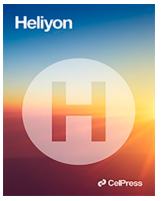
Heliyon 9 (2023) e19585

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



GSDA: Generative adversarial network-based semi-supervised data augmentation for ultrasound image classification

Zhaoshan Liu [a], Qiujie Lv [a,b], Chau Hung Lee [c], Lei Shen [a,*]

[a] Department of Mechanical Engineering, National University of Singapore, 9 Engineering Drive 1, Singapore, 117575, Singapore
[b] School of Intelligent Systems Engineering, Sun Yat-sen University, No. 66, Gongchang Road, Guangming District, 518107, China
[c] Department of Radiology, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore, 308433, Singapore

ARTICLE INFO

Keywords:
Semi-supervised learning
Generative adversarial network
Convolutional neural network
Medical image analysis

ABSTRACT

Medical Ultrasound (US) is one of the most widely used imaging modalities in clinical practice, but its usage presents unique challenges such as variable imaging quality. Deep Learning (DL) models can serve as advanced medical US image analysis tools, but their performance is greatly limited by the scarcity of large datasets. To solve the common data shortage, we develop GSDA, a Generative Adversarial Network (GAN)-based semi-supervised data augmentation method. GSDA consists of the GAN and Convolutional Neural Network (CNN). The GAN synthesizes and pseudo-labels high-resolution, high-quality US images, and both real and synthesized images are then leveraged to train the CNN. To address the training challenges of both GAN and CNN with limited data, we employ transfer learning techniques during their training. We also introduce a novel evaluation standard that balances classification accuracy with computational time. We evaluate our method on the BUSI dataset and GSDA outperforms existing state-of-the-art methods. With the high-resolution and high-quality images synthesized, GSDA achieves a 97.9% accuracy using merely 780 images. Given these promising results, we believe that GSDA holds potential as an auxiliary tool for medical US analysis.

1. Introduction

Medical Ultrasound (US) has become a widely utilized screening and diagnostic tool in clinical practice due to its absence of ionizing radiation, high sensitivity, portability, and relatively low cost [1]. However, there are limitations to be solved. Image quality is easily affected by noise and artifacts, inter-operator variability is considerable, and variability across different US systems is usually high. Due to these, diagnosing medical US images always heavily relies on radiologists. To address the problem, developing an advanced medical US image analysis tool to make medical US diagnosis more objective, accurate, and automatic is essential. In recent years, Deep Learning (DL) has emerged as a powerful tool to automate the extraction of useful information from big data. It has enabled ground-breaking advances in numerous computer vision tasks [2]. For the classification task, the Convolutional Neural Network (CNN) [3] is one of the most dominant methods. However, effectively training a CNN typically requires large datasets, which are often a significant obstacle in the medical field. For one thing, the acquisition of medical images typically necessitates the use of specialized equipment and requires medical experts for annotation. For another, datasets are usually confidential due to privacy concerns. In this

* Corresponding author.
E-mail addresses: e0575844@u.nus.edu (Z. Liu), lvqj5@mail2.sysu.edu.cn (Q. Lv), chau_hung_lee@ttsh.com.sg (C.H. Lee), mpeshel@nus.edu.sg (L. Shen).

https://doi.org/10.1016/j.heliyon.2023.e19585
Received 11 August 2023; Received in revised form 25 August 2023; Accepted 28 August 2023
Available online 4 September 2023
2405-8440/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

case, the Transfer Learning (TL) technique is widely implemented in CNN to relieve the common data shortage. The TL allows the models to be trained on the larger dataset first to relieve the difficulty of model training. However, given the unique characteristics of US images and the acute data shortage, relying solely on TL often fails to guarantee the model performance [4].

To further improve the model performance for US image classification, various Data Augmentation (DA) methods have been widely adopted. Traditional DA methods typically generate images based on a sequence of transformations such as rotation, flip, etc. While such methods are beneficial, they come with several challenges. Firstly, manually designing the type and sequence of transformations largely depends on experience and can often lead to suboptimal results. Secondly, the number of combinations is restricted when leveraging a small number of transformations. Although expanding the number of transformations can potentially address this, excessive transformations might produce meaningless augmented images that drift significantly from the original [5]. Due to this, several advanced DA methods are proposed [6,7] and the Generative Adversarial Network (GAN) [8] is one of the most widely implemented ones. The GAN is widely implemented for medical image synthesis and is composed of a generator (G) and a discriminator (D) playing an adversarial "game". During training, the G synthesizes images based on the data distribution it learned, and the D tries to discriminate whether the images are synthesized or not.

