International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Improving Breast Cancer Performance in CNN by Generating Synthetic Histopathological Images using GAN and Traditional Augmentation

Muhamad Shah Azizie Abd Karim[1] **and Marina Yusoff**[1,2,3*]

[1]*College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia*
[2]*Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) Kompleks Al-Khawarizmi, Universiti Teknologi MARA (UiTM), Shah Alam, Selangor Darul Ehsan Malaysia*
[3]*Faculty of Business, Sohar University, Oman*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: In the pursuit of more accurate cancer detection through breast cancer histopathology (BCH) images, Convolutional Neural Networks (CNNs) have emerged as promising tools. However, CNNs still face limitations, necessitating advancements in classification performance. This research addresses these challenges by harnessing the power of Generative Adversarial Networks (GANs) as data augmentation to optimize CNN models for BCH image classification. This paper addresses the proposed two-stage augmentation strategies based on GAN and the traditional method. The BreakHis dataset was employed to investigate the efficacy of GAN-based data augmentation. The research adopted a transfer learning approach, namely Inception-V3, and VGG16, and fine-tuned them with a single GAN and the two stages augmentation methods. The novel integration of GANs and traditional augmentation enhanced the training dataset, enabling the models to learn from a more diverse and extensive image distribution. Extensive trials demonstrated that the top-performing architecture, Inception-V3 + TradAug, attained a remarkable 97.12% accuracy with 0.1014 loss value, showcasing the effectiveness of the composition of GAN and traditional augmentation in optimizing BCH image classification. The two-stage integration of GANs, such as data augmentation and traditional augmentation, empowers CNN models to identify cancerous conditions accurately. This research signifies a significant step towards enhancing breast cancer classification through advanced AI-driven methodologies.

Keywords: Breast Cancer, Deep Learning, Image Classification, Convolutional Neural Network, Generative Adversarial Network

1. INTRODUCTION

Cancer is a significant health problem that has impacted many lives today. One type of cancer that has affected many women around the world is breast cancer. Analyzing tissue samples from Breast Cancer Histopathology (BCH) images to identify the presence and severity of malignant cells is a vital aspect of the diagnostic process [1]. Implementing an automatic system that can classify these images can be quite a challenging problem due to the high variability and complexity of tissue images. The diagnosis of breast cancer is a time-consuming procedure as it requires the assistance of expert pathologists. In contrast, pathologists base their decisions on various visible characteristics seen on pathology slides, such as the morphological characteristics of nuclei [2]. Therefore, automating this process will be very helpful in the long term. Using a computer-aided diagnosis system is a significant development in trying to curb this challenge. To make the process of cancer diagnosis faster, these systems can be that tiny little spark needed to save a person's life.

Various machine learning algorithms have been applied to diagnise breast cancer images, with promising results [3] [4] [5] [6]. In a similar vein, as highlighted in the study on diagnosing Diabetes Mellitus using machine learning techniques [7], the identification of relevant features and selection of efficient classifiers are pivotal for accurate diagnosis. However, these approaches have their drawback. Applying deep learning for medical images often suffers from limited training data and is very expensive to acquire, which can hinder the training process of the deep learning models [8]. Generating additional synthetic data from existing examples is an effective way to overcome this issue. Deep generative models (DGMs) are multi-layer neural networks trained to approximate complex, high-dimensional probability distributions using a large sample

E-mail address: author 1, author 2, author 3

size [9]. Numerous methodologies exist for constructing deep generative models; one strategy involves leveraging GANs. This research explores using GAN for data augmentation in classifying BCH images.

Traditionally, deep convolutional neural networks primarily require significant data to prevent overfitting [10]. To circumvent this challenge, an advocated technique involves employing data augmentation to amplify the size of the limited dataset. These data need to undergo some form of transformation; otherwise, these networks will not be able to produce acceptable results. Thus, to better train these deep neural networks, there is a need to gather more data, which can be expensive and challenging to do, especially when patients have privacy concerns. Transforming those limited data to produce more variation of the same data is an approach to solve the problem from the root, the training data itself. A CNN model for classifying these histopathological images has yielded promising results [11]. In addition, AlSayyed et al. [12] have also demonstrated the use of pretrained CNN models, VGG16, VGG19, and DenseNet121, in classifying oral photographs. The architectural paradigm of GANs encompasses a dual-model framework, integrating a generative component and a discriminative counterpart, typically realized through the implementation of neural networks.

These networks can capture the distribution of real data examples and create new data using those examples [13]. Leveraging a GAN's capability to generate synthetic augmentations of BCH image samples could potentially enhance a classifier's performance on this task by escalating the volume of the training data. It may be adequate to develop a method that can effectively synthesize high-quality examples of medical images that capture the subtle and complex patterns indicative of cancerous cells and to evaluate the impact of this approach on the accuracy of the classifier. GANs, in the context of medical image augmentation, have also undergone many changes. Chen et al. [14] classify GAN models into three types: GANs based on random latent vectors, image transformations, and classical transformations. GAN is very effective for data augmentation because synthetic examples produced by the GAN can be more realistic and diverse than those generated by traditional data augmentation techniques.

The motivation for doing this research stems from the importance of accurately classifying BCH images to improve patient outcomes. Despite the significant progress made in this area using machine learning algorithms, there is still a need for improved methods that can handle the high variability and complexity of these images. While expanding the training data volume represents an efficacious approach to bolster the performance of machine learning models, conventional data augmentation techniques like rotation, cropping, and scaling may prove inadequate for histopathology image analysis. This inadequacy stems from the intricate and nuanced patterns of cancerous cells, which demand a more sophisticated approach. In this study, the capabilities of GAN as a data augmentation technique will be analyzed and applied to improve the performance of the classification of BCH. The main challenges in classifying BCH images are data scarcity and the limited availability of high-quality training data. This problem can be explained by the fact that annotating images with diagnostic labels is time-consuming and labor-intensive [15].

