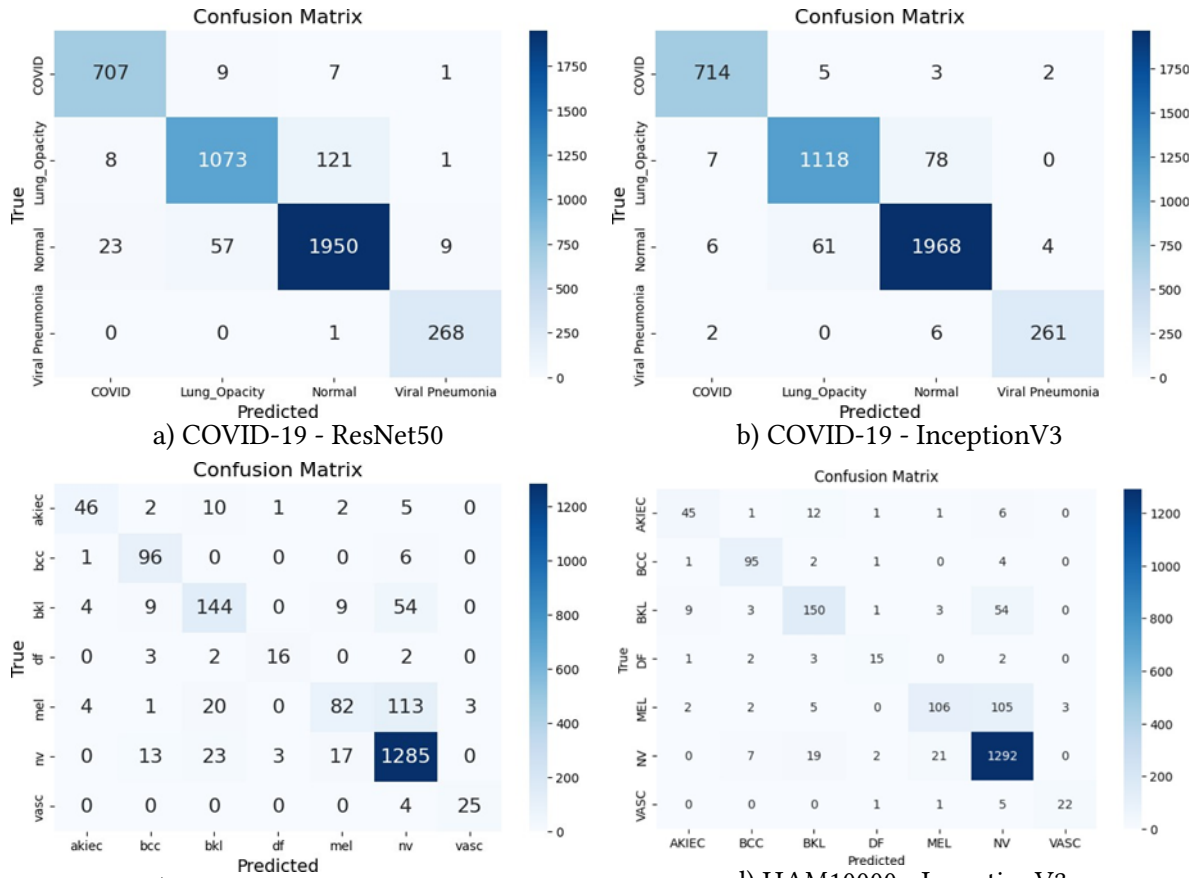
4.2.2. Approach 2: Traditional Data Augmentation with Balancing

Table 4 presents the performance of the ResNet50 and InceptionV3 models on the COVID-19 Radiography and HAM10000 datasets using Approach 2, evaluated in terms of accuracy, recall, precision, and F1-score.

It is noteworthy that the InceptionV3 model slightly surpasses the ResNet50 model in terms of accuracy and F1-score on the COVID-19 Radiography dataset, achieving an accuracy of 95.46% compared to 94.33% for ResNet50. Conversely, for the HAM10000 dataset, it is the ResNet50 model that displays better performance, achieving an accuracy of 83.19% compared to 71.77% for InceptionV3

Table 4
Performance results of the models with Approach 2: Traditional data augmentation with balancing

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	94.40%	94.37%	94.40%	94.40%	84.49%	84.49%	83.68%	83.16%
InceptionV3	95.89%	95.88%	95.89%	95.89%	86.03%	86.03%	85.11%	82.02%



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 3: Confusion matrices for ResNet50 and InceptionV3 on the COVID.