Several previous works [4,9,10] have utilized GAN for DA and classified medical images in a semi-supervised way. However, there are several points to be improved. For one thing, the synthesized images have either low resolution or quality. This can be caused by the basic GAN structure as well as the lack of the TL technique during training GAN. For another, the quality of the synthesized images is not evaluated quantitatively and the data distribution relationship across real and synthesized images is not investigated. Besides, performance across different CNN models is not fully searched. Till now, synthesizing high-resolution and high-quality US images, as well as training a high-performance classification model with a small dataset remain challenging. To solve existing problems, here we propose GSDA consisting of CNN and GAN. The GAN synthesizes and pseudo-labels the artificial US images with high resolution and high quality, whereas the CNN is trained using both real and synthesized images. To enhance image resolution and quality, we adopt the state-of-the-art GAN model SGA [11] and employ the TL technique during its training. To evaluate the synthesized images quantitatively, we implement widely accepted standards Inception Score (IS) [12] and Fre'chet Inception Distance (FID) [13]. We also implement t-SNE for analyzing the data distribution across real and synthesized images. To fully search the performance across different CNN models, we implement intensive experiments on several CNN models and compare the results. Moreover, we also propose a novel evaluation standard, the Training Efficiency Index (TEI), to balance the accuracy and the training time consumption. We evaluate our GSDA on the BUSI dataset [14], and the results show that with high-resolution and high-quality images synthesized, our GSDA can obtain a 97.9% accuracy using merely 780 images. To sum up, our main contributions are:

- We propose a GAN-based semi-supervised DA method GSDA to solve the common data shortage.
- We leverage state-of-the-art GAN to synthesize high-resolution and high-quality US images.
- We evaluate the synthesized image quantitatively and analyze the data distribution between real and synthesized images.
- We propose a novel evaluation standard to balance the classification accuracy and the time consumption.

The rest of this paper is organized as follows: In Section 2, we illustrate the related works of GAN as well as its application on semi-supervised medical image classification. The description of the datasets used, together with the methods proposed are discussed in Section 3. Section 4 shows the core experiment results, detailed analysis, and extensive ablation study. We conclude our work and point out the future perspective in Section 5.

2. Related work

GAN. Many variants of GAN [15–21] have been developed since it was initially proposed. In 2016, Radford et al. [19] developed a DCGAN model, in which the convolution operation is introduced into the GAN. In DCGAN, both the D and G are trained once during each epoch. One year later, the WGAN was proposed by Arjovsky et al. [22], which employs the Wasserstein distance into GAN and uses RMSprop as the optimizer instead of Adam. A variant of it, WGAN-GP, was later proposed [20]. The WGAN-GP adds a gradient penalty and applies layer norm [23] in D. However, training these GAN models always needs a large number of images. Besides, the resolution of synthesized images is relatively low. In 2020, Karras et al. [11] developed a novel SGA network with advanced architectural design, in which Adaptive Data Augmentation (ADA) is introduced in GAN to handle small data regimes. To effectively handle high-resolution images, both the G and D of the SGA are designed with a hierarchical structure.

GAN-based semi-supervised medical image classification. To overcome the common data shortage in the field of medical image classification, several works [4,9,10,24,25] have been proposed to use GAN for DA and classify the images in a semi-supervised way. The existing works can be divided into two approaches. The first is to train GAN solely and use the D of GAN as a classifier [9,24]. The second is to train GAN first and then use separate CNN as a separate classifier [4,10,25]. We opt for the latter approach, as it allows us to employ multiple CNN models and compare their performances. Compared with existing methods, our GSDA has several advantages. First, instead of employing basic GAN models, we implement state-of-the-art GAN model SGA to synthesize images with higher resolution and quality. We observed that most of the existing work using GAN does not introduce the TL technique thus hampering the model performance. We thus implement the TL technique rather than solely training from scratch. Second, besides qualitatively observing the synthesized quality, we employ IS and FID to evaluate the synthesized images quantitatively. We also visualize and analyze the data distribution across real and synthesized images. Third, we implement intensive experiments across different CNN models to search for higher performance. Finally, we propose a new evaluation standard TEI to balance the classification accuracy and the time consumption.