Consequently, machine learning models whose training relied on these datasets may exhibit suboptimal generalization capabilities when confronted with previously unseen images, thus culminating in diminished performance assessments on test data. As a result, machine learning models trained on these datasets may not generalize well to never-before-seen images, leading to poor performance on test data. This makes GANs well-suited to improving the classification of BCH images. The objectives of this research encompass identifying the constraints of conventional data augmentation methods in enhancing CNN classifier performance, constructing a GAN architecture endowed with the capacity to synthesize BCH image samples of elevated fidelity, and employing both GAN and CNN models for BCH image classification.

2. MATERIAL AND METHOD

This section outlines the different methods and techniques used in this study to classify the BCH images. Figure 1 shows the proposed process flow. It details the acquisition and characterization of the dataset, the preprocessing procedures applied, the methods for data augmentation, the approach to data partitioning, the classifier architecture utilized, and the evaluation metrics employed.

A. Data Acquisition and Pre-processing

This research focuses on the BreakHis dataset [16], consisting of 7,909 histopathology images of breast tumors labelled as benign and malignant; it was taken from 82 different patients and divided into 2,480 benign and 5,429 malignant samples, each with size of 700x460 pixels and an 8-bit depth in RGB in PNG format. Initially, both class labels (benign and malignant) were also sorted into different types and patient IDs. Still, for the purpose of this research, both factors were ignored, and only the class labels were considered.

Various adjustments were made to the original dataset to ensure the data can be adequately trained and tested in the model developed for this research. These preliminary steps will include pre-processing the original dataset better to fulfill the input requirement of the GAN model, balancing the amount of data in the class label with the much smaller sample, and dividing the dataset into training, evaluation, and testing.

Data pre-processing refers to transforming the original dataset acquired into a more suitable form for analysis and training. The first step in this process appropriately involved analyzing the data format. Since the dataset is image-based,

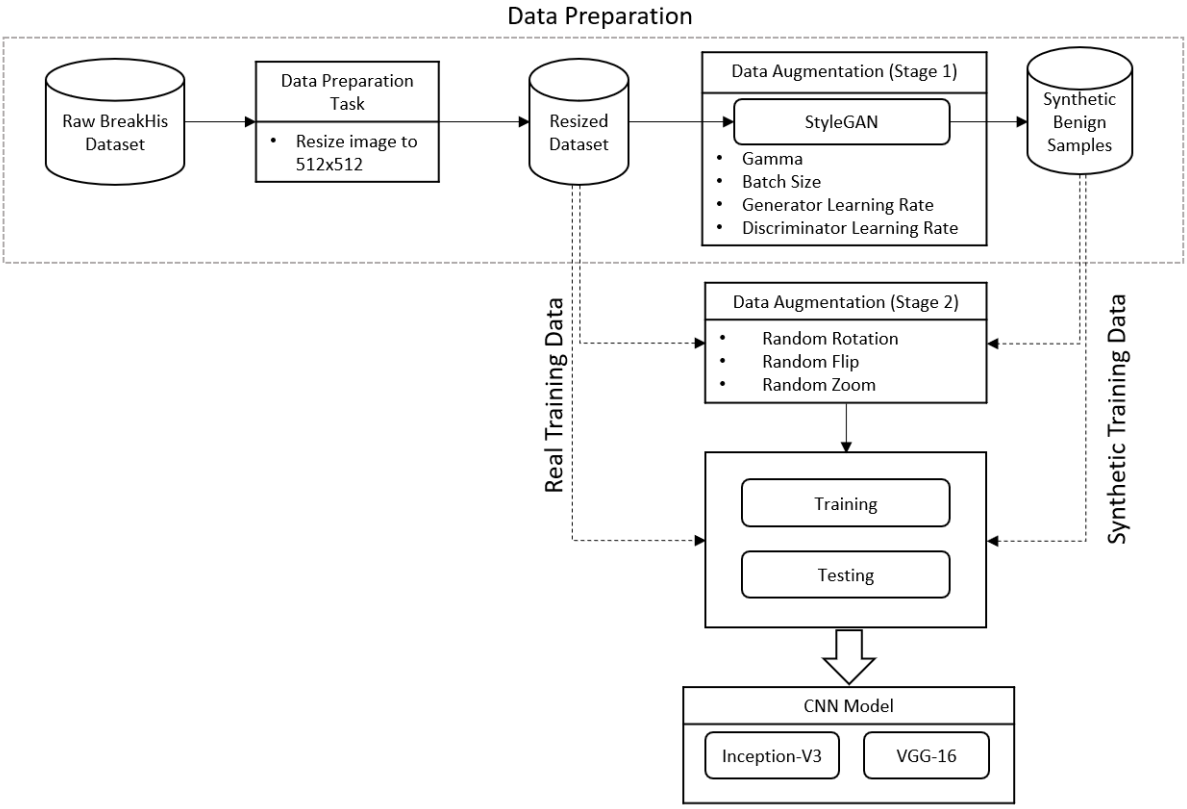

Figure 1. Proposed Process Flow

the image data is stored in PNG format. Further analysis also shows that the images needed to be resized from 700x460 pixels to a much more suitable 512x512 pixels size to accommodate the requirements of the StyleGAN model.

B. Proposed Two Stages Data Augmentation

Since the number of samples between each class label differs, the malignant sample has almost twice the number of benign samples. This will cause the model to be much better at telling apart malignant than benign samples since there is a bias towards the class with the bigger size. The proposed solution involves utilizing GAN to generate synthetic samples of the minority class.

The stage one focuses on GAN for Data Augmentation. Generative models, such as GANs, can generate new and diverse samples from the training data distribution [17]. GANs are useful for data augmentation because they can generate synthetic data examples like the original training examples. This can be particularly useful when the size or diversity of the original training dataset is limited, as it allows for the expansion of the dataset with additional examples that are like the actual data [14]. This research will exclusively concentrate on generating 640 synthetic benign images. The number of samples for this research is shown in Table I. This deliberate effort aims to enhance the overall balance of the dataset, thereby leading to improved outcomes and results.

It is important to note that how well a GAN-based data augmentation approach works will depend on both the quality of the actual GAN model and the specific traits of the training data being used. It may be necessary to carefully design and tune the GAN model to achieve good results, and the synthetic examples produced may sometimes be of different quality than real examples.