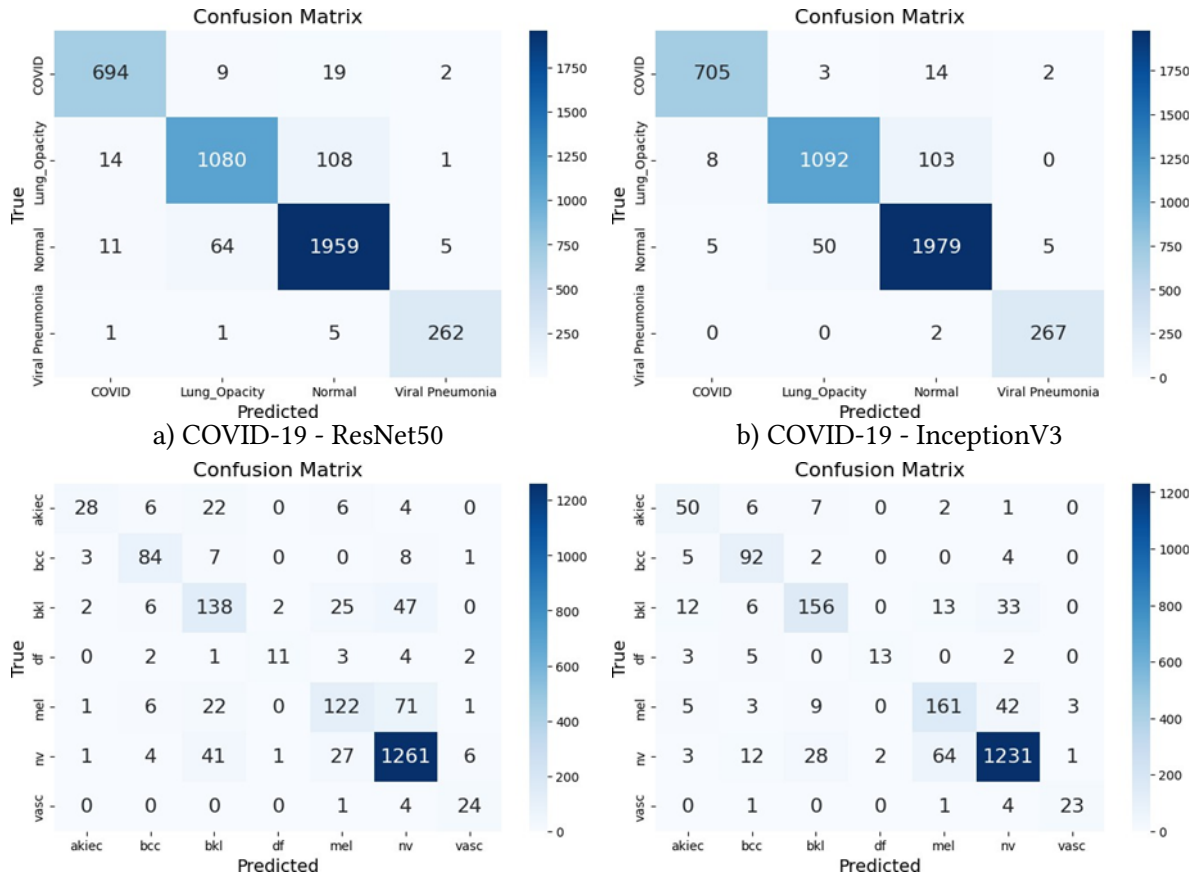
4.2.3. Approach 3: Augmentation using StyleGAN3 with Batch Injection

Table 5 presents the performance of the ResNet50 and InceptionV3 models on the COVID-19 Radiography and HAM10000 datasets using Approach 3, evaluated in terms of accuracy, recall, precision, and F1-score.

Table 5
Performance results of the models with Approach 2: Traditional data augmentation with balancing

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	94.33%	94.31%	94.33%	94.33%	83.19%	82.64%	83.19%	82.77%
InceptionV3	95.46%	95.45%	95.47%	95.48%	86.08%	86.19%	86.08%	86.54%

We observe that the InceptionV3 model slightly outperforms the ResNet50 model in terms of accuracy and F1-score on the COVID-19 Radiography dataset, achieving an accuracy of 95.46% compared to 94.33% for ResNet50. However, on the HAM10000 dataset, the InceptionV3 model also achieves better results with an accuracy of 86.08% compared to 83.19% for ResNet50 (see Figure 4 for the confusion matrices).



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 4: Confusion matrices for ResNet50 and InceptionV3 on the COVID-19 and HAM10000 datasets using Approach 3.