3. Materials and methods

3.1. Datasets

We use the BUSI dataset for training and several example figures accessed from it are illustrated in Fig. 1. The BUSI dataset is a breast cancer dataset collected among 600 female patients between 25 and 75 years old in 2018. The data is collected using the LOGIQ E9 US and LOGIQ E9 Agile US systems. The BUSI dataset contains 780 images and is divided into three subsets, including benign, malignant, and normal, as illustrated in Fig. 1a–c, respectively. Each subset corresponds to different breast cancer conditions. The benign, malignant, and normal subsets contain 437, 210, and 133 images, respectively. The average resolution of the images is around 500 × 500. Besides the BUSI dataset, four large datasets are selected as the source datasets of the TL technique. For synthesis, Flickr-Faces-HQ (FFHQ) [26], Large-Scale CelebFaces Attributes (CelebA) [27], and Large-scale Scene Understanding Challenge (LSUN) DOG [28] are leveraged. FFHQ is a face dataset with 70 K images at the resolution of 1024 × 1024. There is considerable variation in age, ethnicity, and image background among all images. CelebA is a large-scale face attributes dataset with 200 K celebrity face images collected from 10,177 identities. Large pose variations and background clutter are covered. LSUN DOG contains 5 M images of the category dog. For classification, ImageNet [29] is utilized. ImageNet, with its 14 M images, is organized according to the nouns of the WordNet hierarchy, with each node represented by numerous images.

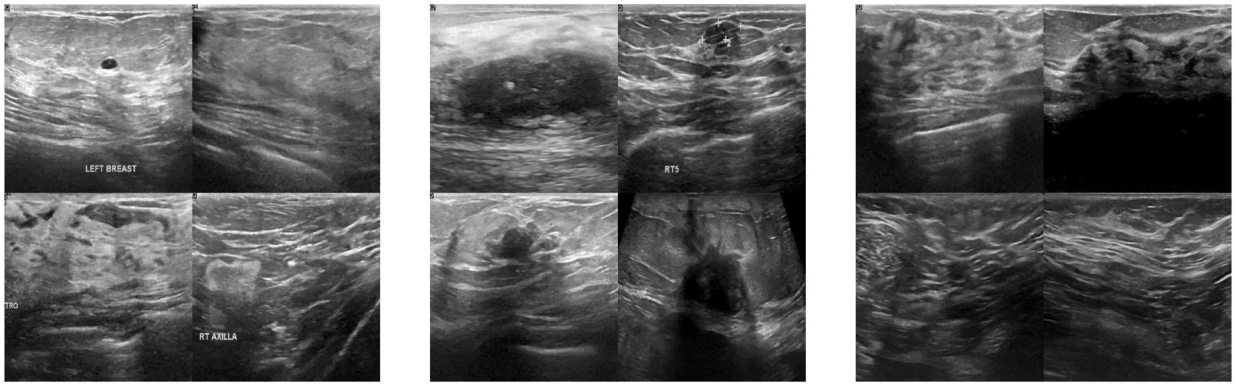
3.2. SGA

We leverage SGA to synthesize medical US images. The SGA features a $G-D$ architecture and includes an ADA block. Both G and D in the SGA follow a hierarchical structure, in which the resolution for G progresses from low to high and vice versa for D. The detailed structure of G and D of SGA can be found in Fig. 2b. The ADA is composed of eighteen transformations grouped into six groups, including pixel blitting, more general geometric transformations, color transforms, image-space filtering, additive noise [30], and cutout [31]. The set of transformations is employed in a fixed order with a strength $p \in [0,1]$. The p is adaptively controlled based on the degree of overfitting. The evaluation of overfitting is to utilize a separate validation set and observe its behavior with respect to the training set. Let us denote the outputs of D by $D_{\text{train}}, D_{\text{validation}}$, and $D_{\text{synthesized}}$, for the training set, validation set, and synthesized images, respectively, and their mean over N consecutive batches by $\mathbb{E}[\cdot]$, the overfitting can be computed using the below equation:

$$r_v = \frac{\mathbb{E}[D_{\text{train}}] - \mathbb{E}[D_{\text{validation}}]}{\mathbb{E}[D_{\text{train}}] - \mathbb{E}\left[D_{\text{synthesized}}\right]} r_t = \mathbb{E}[\text{sign}(D_{\text{train}})] \tag{1}$$

where $r = 0$ represents no overfitting and $r = 1$ shows completely overfitting. r_v shows the output for the validation set relative to the training set and synthesized images, and r_t estimates the portion of the training set with positive D outputs. For the adaptively controlled p, it is initialized to zero and adjusted once every four mini batches based on the Equation (1). In case the results indicate too much/little overfitting occurs, the p is adjusted by incrementing/decrementing a fixed amount.