There are no GAN models specifically designed for synthetic image generation of BCH. However, some GAN models are designed to be generalizable and can be used in many domains. For example, StyleGAN and its variants have been used to generate synthetic images that can be added to the original dataset, increasing its size [18]. Figure 2 shows an overview of StyleGAN's architecture.

StyleGAN is an architecture designed to generate high-quality images with realistic details. It was introduced to improve upon the limitations of traditional GANs. The primary innovation of StyleGAN lies in its use of a style-based generator and a mapping network [19]. Unlike traditional GANs, where the input noise directly influences each layer, StyleGAN uses a separate latent vector called 'style vector'. These style vectors are transformed from a common input through a mapping network.

TABLE I. BreakHis data samples

Image Label	No. of Samples
Benign	2480
Malignant	5429
Benign (Synthetic)	640
Malignant (Synthetic)	0

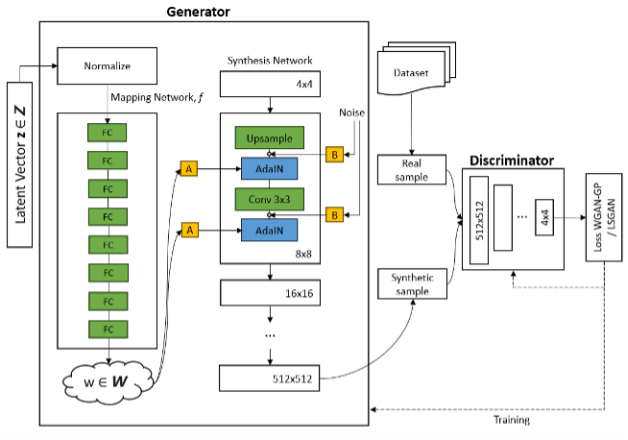

Figure 2. The architecture of StyleGAN

StyleGAN offers several unique features and advantages that set it apart from other GANs making it an attractive choice for this research. A key reason was that StyleGAN is built to create highly realistic, high-quality images packed with fine details. Its hierarchical setup and style-based generator allow for precise control over different aspects of the image-generation process, like textures, colors, and shapes. This results in visually appealing images that closely resemble real-world images. Moreover, StyleGAN has consistently demonstrated state-of-the-art image quality and diversity performance compared to many other GAN architectures. Its innovative design and attention to local and global realism have yielded impressive results in various domains. By training the StyleGAN model with the BreakHis dataset, its capabilities to produce realistic images of BCH were measured and recorded.

This research utilized a repository that faithfully implements StyleGAN in PyTorch [20]. A significant drawback of the initial version of StyleGAN is its tendency to produce blob-like artifacts in certain instances. StyleGAN2 was subsequently developed to address this issue. StyleGAN2-ADA builds upon StyleGAN2 by introducing adaptive data augmentation during the training process. In this research phase, applying transfer learning on a StyleGAN2 model pre-trained on the BreCaHAD [21] dataset, the model can be further trained on the BreakHis dataset. Due to hardware limitations, the resolution was set to produce images of size 512x512 and only trained using a single GPU. After various iterations, the quality of the output images was then recorded.

Stage two is on traditional augmentation (TradAug). Geometric augmentations, such as reflecting, cropping, and translating, are common and acceptable approaches for image data augmentation [22]. Traditional data augmentation techniques, often dependent on fundamental geometric transformations, exhibit limitations. Although they can expand the training dataset, enhancements are necessary to enhance their capacity for generating diverse and representative samples reflective of the underlying data distribution. This research involves labeled data comprising benign and malignant samples, as outlined in Table I, which will be evaluated with and without traditional augmentation methods. Specifically, the images will be applied with a random transformation of vertical and horizontal flipping, rotation, and zoom during training. A regularization technique will also be used by adding a dropout layer to the model architecture.

C. Data Splitting

The final steps in the data preparation and balancing are splitting the dataset further into training, evaluation, and testing subsets. The testing set will act as unseen data to evaluate the final performance of the model. The training set is used to learn the model's parameters; the evaluation partition is employed to calibrate the model's parametric configurations and determine the preeminent architecture. The testing subset, in turn, facilitates the assessment of the conclusive model's performance capabilities on unseen data samples during the training and evaluation phases. In this research endeavor, the training set constitutes 80% of the available data, the evaluation set comprises 10%, and the remaining 10% is allocated for testing purposes.

D. Classifier

After the GAN model was determined to produce good enough synthetic BCH images, these images were used in conjunction with the original training data, which involves building CNN models capable of classifying BCH images as benign or malignant. The methodological approach shall entail leveraging deep transfer learning techniques to fine-tune a pre-trained convolutional neural network architecture, Inception-V3 and VGG16, on the BreakHis dataset. Deep transfer learning entails transferring and adjusting the weights of a standard model, initially trained on large datasets, to perform similar tasks on a smaller dataset.

These convolutional neural network architectures have demonstrated proficient performance across many image classification undertakings, having trained on the extensive