4.2.4. Approach 4: Augmentation using StyleGAN3 and Balancing with Traditional Methods

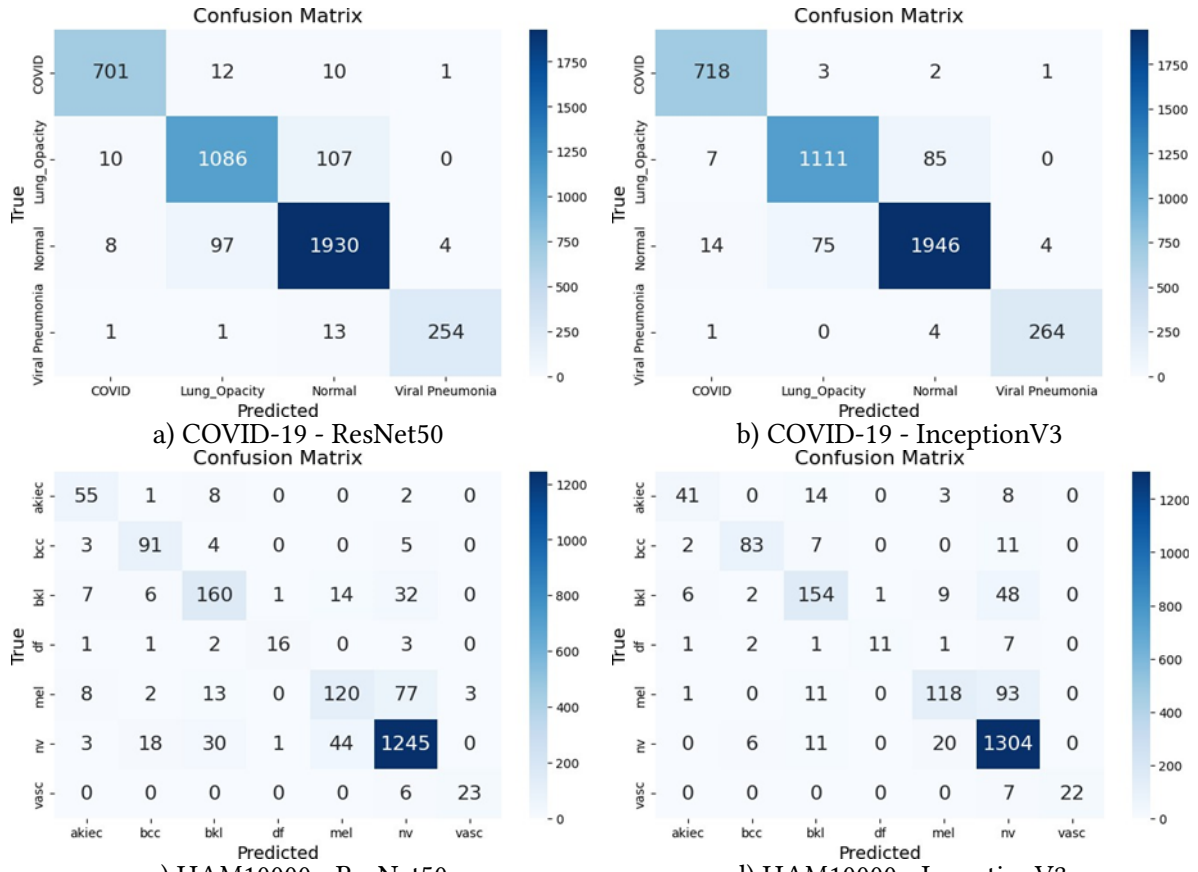
In this approach, we tested different levels of data augmentation, namely 10%, 20%, 30%, and 40%.

4.2.5. With 10% Augmentation

Table 6 presents the performance of the ResNet50 and InceptionV3 models on the COVID-19 Radiography and HAM10000 datasets with 10% data augmentation using Approach 4, evaluated in terms of accuracy, recall, precision, and F1-score.

Table 6
Performance results of the models with Approach 1: Traditional augmentation

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	93.76%	93.77%	93.77%	93.78%	85.28%	84.97%	85.29%	84.95%
InceptionV3	95.37%	95.36%	95.37%	95.36%	86.43%	85.61%	86.43%	85.96%



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 5: Confusion matrices for ResNet50 and InceptionV3 on the COVID-19 and HAM10000 datasets using Approach 4 with 10% augmentation.

Table 7

Model performance results with approach 4: augmentation by StyleGAN3 and balancing with traditional methods

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	94.49%	94.48%	94.50%	94.49%	83.99%	82.51%	83.99%	83.26%
InceptionV3	95.96%	95.96%	95.96%	95.96%	86.43%	86.20%	86.43%	86.14%

It is observed that the InceptionV3 model continues to outperform the ResNet50 model in terms of accuracy and F1-score with this approach on the COVID-19 Radiography dataset, achieving an accuracy of 95.84% compared to 94.49% for ResNet50. Furthermore, on the HAM10000 dataset, InceptionV3 also demonstrates better performance with an accuracy of 86.43% versus 83.99% for ResNet50.