The three subsets of the BUSI dataset are each used to train the SGA. Real images are preprocessed to a resolution of 256 × 256. The resolution of synthesized images is also set as 256 × 256 to balance the image quality and time consumption. The loss function implemented is the non-saturating logistic loss $f(x) = \log(\text{sigmoid}(x))$ [8]. The D loss is computed as $-f(x)$, where the G loss is computed using $-f(x)$ and $-f(-x)$. The optimizer is Adam and the learning rate is 0.0025. The number of iterations is 4000 with a batch size of 32. SGA is trained using four different settings, including training from scratch and using the TL technique with three different source datasets to demonstrate the impact of the TL technique.



(a) Benign (b) Malignant (c) Normal

Fig. 1. Example figures accessed from the BUSI dataset. (a) Benign, (b) malignant, and (c) normal.

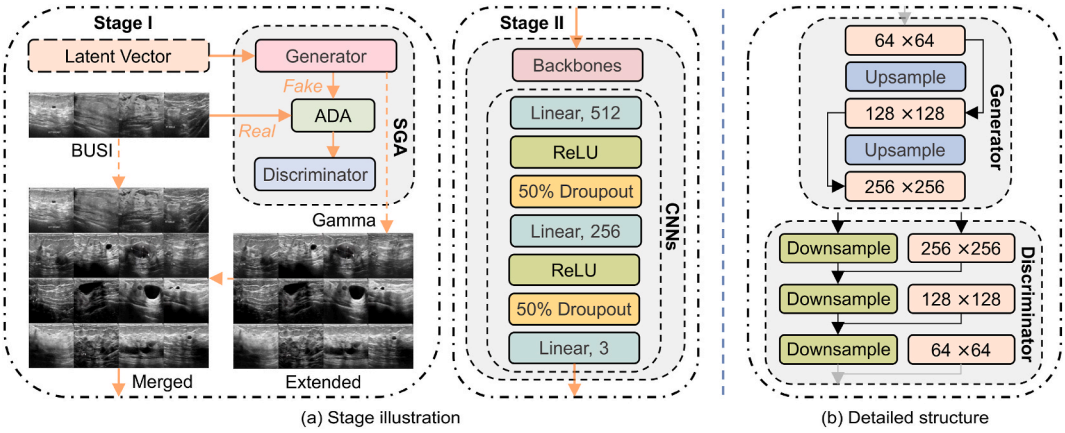


Fig. 2. The proposed GSDA is composed of two stages. In the first stage, SGA is trained using the BUSI dataset to capture the real image data distribution and synthesize artificial images. The synthesized extended datasets are then merged with the BUSI dataset to compose merged datasets. In the second stage, different CNN models are trained using the merged datasets. ADA stands for adaptive discriminator augmentation. Solid and dotted orange arrows show the training stream and data stream, respectively. Grey arrows point toward the omitted similar structures at different resolutions. (a) Stage illustration, and (b) detailed structure of G and D of SGA.

3.3. GSDA

As shown in Fig. 2a, our GSDA is composed of two stages. In the first stage, we train SGA using the BUSI dataset to capture the real image data distribution for image synthesis. In the second stage, we train different CNN models using the merged datasets. We construct seven CNN models with the unique classification head using VGGNet [32], ShuffleNet [33], ResNeXt [34], ResNet [35], MobileNet [36], InceptionNet [37], and DenseNet [38] as the backbone. The classification head consists of two linear layers with a ReLU activation function, two dropout layers, and a linear layer with three output nodes, which equals the number of subsets. It is worth noting that we denote the combination of the SGA and different CNN models used in different groups of experiments as different SGA-CNN pairs. This results in seven SGA-CNN pairs, which are SGA-VGG, SGA-Shuffle, SGA-ResNeXt, SGA-Res, SGA-Mobile, SGA-Inception, and SGA-Dense.

We utilize SGA to synthesize medical US images. We endow the synthesized images with pseudo-labels that are the same as the real images. Specifically, the images synthesized by the SGA trained with the benign subset are pseudo-labeled benign, and the same process is performed for the malignant subset and the normal subset. We use the synthesized images to compose the extended datasets. The size of each extended dataset equals the integer γ multiple of the BUSI dataset. To keep class balance, the proportion of the three subsets in the extended datasets is the same as that of the BUSI. For each SGA-CNN pair, the maximum value of γ is determined experimentally based on our proposed evaluation standard TEI. For the detailed algorithm on how the maximum value of γ is determined, see Section 3.4. We add the extended datasets to the BUSI dataset to compose the merged datasets.