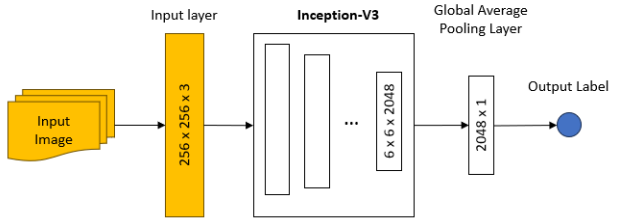

Figure 3. Architecture of Inception-V3

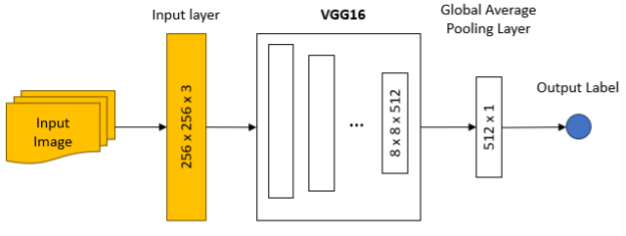

Figure 4. Architecture of VGG16

ImageNet dataset. Additionally, separate CNN models will be trained with and without using the generated synthetic images to conduct a comparative evaluation and quantify the ramifications exerted by the synthesized image data. This approach allows for a comprehensive assessment of the generated data's effectiveness in enhancing the BCH classification's performance. The research will use both models, and their performances will be recorded. Figure 3 and Figure 4 showcase the architecture of Inception-V3 and VGG16, respectively, alongside additional layers added for the use case of this research. The implementation of Inception-V3 is similarly discussed in a study by Ali et al. [23], but VGG16 is discussed similarly by Lee Ji-Hee and Lee Jee-Hyong [24]. Incorporating pre-trained deep neural networks like Inception-V3 and VGG16 into this research provides a significant advantage in classifying BCH images. These advantages include the robust feature extraction system of both models. The Inception-V3 model is known for its ability to capture multi-scale features using a combination of convolutional filters with different kernel sizes. This is particularly useful since this research involves images captured at different magnification factors and complex structures. Inception-V3's inception modules can efficiently capture intricate details, providing more comprehensive feature representations. In addition, VGG16 is recognized for its simple yet effective architecture with multiple convolutional layers of small filter sizes. This architecture facilitates the extraction of finer details from images. Since this research requires precise texture and pattern analysis, VGG16's deep layers can excel in capturing these features.

E. Performance Evaluation

Once all models have been developed, the next phase evaluates and tests them to determine their performance. This research has evaluated GAN and CNN models for their ability to perform a specific task related to BCH images. Both models were trained on the same dataset, and their performances were evaluated using different metrics.

The evaluation of the GAN architecture will prioritize metrics that quantify the caliber of the synthesized image data, such as the Fréchet Inception Distance (FID) and Kernel-Inception Distance (KID) measures. These metrics facilitate the assessment of similarity between the generated and authentic image samples, furnishing a quantitative gauge of the GAN's proficiency in encapsulating the underlying distribution inherent to the training data corpus. The evaluation of these metrics is of paramount importance for ascertaining the efficacy of image generative models.

The FID is a metric devised to evaluate image generative model fidelity. It is derived from feature representations computed by a pre-trained Inception-V3 network architecture. It assumes these features adhere to a Gaussian distribution and employs the Fréchet distance between two multivariate Gaussian distributions to quantify the discrepancy between authentic and synthesized image samples. A lower FID score indicates a more proficient generative model regarding output quality.

To calculate the FID, one must extract the features of real and generated images on the coding layer of the Inception-V3 network [25]. The images' activations are aggregated into a multivariate Gaussian distribution by computation of the mean and covariance. The FID between the two Gaussians is then calculated as in Equation (1).

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\sum c_r + c_g - 2\left(\sqrt{c_r}\,\sqrt{c_g}\right)\right) \quad (1)$$

where:

- μ_r is the mean of the real data distribution,
- μ_g is the mean of the generated data distribution,
- c_r and c_g are the covariance matrices of the real and generated data distributions, respectively, and
- Tr denotes the trace operator.

In addition, the KID is an image quality metric proposed to replace the popular FID. Both measures assess the disparity between the representation spaces of a pre-trained Inception-V3 network on ImageNet, comparing the generated and training distributions. KID is more straightforward to implement, can be estimated per batch, and is computationally lighter than FID. Using a polynomial kernel, KID measures the squared Mean Discrepancy (MMD) between the Inception representations of actual and generated samples. It also measures the distance between probability distributions [26].

In examining the CNN model, the emphasis lies on assessing its classification efficacy, gauged through metrics like accuracy, precision, recall, and F1-score. This evaluation is grounded in a confusion matrix and the outcomes of a loss function. These metrics assess the ability of the CNN model to accurately classify the images into different categories, providing a measure of its ability to learn and generalize from the training data. Equation (2), (3), (4) and (5) are the performance measures of the classification.

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{FP + TP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1 Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

For binary classification problems, binary cross-entropy is a standard loss function used to measure the difference between the predicted probabilities and the actual labels. It is widely used in training logistic regression and neural network models for binary classification tasks [27]. Minimizing the binary cross-entropy during training means that the model is trying to maximize the likelihood of predicting the true labels correctly for each instance.

Both evaluations are conducted on the data's validation set and test set, and the results are compared to determine which model is better suited for the task at hand. Based on the results, the research will be able to conclude which one model is superior to the other or that the performance of the models is comparable, which is to find the best models for generating synthetic images of BCH and an improved CNN model that are trained alongside the synthetic images.

3. Experimental Results

A. Performance of GAN

This preliminary stage aims to train a GAN model to produce high-fidelity images of BCH. Two key evaluation metrics, namely the FID and KID scores, were employed to assess the training process and the model performance. In Table II, you can find the parameter settings used for training the StyleGAN model. In this experiment, 'kimg' represents the number of images in thousands that have been cycled during the training run, specifically denoting the number of benign samples for each configuration. To enhance the performance of the StyleGAN model, various parameters, such as the gamma value and batch size, were adjusted to fit better the dataset used in this research.

Table III presents the FID and KID scores obtained for various γ values in each configuration. γ of 1.0 resulted in the lowest FID score of 16.1145 and KID score of 0.0030. Throughout these computations, the batch size remained constant at 16. Table IV presents the FID and KID scores obtained for various batch sizes in each configuration. Batch size 32 resulted in the lowest FID and KID score of 15.9542 and 0.0039 respectively. Throughout these computations, the γ value remained constant at 1.0.

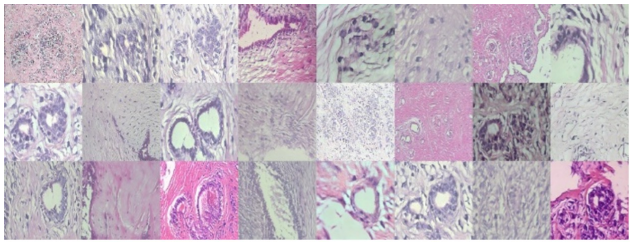

Figure 5. Samples of generated benign BCH images after training

The inconsistency in the amount of image processed is the different amount of time when the model converged to a certain level of performance. Continuing the training process might not be cost-effective regarding computational resources and time, as it is unlikely to yield a considerable enhancement in the generated image quality beyond the observed threshold.