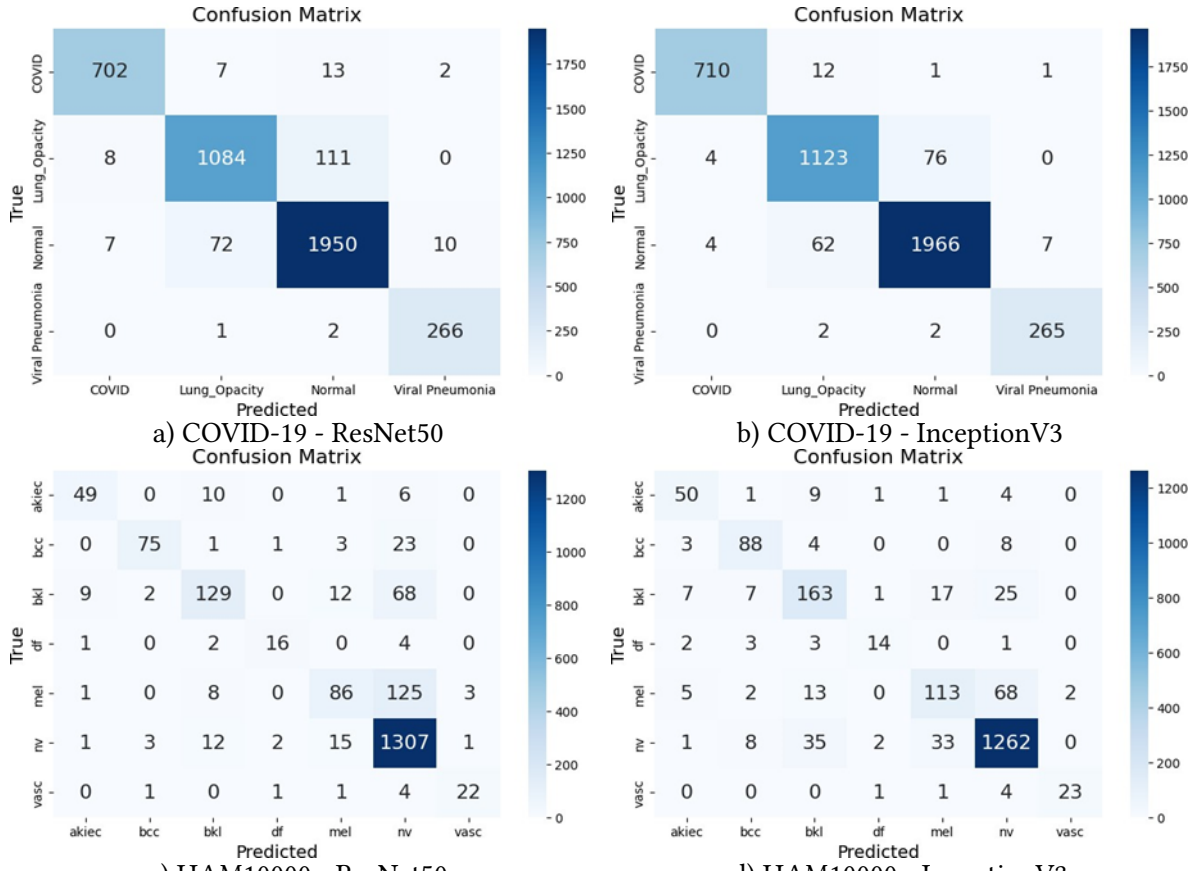
4.2.6. With 20% Augmentation

Table 8 presents the performance of the InceptionV3 and ResNet50 models on the COVID-19 Radiography and HAM10000 datasets with a 20% data augmentation using approach 4.

Table 8

Performance results of the models with approach 4: augmentation using StyleGAN3 and balancing with traditional methods

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	93.76%	93.75%	93.77%	93.75%	84.88%	84.31%	84.89%	84.22%
InceptionV3	95.58%	95.60%	95.58%	95.62%	85.73%	85.33%	85.74%	85.65%



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 6: Confusion matrices for ResNet50 and InceptionV3 on the COVID-19 and HAM10000 datasets with approach 4 at 30%.

It can be observed that in the COVID-19 Radiography dataset, InceptionV3 outperforms ResNet50 with an accuracy of 95.58% compared to 93.76%. The F1 scores, recall, and precision further confirm this trend, indicating better overall performance for InceptionV3.

Regarding the HAM10000 dataset, although the gap is smaller, InceptionV3 maintains an advantage with an accuracy of 85.73% compared to 84.88% for ResNet50. This difference demonstrates better handling of complex classes by InceptionV3.