We train pre-trained CNN models using the merged datasets. The merged datasets are divided randomly into a training set and a validation set with a ratio of 8:2. The resolution of images is preprocessed to 224×224 or 299×299 (InceptionNet) due to different model architecture designs. The loss function is cross-entropy. The optimizer used is Adam and the learning rate equals 0.003. The weight decay (WD) is set to 0.001 if applicable. The number of epochs is 60 and the batch size is 32. For a given γ, each CNN is trained under two groups of settings, depending on whether the WD is chosen or not. Traditional DA methods, including RandomResizedCrop and RandomHorizontalFlip, are implemented.

3.4. Evaluation standards

To evaluate the quality of images synthesized by the SGA, we use two ways of evaluation. The first is through qualitative observation where the overall quality and basic details are directly distinguishable. The second is quantitative assessment using IS and FID. The IS and FID are two prevalent evaluation standards for image synthesis and can be calculated via:

$$IS = \exp\left(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x) \parallel p(y))\right) \tag{2}$$

$$FID = \parallel m - m_w \parallel_2^2 + Tr\left(C + C_w - 2(CC_w)^{\frac{1}{2}}\right) \tag{3}$$

where $x \sim p_g$ represents sample x from p_g, D_{KL} represents the KL divergence. w represents the real-world data, m denotes the mean value, C shows the covariance matrix, and Tr represents the trace. A lower IS indicates worse model performance, whereas a lower FID indicates better performance. From Equation (2) and Equation (3), it is observed that the real images are not taken into consideration when calculating IS. In such scenarios, the model might achieve a high IS by simply replicating the real images. To make the evaluation

more convincing, we regard FID as the main standard and take IS as a reference. Both FID and IS are calculated every 200 iterations.

For the SGA-CNN pair, the training time t_γ increases significantly with γ. We thus evaluate the model performance in two-fold. First, we employ classification accuracy acc_γ for evaluation. This is the most distinct and effective evaluation standard. Second, we propose a new standard TEI to balance the classification accuracy and the time consumption. Given γ, TEI can be calculated using:

$$TEI = \begin{cases} \ln\left(t_\gamma - t_{\gamma=0}\right)^{-1} \cdot \left(acc_\gamma - acc_{\gamma=0}\right) & \gamma > 1, \\ 0 & \gamma = 0. \end{cases} \tag{4}$$

where $t_{\gamma=0}$, $acc_{\gamma=0}$ denotes the training time and accuracy when training using the BUSI dataset. From Equation (4), we can find that the TEI indicates the ability of the model to attain improved accuracy within a limited training time. The higher the TEI, the better the ability. It is worth noting that for each SGA-CNN pair, we determine the maximum value of γ experimentally based on the proposed TEI. The specific procedure is (1) Initialize γ as 1, (2) calculate TEI, (3) increase γ by 1, (4) calculate new TEI, (5) compare new TEI with the previous one, (6) if TEI increases, repeat (3), (4), and (5) until TEI stops increasing. The pseudo-code of the proposed algorithm is illustrated in Algorithm 1.

Algorithm 1
Determination of the maximum value of γ for SGA-CNN pairs

Require: Extended multiple $\gamma \in [0, n]$, GAN model SGA, i_{th} CNN model $CNN_i, i \in [1, 7]$, dataset $D = (X_{image}, Y)$, time t;
Ensure: $TEI_{max}|\gamma$;
$TEI_{\gamma=0} = 0$;
for all i **do**
  Initialize $\gamma = 1$;
  while $TEI_{\gamma,i} > TEI_{\gamma-1,i}$ **do**
    $t_{start} = start\ time$;
    $X^{SGA}_{image} = SGA(X_{image})$;
    $y = CNN_i(X^{SGA}_{image} \bigcup X_{image})$;
    $acc_{\gamma,i} = accuracy(y, Y)$;
    $t_{end} = end\ time$;
    $t_{\gamma,i} = t_{end} - t_{start}$;
    Compute $TEI_{\gamma,i} = \ln\left(t_{\gamma,i} - t_{\gamma-1,i}\right)^{-1} \cdot (acc_{\gamma,i} - acc_{\gamma-1,i})$;
    $\gamma = \gamma + 1$;
  end while
  Out γ when $TEI_{max}|\gamma$;
end for

4. Results and analysis

4.1. Unsupervised synthesis

Several medical US images synthesized by SGA are shown in Figs. 3 and 4a, for using the TL technique and training from scratch, respectively. We also show images synthesized by DCGAN, WGAN, and WGAN-GP in Fig. 4b–d for comparison. The synthesized images from SGA exhibit noticeably higher quality compared to those from other GAN models. The SGA effectively mimics the medical annotations (white in figures) in the BUSI dataset, while the commonly used DCGAN, WGAN, and WGAN-GP cannot even handle the task