As observed in Figure 5, the generated benign BCH samples exhibit a remarkably high quality and visual fidelity. Using such high-quality synthetic data for dataset expansion can prove particularly beneficial in scenarios, where obtaining a large, manually labeled dataset is challenging and costly. The quality and diversity exhibited by the StyleGAN-generated images make them a worthy idea for data augmentation, and these images are then used in fine-tuning a CNN model.

B. Performance of STYLEGAN + CNN

To leverage the pre-trained Inception-V3 and VGG16 model's knowledge, new layers were added on top of the pre-trained models, which will be referred to as the base models. To retain the knowledge captured by the base models, we froze the weights of the base model layers. This ensured that the lower-level image features remained unchanged during training. The training focuses on only changing the newly added binary classification layers using the target BreakHis dataset. By fine-tuning these layers, the model could adjust its parameters to suit the characteristics of the specific dataset while benefiting from the general features learned from the base layers. Table V shows the parameters and values used for the CNN models to classify BCH images. The experiment involves five epochs, 10, 20, 30, 40, and 50, and uses two pre-trained models called Inception-V3 and VGG16.

Table VI and Table VII show transfer learning results with Inception-V3 and VGG16 models obtained at the specified epochs. These models can be considered the baseline since no additional data augmentation techniques were used during the training. Additionally, Table XI shows the classification results of both models.

TABLE II. Parameter settings of StyleGAN

Parameter	Values
Training Data	2480 (Real) Benign Images
Configuration	StyleGAN2
No. of GPU	1
Learning Rate (Generator)	0.0025
Learning Rate (Discriminator)	0.0025
Image Input Size	512x512
Image Output Size	512x512
Gamma	0.5, 1.0, 1.5, 1.55, 2.0, 5.0
Batch	16, 32, 64, 128

TABLE III. Results of StyleGAN training for training gamma values

Image Processed (kimg)	Gamma, γ	FID Score	KID Score
240	0.5	21.1684	0.0049
300	**1.0**	**16.1145**	**0.0030**
300	1.5	17.5853	0.0031
300	1.55	18.4753	0.0041
300	2.0	18.7472	0.0047
120	5.0	17.7404	0.0033
320	6.55	18.0501	0.0039

TABLE IV. Results of StyleGAN training for various batch sizes

Image Processed (kimg)	Batch Size, B_{size}	FID Score	KID Score
620	16	17.1803	0.0040
600	**32**	**15.9542**	**0.0039**
644	64	16.4828	0.0037
690	128	18.6250	0.0052

TABLE V. Parameter settings of CNN

Parameter	Values
Epoch	10, 20, 30, 40, 50
Batch Size	64
Learning Rate	0.0001
Transfer Learning Models	Inception-V3, VGG16

Table VIII and Table IX show transfer learning results where traditional data augmentation (TradAug), such as image rotation, vertical and horizontal flipping, and zooming, were used during the training process, for Inception-V3 and VGG16 models obtained at the specified epochs. In addition, a dropout layer was added as a form of regularization that perturbs the neural network's architecture during training by randomly setting some activations to zero. Additionally, Table XI also shows the classification results of both models.

To help improve the performance of the CNN models even further, the generated StyleGAN output was utilized as a new dataset and will be trained alongside the original dataset. Table X presents a summary of the outcomes of the four models prior to their training on the StyleGAN dataset. During this training, only the first 100 layers of the Inception-V3 model and the first 16 layers of the VGG16 models remained frozen during the training process. By freezing these layers, the model effectively prevents itself from altering the lower-level features that were already learned from the large-scale dataset on which Inception-V3 and VGG16 were trained on before.

Table XII shows the parameters and values that were used for the training of the CNN models to classify BCH images as they are being fine-tuned with dataset generated by the StyleGAN model. The experiment only runs for 10 epochs and uses the four models trained in the previous experiment. Table XIII and Table XIV provide an overview

TABLE VI. Computation results of Inception-V3

Epoch	Accuracy			Loss		
	Training	Validation	Testing	Training	Validation	Testing
10	0.8187	0.8061	0.8164	0.4052	0.4116	0.4142
20	0.8430	0.8331	0.8372	0.3627	0.3777	0.3729
30	0.8551	0.8442	0.8516	0.3389	0.3583	0.3504
40	0.8658	0.8405	0.8594	0.3228	0.3561	0.3342
50	**0.8707**	**0.8282**	**0.8633**	**0.3102**	**0.3578**	**0.3211**

TABLE VII. Computation results of VGG16

Epoch	Accuracy			Loss		
	Training	Validation	Testing	Training	Validation	Testing
10	0.8183	0.8241	0.7904	0.4211	0.4082	0.4618
20	0.8650	0.8672	0.8490	0.3254	0.3393	0.3593
30	0.8856	0.8782	0.8633	0.2896	0.2952	0.3189
40	0.8954	0.8819	0.8854	0.2684	0.2825	0.2940
50	**0.9009**	**0.8868**	**0.8867**	**0.2539**	**0.2642**	**0.2771**

TABLE VIII. Computation results of Inception-V3 + TradAug

Epoch	Accuracy			Loss		
	Training	Validation	Testing	Training	Validation	Testing
10	0.8038	0.8037	0.8268	0.4268	0.4213	0.4094
20	0.8307	0.8282	0.8320	0.3824	0.3883	0.3711
30	0.8381	0.8454	0.8529	0.3587	0.3691	0.3504
40	0.8483	0.8368	0.8529	0.3442	0.3653	0.3361
50	**0.8586**	**0.8380**	**0.8542**	**0.3346**	**0.3662**	**0.3249**

TABLE IX. Computation results of VGG16 + TradAug

Epoch	Accuracy			Loss		
	Training	Validation	Testing	Training	Validation	Testing
10	0.7348	0.7823	0.7760	0.5948	0.4880	0.5052
20	0.7838	0.8155	0.8346	0.4946	0.4073	0.4170
30	0.8089	0.8487	0.8620	0.4296	0.3569	0.3732
40	0.8295	0.8585	0.8568	0.3896	0.3343	0.3473
50	**0.8289**	**0.8672**	**0.8659**	**0.3903**	**0.3197**	**0.3287**

TABLE X. Computation results of CNN for different models (without StyleGAN)