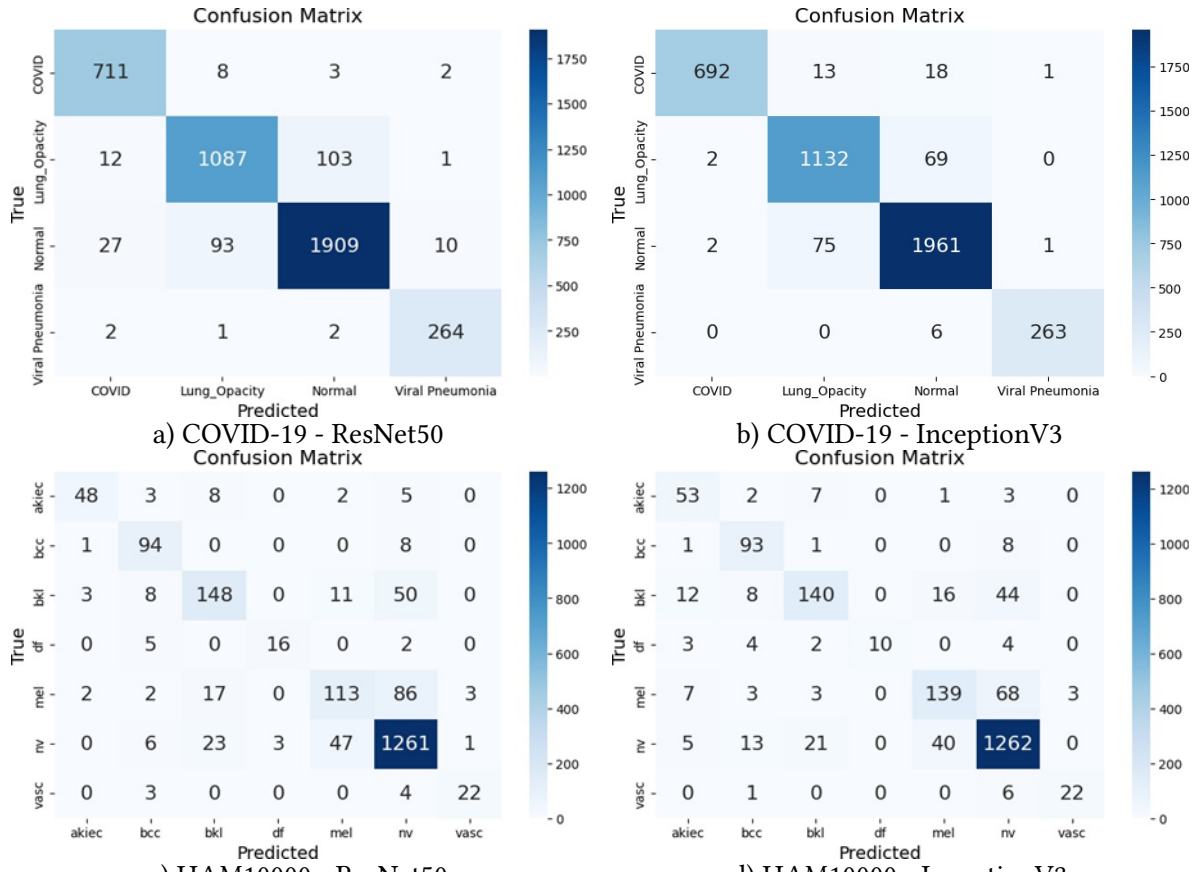
4.2.7. With 30% Augmentation

Table 9 presents the performance of the ResNet50 and InceptionV3 models on the COVID-19 Radiography and HAM10000 datasets with approach 4, evaluated in terms of accuracy, recall, precision, and F1-score.

Table 9
Performance results of the models with approach 4: augmentation using StyleGAN3 and balancing with traditional methods

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	94.49%	94.48%	94.50%	94.49%	83.99%	82.51%	83.99%	83.26%
InceptionV3	95.96%	95.96%	95.96%	95.96%	86.43%	86.20%	86.43%	86.14%

It can be noted that the InceptionV3 model continues to surpass the ResNet50 model in terms of accuracy and F1-score with this approach on the COVID-19 Radiography dataset, achieving an accuracy of 95.84% compared to 94.49% for ResNet50. Furthermore, on the HAM10000 dataset, InceptionV3 also demonstrates better performance with an accuracy of 86.43% compared to 83.99% for ResNet50.



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 7: Confusion matrices for ResNet50 and InceptionV3 on the COVID-19 and HAM10000 datasets with approach 4 at 20%.