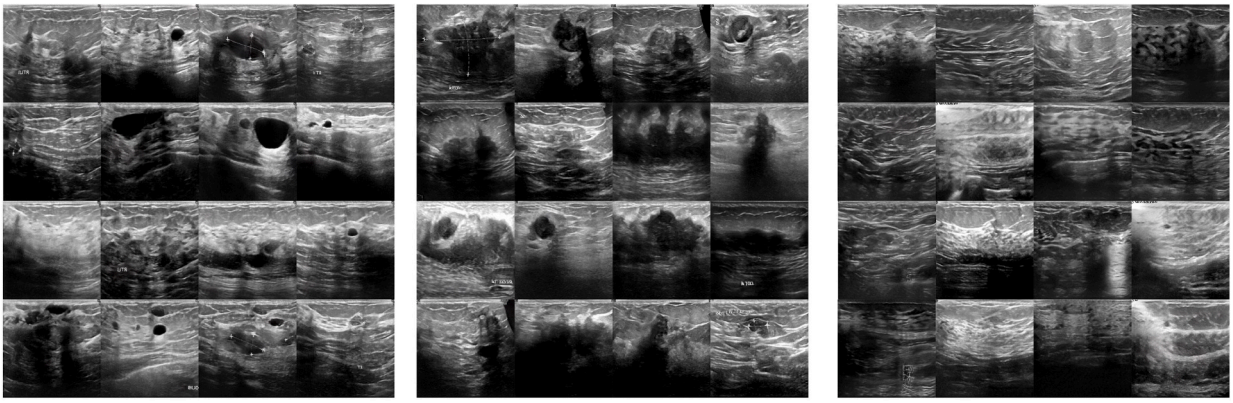


Fig. 3. Images synthesized by SGA with the TL technique. (a) Benign, (b) malignant, and (c) normal.

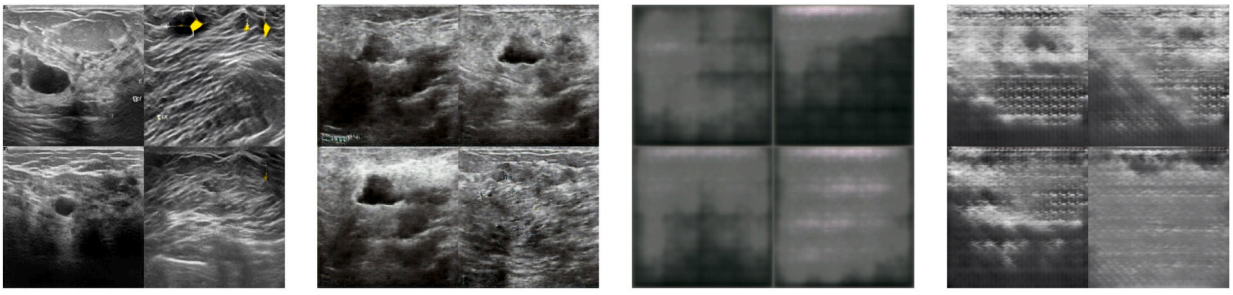


Fig. 4. Normal images synthesized by different GAN models. Models are trained from scratch. (a) SGA, (b) DCGAN, (c) WGAN, and (d) WGAN-GP.

well at this resolution. Besides, the quality of the synthesized images significantly improves with the introduction of the TL technique. On a qualitative note, Fig. 3a–c do not exhibit the evident yellow flaws that are present in Fig. 4a. Quantitatively, better FID and IS are observed, as shown in Fig. 5a and b, respectively. From the observation, we can find that the TL technique not only enhances performance but also improves stability. As the TL technique can improve the performance of SGA essentially, we set the TL technique as the default when developing SGA-CNN pairs. It is worth noting that the TL source dataset used here is FFHQ. For the comparison of FID and IS across different TL experimental groups, see the corresponding ablation study in Section 4.3.