CNN Model	Accuracy			Loss		
	Training	Validation	Testing	Training	Validation	Testing
Inception-V3	0.8707	0.8282	0.8633	0.3102	0.3578	0.3211
Inception-V3 + TradAug	0.8586	0.8380	0.8542	0.3346	0.3662	0.3249
VGG16	0.9009	0.8868	0.8867	0.2539	0.2642	0.2771
VGG16 + TradAug	0.8289	0.8672	0.8659	0.3903	0.3197	0.3287

TABLE XI. Classification results of CNN for different models (without StyleGAN)

CNN Model	Benign			Malignant		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Inception-V3	0.8827	0.6810	0.7688	0.8744	0.9608	0.9156
Inception-V3 + TradAug	0.9023	0.6767	0.7734	0.8737	0.9683	0.9186
VGG16	0.8558	0.7672	0.8091	0.9036	0.9440	0.9234
VGG16 + TradAug	0.8418	0.7112	0.7710	0.8829	0.9422	0.9116

of the experimental result alongside StyleGAN, trained with parameters of $\gamma = 1.0$ and $\gamma = 6.55$.

4. Discussion

A. Discussion of CNN + TradAug

As shown in Table XIII, the results obtained have shown that Inception-V3 and VGG16 achieved an accuracy of 86.33% and 88.67%, respectively, without employing any supplementary data augmentation techniques. Meanwhile, using traditional data augmentation, the results show that accuracy decreased to 85.42% and 86.59%, respectively. A possible cause for this reduction is adjusting hyper-parameters as more complex augmentation is introduced [28]. Due to the nature of the experiment, where data augmentation is applied to a pre-trained model, it becomes essential to carefully tune the augmentation parameters to strike the right balance between introducing diversity and avoiding overfitting. Adding new hyperparameters related to augmentation might require extensive tuning to achieve optimal performance.

Moreover, the complexity of augmentations can potentially introduce unrealistic or noisy variations in the data, making it more challenging for the model to learn meaningful features [29]. This phenomenon could further contribute to the decline in classification accuracy compared to the scenario without augmentation. Additionally, the target dataset for fine-tuning or transfer learning may be limited in size, and the introduction of complex augmentations could exacerbate the issue of insufficient data [28]. The model might struggle to generalize well with the augmented data, leading to a decrease in accuracy. However, leveraging GAN is one way to overcome this limitation and potentially increase accuracy.

B. CNN with StyleGAN + TradAug

This section discusses the results of CNN with StyleGAN+ TradAug for training and testing. Figure 7 shows the point during the training process where StyleGAN data was applied. The performance shown included the training and validation accuracy and loss from epoch 50 until 60.

Figure 7 showcases that 3 of the 274 benign images in the test dataset are still falsely labelled as malignant. Meanwhile, 21 of the 558 malignant images are still incorrectly labeled. This highlights that even for the best models, there still remain instances of misclassification. Despite the overall effectiveness of the models, these findings underscore the persistent challenge of accurately distinguishing between benign and malignant images in some cases.

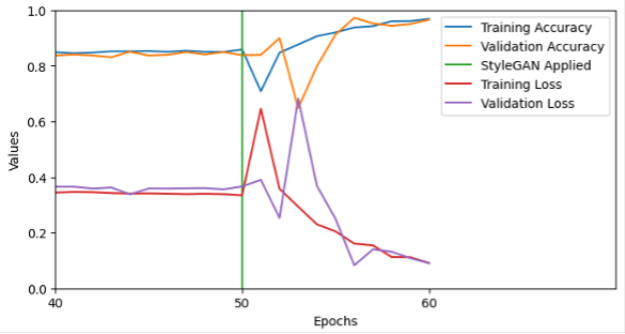

Figure 6. Training and validation accuracy and loss graph, for Inception-V3 + StyleGAN + TradAug

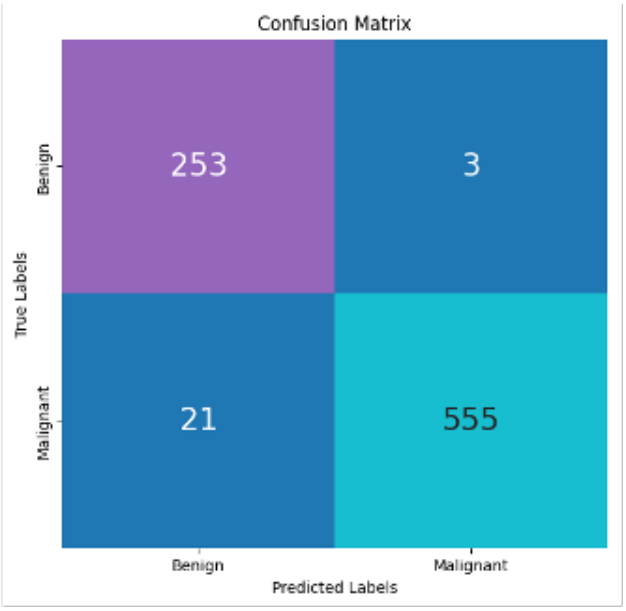

Figure 7. Confusion matrix of the best CNN Model using Inception-V3 + TradAug + StyleGAN

The findings in Table XIII and Table XIV revealed that the model achieved excellent results on the Inception-V3 model with TradAug applied and trained alongside

TABLE XII. Parameter settings of CNN + StyleGAN

Parameter	Values
Epoch	10
Batch Size	64
Learning Rate	0.0001
Transfer Learning Models	Inception-V3, VGG16, Inception-V3 + TradAug, VGG16 + TradAug
Dataset	Original (2480 Benign, 5429 Malignant) + GAN (640 Benign, 0 Malignant)

TABLE XIII. Computation results of CNN for different models and StyleGAN parameters