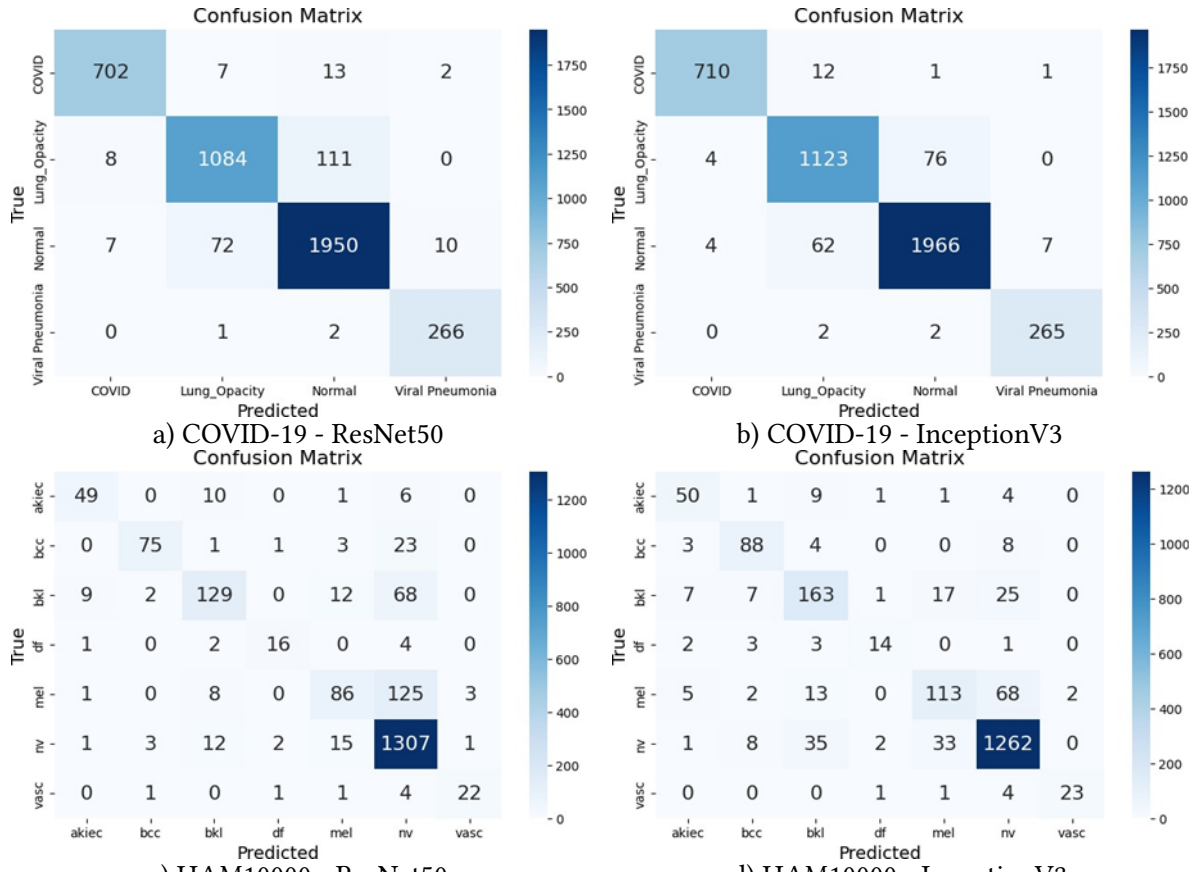
4.2.8. With 40% augmentation

The table 10 presents the performance of the InceptionV3 and ResNet50 models on the COVID-19 Radiography and HAM10000 datasets with a 40% data augmentation using approach 4.

Table 10
Performance results of the InceptionV3 model with approach 4: augmentation by StyleGAN3 and balancing with traditional methods

Model	COVID				HAM10000			
	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec
ResNet50	93.24%	93.21%	93.25%	93.27%	82.94%	81.58%	82.94%	81.93%
InceptionV3	95.18%	95.18%	95.18%	95.19%	85.38%	84.99%	85.39%	84.91%

The results show that InceptionV3 continues to exhibit better performance compared to ResNet50 in terms of accuracy and F1-score. For the COVID-19 dataset, InceptionV3 achieves an accuracy of 95.18%, while ResNet50 shows an accuracy of 93.24%. Regarding the HAM10000 dataset, InceptionV3 achieves an accuracy of 85.38%, compared to 82.94% for ResNet50, confirming the effectiveness of InceptionV3 for this approach.



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 8: Confusion matrices for ResNet50 and InceptionV3 on the COVID-19 and HAM10000 datasets with approach 4 at 30%.