To prove the effectiveness of the proposed image synthesis method, we employ t-SNE to visualize the data distribution across real and synthesized images in Fig. 6. The visualization is performed using SGA-VGG without decay as it outperforms other combinations. Features are extracted prior to the classification head, and each category comprises a hundred randomly sampled images. The results demonstrate that the distributions of both real and synthesized images are closely aligned, and a nearly overlapping distribution attests to the effectiveness of the proposed synthesis method. Notably, several outliers are observed in both synthesized benign and malignant categories. This can be caused by either CNN prediction or SGA synthesize deviation. Nevertheless, the number of such outliers is limited and thus does not influence the overall results. In Fig. 7, we illustrate how the TL technique aids the model training and the images synthesized during the process. When the TL technique is employed, as evident in Fig. 7a, the *G* of SGA inherits the weights learned from the FFHQ dataset. With pre-learned weights, the *G* can learn the distribution of the BUSI dataset quickly. By the 32nd iteration, the *G* can already synthesize the BUSI-like images. The lowest FID is reached at the 1000th iteration. However, in the case of lacking the TL technique, the weights are initialized randomly at the beginning of the training, as illustrated in Fig. 7b. The *G* starts to learn some representations at around 64 to 128 iterations, and the lowest FID is reached at the 3200th iteration. This indicates a substantially longer training time consumption compared to scenarios utilizing the TL technique. Worse still, even with 3200 iterations, the *G* trained from scratch exhibits severe mode collapse. In other words, its diversity is significantly lower compared to that achieved with TL from FFHQ.

4.2. *Semi-supervised classification*

The classification accuracy of different SGA-CNN pairs is shown in Table 1. We find that the SGA-VGG pair without WD achieves the highest accuracy at 97.9%. For the SGA-VGG pair with WD, we obtain an accuracy of 97.3%. While the SGA-VGG pairs achieve the

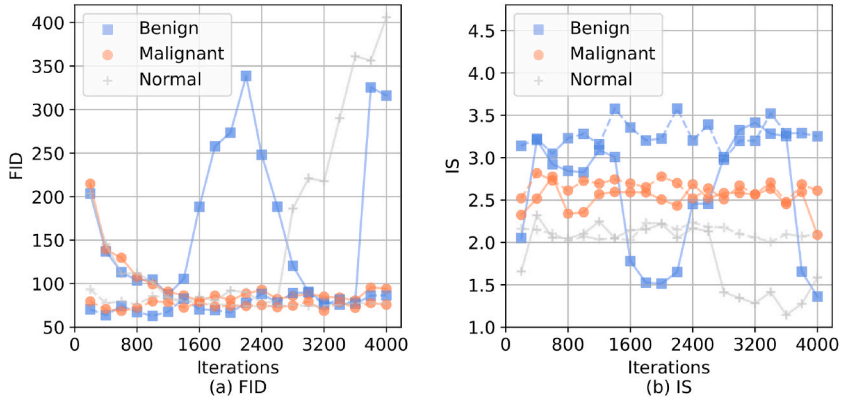


Fig. 5. FID and IS recording during training. The solid lines stand TL from FFHQ, while dotted lines illustrate training from scratch. (a) FID, and (b) IS.

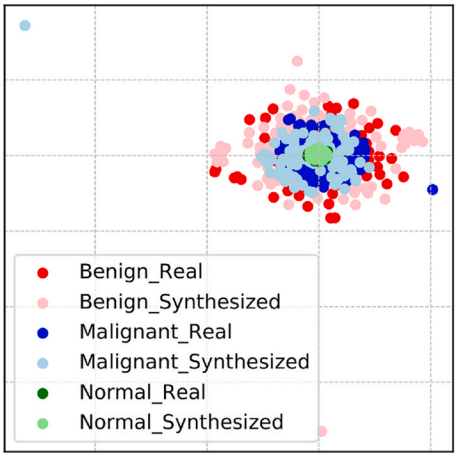


Fig. 6. Data contribution visualization using t-SNE across real and synthesized images.

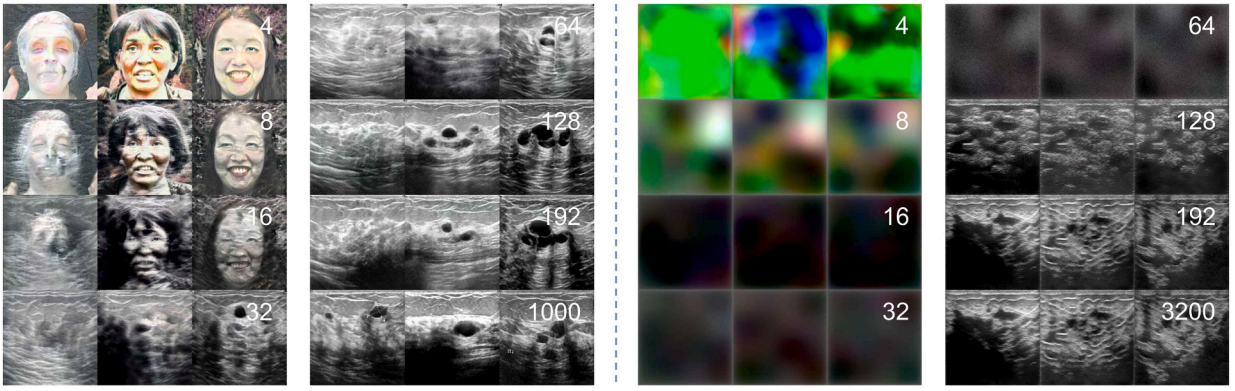


Fig. 7. Images synthesized by SGA at different stages during training. Only for illustration. Numbers represent the number of iterations trained. (a): with TL, transfer from FFHQ, and (b): without TL, training from scratch.