CNN Model	Data Augmentation	StyleGAN Parameters	Accuracy			Loss		
			Training	Validation	Testing	Training	Validation	Testing
Inception-V3	StyleGAN	γ = 1.0, B_{size} = 32	0.9663	0.9291	0.9363	0.1145	0.1507	0.1707
	StyleGAN	γ = 6.55, B_{size} = 16	0.9688	0.8365	0.8558	0.0954	0.5242	0.5163
	StyleGAN + TradAug	**γ = 1.0, B_{size} = 32**	**0.9682**	**0.9651**	**0.9712**	**0.0913**	**0.0904**	**0.1014**
	StyleGAN + TradAug	γ = 6.55, B_{size} = 16	0.9650	0.7560	0.8293	0.0982	0.3830	0.3657
VGG16	StyleGAN	γ = 1.0, B_{size} = 32	0.9846	0.8353	0.8654	0.0494	0.4084	0.3989
	StyleGAN	γ = 6.55, B_{size} = 16	0.9788	0.9471	0.9567	0.0767	0.0965	0.1027
	StyleGAN + TradAug	γ = 1.0, B_{size} = 32	0.9213	0.9399	0.9459	0.1990	0.1722	0.1546
	StyleGAN + TradAug	γ = 6.55, B_{size} = 16	0.9290	0.9279	0.9399	0.1901	0.1705	0.1618

TABLE XIV. Classification results of CNN for different models and StyleGAN parameters

CNN Model	Data Augmentation	StyleGAN Parameters	Benign			Malignant		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
Inception-V3	StyleGAN	γ = 1.0, B_{size} = 32	0.8395	0.9805	0.9045	0.9906	0.9167	0.9522
	StyleGAN	γ = 6.55, B_{size} = 16	0.6838	0.9883	0.8083	0.9935	0.7969	0.8844
	StyleGAN + TradAug	**γ = 1.0, B_{size} = 32**	**0.9234**	**0.9883**	**0.9547**	**0.9946**	**0.9635**	**0.9788**
	StyleGAN + TradAug	γ = 6.55, B_{size} = 16	0.6454	0.9883	0.7809	0.9932	0.7587	0.8602
VGG16	StyleGAN	γ = 1.0, B_{size} = 32	0.6967	0.9961	0.8199	0.9979	0.8073	0.8925
	StyleGAN	γ = 6.55, B_{size} = 16	0.8929	0.9766	0.9328	0.9891	0.9479	0.9681
	StyleGAN + TradAug	γ = 1.0, B_{size} = 32	0.9237	0.8984	0.9109	0.9554	0.9670	0.9612
	StyleGAN + TradAug	γ = 6.55, B_{size} = 16	0.8992	0.9062	0.9027	0.9582	0.9549	0.9565

the StyleGAN (γ = 1.0, Batch = 32) dataset with an accuracy of 97.12% and a loss of 0.1014. Interestingly, for VGG16, the model without data augmentation performs better with StyleGAN (γ = 6.55, Batch = 16), with an accuracy of 95.67% and loss of 0.1027, indicating that combining StyleGAN with TradAug may not result in a synergistic enhancement of the model's performance. It suggests that while using GAN alone positively impacts the CNN model's overall performance, it may plateau after a certain point. The possible cause might be that the data is only as good as the underlying data they operate on [30]. However, when combined with the traditional data augmentation, as it is being done in our two stages of data augmentation pipeline, it is confirmed that the possibility of significantly improving the final performance exists. The reason might be attributed to the broader array of potential points offered by traditional augmentation, thereby diminishing the gap between training, validation, and test samples and subsequently mitigating the risk of overfitting [10].

Inception-V3 has resulted in the absence of data augmentation might have contributed to mitigating the risk of overfitting and preserving the pre-trained models' transferability. However, the situation appears to be the opposite for VGG-16. These findings suggest that the hyperparameter needs to be adjusted as it becomes more prevalent as more complex augmentation is introduced, as mentioned in the previous analysis. This anomaly can also be explained by the insufficient epochs to train the model, which might allow the model to adapt better to the new dataset. In addition, this experiment's findings also suggest that these models possess sufficient representation power, enabling effective transfer learning even with a limited labeled dataset. These insights can guide practitioners in selecting appropriate strategies for maximizing the performance of pre-trained models in specific classification tasks.

5. Conclusions

This research successfully employed two stages, StyleGAN and traditional augmentation to augment the training dataset for BCH image classification. It leverages StyleGAN

architecture to generate high-quality synthetic BCH images with the conventional augmentation method. The inception-V3 model with TradAug + StyleGAN (γ = 1.0, Batch = 32) yielded an accuracy of 97.12% and loss of 0.1014. It outperforms VGG16 with the same parameter, TradAug + StyleGAN (γ = 1.0, Batch = 32), yielding an accuracy of 94.59% and a loss of 0.1546. The research highlights the effectiveness of the two stages of data augmentation, GAN, and TradAug, in improving the accuracy of CNN classifiers in discerning BCH images. Nevertheless, it also emphasizes the importance of having more extensive and diverse datasets to enhance the model's generalizability. Additionally, comprehensive hyperparameter optimization is crucial to exploit the potential of GAN-based fine-tuning fully. Addressing these aspects holds promise for advancing AI-driven breast cancer classification in medical applications, ensuring more robust and reliable diagnostic tools for clinical practice. By bridging the gap between innovative AI techniques and medical image analysis, this research contributes to advancing breast cancer diagnosis.

References

[1] S. H. Gheshlaghi, C. N. E. Kan, and D. H. Ye, "Breast cancer histopathological image classification with adversarial image synthesis," 2021.

[2] A. Patil, D. Tamboli, S. Meena, D. Anand, and A. Sethi, "Breast cancer histopathology image classification and localization using multiple instance learning," 2019.

[3] D. Houfani, S. Slatnia, O. Kazar, I. Remadna, H. Saouli, G. Ortiz, and A. Merizig, "An improved model for breast cancer diagnosis by combining pca and logistic regression techniques," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, p. 701 – 716, 2023, cited by: 1; All Open Access, Gold Open Access. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160393174&doi=10.12785%2fijcds%2f130156&partnerID=40&md5=c25eba47d47704ccc6c7e1245c0be665

[4] M. M. A. Rahhal, "Breast cancer classification in histopathological images using convolutional neural network," *international journal of advanced computer science and applications*, vol. 9, 2018.