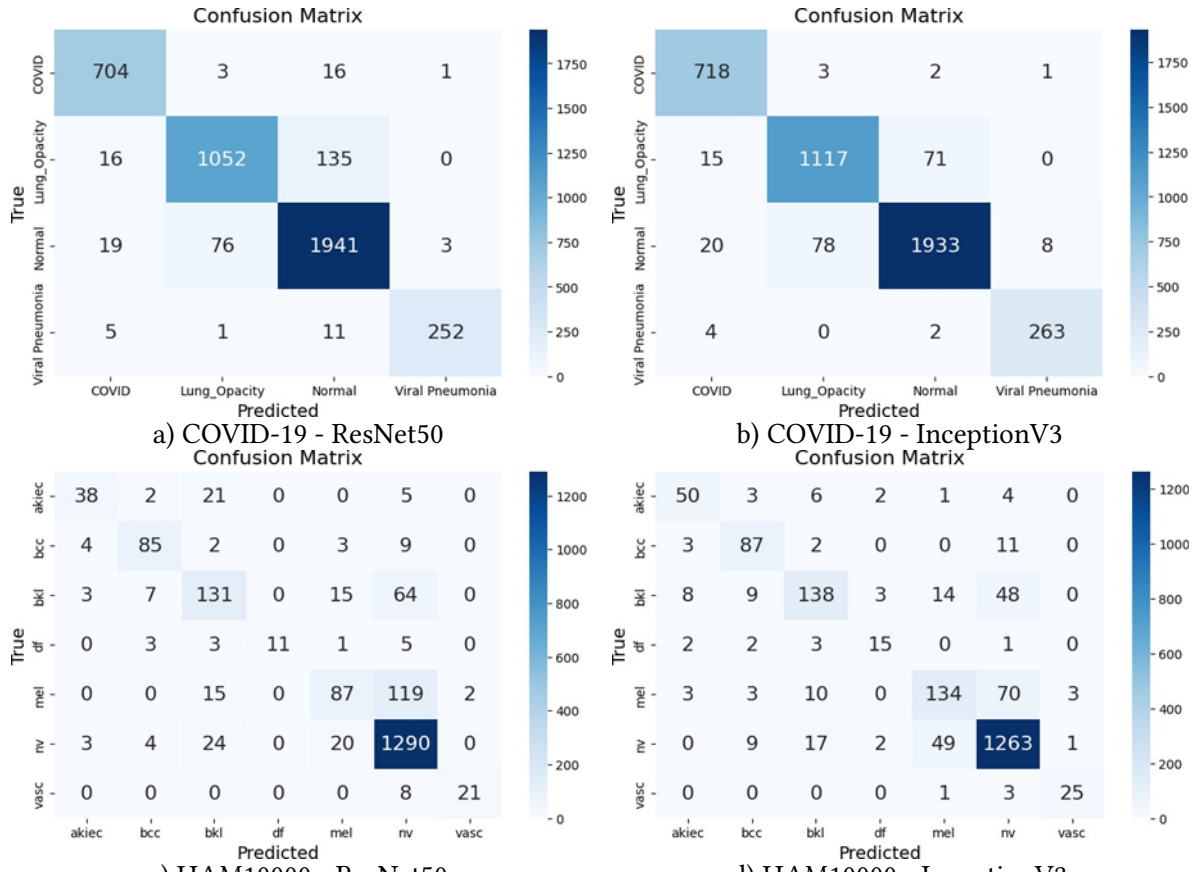
4.3. Discussion and comparison

For the COVID-19 Radiography dataset, approach 4, which includes a 30% increase in generated data, proved to be the most effective. It achieved an accuracy of 95.96% for the InceptionV3 model and 94.49% for ResNet50, outperforming all other tested approaches.

In the HAM10000 dataset, approach 4 also delivered the best results, especially for InceptionV3, which reached an accuracy of 86.43%, surpassing other methods.

Approach 1 relies on traditional data augmentation, but its limitations quickly become apparent. On the COVID-19 Radiography dataset, it allows InceptionV3 to achieve an accuracy of 94.82% and ResNet50 91.26%, while on the HAM10000 dataset, the performances are 83.09% for InceptionV3 and 82.59% for ResNet50. However, this unbalanced approach does not provide significant improvements and may even lead to decreased performance due to the persistent class imbalance. This is where approach 2, which incorporates class balancing along with data augmentation, proves to be more effective. Indeed, on the COVID-19 Radiography dataset, it achieves an accuracy of 95.89% for InceptionV3 and 94.40% for ResNet50, and on HAM10000, the results are also improved, with 86.03% for InceptionV3 and 84.49% for ResNet50. This demonstrates that the addition of class balancing allows for better model generalization, particularly on unbalanced datasets, making approach 2 more effective than approach 1.

In approach 3, although the generated images are of good quality, the main issue lies in the constant variation of the images injected into the batches at each training step. This fluctuation prevents the model from converging effectively, as it cannot adapt well to changing data. The lack of consistency in the batches disrupts the model's learning, limiting overall performance improvements. In comparison, approach 2, which uses static data balancing, offers greater stability and enables the model to converge better, thus explaining its superior results.



a) COVID-19 - ResNet50 b) COVID-19 - InceptionV3

c) HAM10000 - ResNet50 d) HAM10000 - InceptionV3

Figure 9: Confusion matrices for ResNet50 and InceptionV3 on the COVID-19 and HAM10000 datasets with approach 4.

5. Conclusion

Conclusion Data augmentation is key to optimizing the performance of deep learning models in medical imaging, especially in the presence of imbalanced and challenging datasets. In this work, the application of StyleGAN in generating synthetic images was investigated and its impact on the training of classification models evaluated.

Our experiments demonstrated that while the synthetic images generated by StyleGAN are realistic and useful for augmenting datasets, they alone are not sufficient to surpass the performance obtained using traditional data augmentation methods. The diversity and realism of the generated images remain a concern, especially with the complexity and variability of medical images, which are not always well-modeled by generative models.

However, this study shows the potential of GANs for enhancing medical classification datasets. While ResNet50 and InceptionV3 have worked effectively, other architectures and fine-tuning strategies can potentially improve model robustness.

Prospective Pathways To supplement this study, there are numerous paths that can be taken:

- Hybrid methods: Combining GANs with other generation techniques, i.e., variational auto-encoders, would further improve synthetic data quality. Systematic clinical validation: Testing these techniques in actual clinical environments to determine their feasibility.
- Combining imbalance handling techniques: Addressing dataset imbalance by integrating two approaches, such as undersampling the majority classes through pruning while simultaneously oversampling the minority classes, can improve model generalization and mitigate bias.

Table 11
Model performance results across different approaches

		COVID-19 Radiography				HAM10000			
	Model	A	F1	R	P	A	F1	R	P
	Approach 1								
	ResNet50	91.26%	91.20%	91.26%	91.50%	82.59%	81.27%	82.59%	81.48%
	InceptionV3	94.82%	94.84%	94.83%	94.88%	83.09%	82.02%	83.09%	82.49%
	Approach 2								
	ResNet50	94.40%	94.37%	94.40%	94.40%	84.49%	84.49%	83.68%	83.16%
	InceptionV3	95.89%	95.88%	95.89%	95.89%	86.03%	86.03%	85.11%	82.02%
	Approach 3								
	ResNet50	94.33%	94.31%	94.33%	94.33%	83.19%	82.64%	83.19%	82.77%
	InceptionV3	95.46%	95.45%	95.47%	95.48%	86.08%	86.19%	86.08%	86.54%
	Approach 4								
10%	ResNet50	93.76%	93.77%	93.77%	93.78%	85.28%	84.97%	85.29%	84.95%
	InceptionV3	95.37%	95.36%	95.37%	95.36%	86.43%	85.61%	86.43%	85.96%
20%	ResNet50	93.76%	93.75%	93.77%	93.75%	84.88%	84.31%	84.89%	84.22%
	InceptionV3	95.58%	95.60%	95.58%	95.62%	85.73%	85.33%	85.74%	85.65%
30%	ResNet50	94.49%	94.48%	94.50%	94.49%	83.99%	82.51%	83.99%	83.26%
	InceptionV3	95.96%	95.96%	95.96%	95.96%	86.43%	86.20%	86.43%	86.14%
40%	ResNet50	93.24%	93.21%	93.25%	93.27%	82.94%	81.58%	82.94%	81.93%
	InceptionV3	95.18%	95.18%	95.18%	95.19%	85.38%	84.99%	85.39%	84.91%

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. Garcea, A. Serra, F. Lamberti, L. Morra, Data augmentation for medical imaging: A systematic literature review, Computers in Biology and Medicine 152 (2023) 106391.

[2] W. Chen, K. Yang, Z. Yu, Y. Shi, C. P. Chen, A survey on imbalanced learning: latest research, applications and future directions, Artificial Intelligence Review 57 (2024) 137.

[3] P. Kumar, R. Bhatnagar, K. Gaur, A. Bhatnagar, Classification of imbalanced data: review of methods and applications, in: IOP conference series: materials science and engineering, volume 1099, IOP Publishing, 2021, p. 012077.

[4] S. Yadav, G. P. Bhole, Handling imbalanced dataset classification in machine learning, in: 2020 IEEE Pune Section International Conference (PuneCon), IEEE, 2020, pp. 38–43.

[5] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural networks 106 (2018) 249–259.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[7] F. Touazi, D. Gaceb, M. Chirane, S. Herzalleh, Two-stage approach for semantic image segmentation of breast cancer: Deep learning and mass detection in mammographic images, in: International Conference on Informatics & Data-Driven Medicine, 2023, pp. 1–13.

[8] M. Khaled, F. Touazi, D. Gaceb, Improving breast cancer diagnosis in mammograms with progres- sive transfer learning and ensemble deep learning, Arabian Journal for Science and Engineering (2024) 1–24.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems 27, 2014, pp. 2672–2680.

[10] J. Islam, Y. Zhang, GAN-based synthetic brain PET image generation, Brain Informatics 7 (2020).

[11] A. Teramoto, H. Tsukamoto, H. Kiriyama, J. Fujita, H. Yamamoto, T. Tsukiji, Y. Imaizumi, T. Toyama, T. Oda, T. Kudo, Deep learning approach to classification of lung cytological images:

Two-step training using actual and synthesized images by progressive growing of generative adversarial networks, PLOS ONE 15 (2020) e0229951.
[12] A. Waheed, M. G. G. Khan, M. Ali, M. A. G. Javed, J. Liao, A. I. T. S., G. Yang, CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection, IEEE Access 8 (2020) 91916–91923.
[13] V. Sandfort, K. Yan, P. J. Pickhardt, R. M. Summers, Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks, Scientific Reports 9 (2019) 16884.
[14] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
[16] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, Advances in neural information processing systems 34 (2021) 852–863.
[17] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P. R. Pinheiro, Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection, IEEE Access 8 (2020) 91916–91923.
[18] S. V. J, J. F. D, Deep learning algorithm for covid-19 classification using chest x-ray images, Computational and Mathematical Methods in Medicine 2021 (2021) Article ID 9269173, 10 pages.
[19] B. Ahmad, S. Jun, V. Palade, Q. You, L. Mao, M. Zhongjie, Improving skin cancer classification using heavy-tailed student t-distribution in generative adversarial networks (ted-gan), Diagnostics 11 (2021) 2147.
[20] Q. Su, H. N. A. Hamed, M. A. Isa, X. Hao, X. Dai, A gan-based data augmentation method for imbalanced multi-class skin lesion classification, IEEE Access 12 (2024) 16498–16513.
[21] A. Bilal, J. Sun, Q. You, V. Palade, Z. Mao, Brain tumor classification using a combination of variational autoencoders and generative adversarial networks, Biomedicines 10 (2022) 1–19.
[22] G. C. Oliveira, G. H. Rosa, D. C. G. Pedronette, J. P. Papa, H. Kumar, L. A. Passos, D. Kumar, Which generative adversarial network yields high-quality synthetic medical images: Investigation using amd image datasets, arXiv preprint arXiv:2203.13856v1 (2022).
[23] P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific data 5 (2018) 1–9.
[24] U. Rehman, T. COVID, Radiography database| kaggle, 19. URL: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.