[5] S. S. M. Noh, N. Ibrahim, M. M. Mansor, and M. Yusoff, "Hybrid filtering methods for feature selection in high-dimensional cancer data," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 6, p. 6862 – 6871, 2023, cited by: 0; All Open Access, Gold Open Access. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85172876734&doi=10.11591%2fijece.v13i6.pp6862-6871&partnerID=40&md5=9c67f896d8b0d1dc507b1f030cb1cab3

[6] M. Yusoff, T. Haryanto, H. Suhartanto, W. A. Mustafa, J. M. Zain, and K. Kusmardi, "Accuracy analysis of deep learning methods in breast cancer classification: A structured review," *Diagnostics*, vol. 13, no. 4, 2023, cited by: 9; All Open Access, Gold Open Access, Green Open Access. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85148956575&doi=10.3390%2fdiagnostics13040683&partnerID=40&md5=da7c6f5689f7215c7d68266f8179b414

[7] M. Alzyoud, R. Alazaidah, M. Aljaidi, G. Samara, M. Qasem, M. Khalid, and N. Al-Shanableh, "Diagnosing diabetes mellitus using machine learning techniques," *International Journal of Data and Network Science*, vol. 8, pp. 179–188, 2024.

[8] X. Li, Y. Jiang, J. J. Rodriguez-Andina, H. Luo, S. Yin, and O. Kaynak, "When medical images meet generative adversarial network: recent development and research opportunities," *Discover Artificial Intelligence*, vol. 1, p. 5, 2021. [Online]. Available: https://doi.org/10.1007/s44163-021-00006-0

[9] L. Ruthotto and E. Haber, "An introduction to deep generative modeling," 2021. [Online]. Available: https://arxiv.org/abs/2103.05180

[10] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, p. 60, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[11] M. Ahmed and M. R. Islam, "Breast cancer classification from histopathological images using convolutional neural network." IEEE, 12 2021, pp. 1–4.

[12] A. AlSayyed, A. Taqateq, R. Al-Sayyed, D. Suleiman, S. Shukri, E. Alhenawi, and A. Albsheish, "Employing cnn ensemble models in classifying dental caries using oral photographs," *International Journal of Data and Network Science*, vol. 7, pp. 1535–1550, 2023.

[13] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[14] Y. Chen, X. H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan, "Generative adversarial networks in medical image augmentation: A review," *Computers in Biology and Medicine*, vol. 144, 2022.

[15] M. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, "Towards a better understanding of annotation tools for medical imaging: a survey," *Multimedia Tools and Applications*, vol. 81, pp. 25 877–25 911, 2022. [Online]. Available: https://doi.org/10.1007/s11042-022-12100-1

[16] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, 2016.

[17] S. Motamed, P. Rogalla, and F. Khalvati, "Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images," *Informatics in Medicine Unlocked*, vol. 27, p. 100779, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914821002501

[18] Razvan, D. A. V., B. A. K., B. Martin, R. N. S., G. P. H. Sungmin, and Marinescu, "3d-stylegan: A style-based generative adversarial network for generative modeling of three-dimensional medical images," Ilkay, Z. Dajiang, Y. Yixuan, M. Anirban, H. Nicholas, H. S. Xiaolei, N. Hien, S. Raphael, X. Y. E. Sandy, and Oksuz, Eds. Springer International Publishing, 2021, pp. 24–34.

[19] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019.

[20] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," 2021.

[21] A. Aksac, D. J. Demetrick, T. Ozyer, and R. Alhajj, "Brecahad:

a dataset for breast cancer histopathological annotation and diagnosis," *BMC Research Notes*, vol. 12, p. 82, 2019. [Online]. Available: https://doi.org/10.1186/s13104-019-4121-7

[22] J. Wang, L. Perez *et al.*, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, pp. 1–8, 2017.

[23] L. Ali, F. Alnajjar, H. Jassmi, M. Gochoo, W. Khan, and M. Serhani, "Performance evaluation of deep cnn-based crack detection and localization techniques for concrete structures," *Sensors*, vol. 21, p. 1688, 3 2021.

[24] J.-H. Lee and J.-H. Lee, "A reliable defect detection method for patterned wafer image using convolutional neural networks with the transfer learning," *IOP Conference Series: Materials Science and Engineering*, vol. 647, p. 12010, 3 2019.

[25] Y. Liu, Y. Zhou, X. Liu, F. Dong, C. Wang, and Z. Wang, "Wasserstein gan-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology," *Engineering*, vol. 5, pp. 156–163, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809918301127

[26] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," 2018.

[27] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, 3 2020.

[28] Y. Li, Y. Shen, H. Jiang, W. Zhang, J. Li, J. Liu, C. Zhang, and B. Cui, "Hyper-tune: towards efficient hyper-parameter tuning at scale," *Proc. VLDB Endow.*, vol. 15, pp. 1256–1265, 2 2022. [Online]. Available: https://doi.org/10.14778/3514061.3514071

[29] D. Gudovskiy, L. Rigazio, S. Ishizaka, K. Kozuka, and S. Tsukizawa, "Autodo: Robust autoaugment for biased data with label noise via scalable probabilistic implicit differentiation." IEEE Computer Society, 6 2021, pp. 16 596–16 605.

[30] T. Pinetz, J. Ruisz, and D. Soukup, "Actual impact of gan augmentation on cnn classification performance." *ICPRAM*, vol. 25, pp. 15–23, 2019.

Marina Yusoff is currently a deputy director and senior fellow researcher at the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Associate Professor of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia. She has a Ph.D. in Information Technology and Quantitative Sciences (Intelligent Systems). She previously worked as a Senior Executive of Information Technology at SIRIM Berhad, Malaysia. She is most interested in multidisciplinary research, artificial intelligence, nature-inspired computing optimization, and data analytics. She applied and modified AI methods in many research and projects, including recent hybrid deep learning, particle swarm optimization, genetic algorithm, ant colony, and cuckoo search for real-world problems and medical and industrial projects. Her recent projects are data analytic optimizer, audio, and image pattern recognition. She has many impact journal publications and contributes as an examiner and reviewer to many conferences, journals, and universities' academic activities.

Muhamad Shah Azizie Abd Karim is a passionate Machine Learning Developer, holding a degree in Intelligent System Engineering, with a drive for innovation, adept at exploring AI and data science frontiers. Experienced in programming and machine learning, skilled in automation and problem-solving. Committed to making a meaningful impact and advancing technological innovation.