highest accuracy, several pairs demonstrate a greater improvement in accuracy with limited time consumption, reaching higher TEI. For instance, the SGA-Dense pair with WD reaches a TEI of 3.06, and the SGA-Mobile pair without WD obtains a TEI of 3.04. The intensive experiment results indicate that our method is universally suitable for all CNN models without any selection bias. We use the SGA-VGG pair without WD when comparing the performance with existing methods. In Table 2, we compare our GSDA with the state-of-the-art methods using the same dataset. We divide the existing methods into two categories, depending on whether the method is semi-supervised or not. From the table, we can find that the proposed GSDA reaches the highest accuracy and overperforms existing methods, even for comparison with binary classification and training with extra data. This demonstrates the effectiveness of GSDA, establishing it as a new state-of-the-art milestone.

To provide guidance for practical applications, we plot the comparison of the accuracy-time curve across different SGA-CNN pairs in Fig. 8a and b, with and without the TL technique, respectively. It is worth noting that though the maximum value of γ varies among different SGA-CNN pairs, we conduct additional experiments and illustrate the results based on the maximum γ across all SGA-CNN pairs, which is eight as illustrated in Table 1. In other words, each SGA-CNN pair has eight groups of experiments for both WD settings, respectively. This standardizes the results, making them more amenable to comparison. The results reveal that the SGA-VGG pair achieves the highest accuracy with the least time consumption, irrespective of the presence of WD. This illustrates that the SGA-VGG pair should be priority considered when deploying the classification task in practice. It is worth noting that the training of several pairs can become unstable without the WD, indicated in points A, B, and C in Fig. 8b. This suggests that WD contributes to stabilizing the training to a certain degree.

4.3. Ablation studies

The structure ablation study is implemented by comparing the classification performance in the case of whether the SGA is implemented or not. The performance of the CNN models without SGA is detailed in Table 3. A comparison between Tables 1 and 3

Table 1
Classification accuracy across SGA-CNN pairs. *macc* shows the maximum accuracy across all γ. ↑ means the higher, the better, and ↓ inverse. Bold numbers show the best results.

WD	Standard	SGA-VGG	SGA-Shuffle	SGA-ResNeXt	SGA-Res	SGA-Mobile	SGA-Inception	SGA-Dense
✓	acc_γ ↑	97.2%	91.6%	84.5%	85.4%	94.9%	81.7%	90.0%
✓	*macc* ↑	**97.3%**	91.6%	84.5%	85.4%	94.9%	82.7%	90.4%
✓	t_γ ↓	1480.8s	1458.2s	1467.5s	1457.6s	1460.3s	1385.6s	1263.7s
✓	γ	7	7	7	7	7	4	5
✓	TEI_γ ↑	2.18	2.75	2.55	2.14	2.86	2.28	**3.06**
×	acc_γ ↑	97.8%	95.0%	81.2%	86.6%	94.9%	82.4%	88.8%
×	*macc* ↑	**97.9%**	95.0%	81.2%	86.6%	95.0%	82.4%	89.1%
×	t_γ ↓	1478.4s	1456.5s	1251.9s	1455.5s	1460.3s	1175.0s	1260.1s
×	γ	7	7	5	7	7	3	5
×	TEI_γ ↑	2.18	2.96	1.69	2.14	**3.04**	2.46	2.06

Table 2
Performance comparison on the BUSI dataset between GSDA and state-of-the-art methods. * shows binary classification. ** stands including additional training data. SSL presents whether the methods belong to semi-supervised or not.

Ref.	Year	SSL	Methods	*macc* ↑
[39]	2021	×	Multi-CNN Hybrid Structure	95.6%
[40]	2021	×	ResNet	88.9%
[41]	2021	×	ResNet + Binary Grey Wolf Optimization + Support Vector Machine	84.9%
[42]	2022	×	YOLO	95.3%
[43]	2022	×	CNN + Genetic Algorithm **	92.8%
[44]	2023	×	ShuffleNet-ResNet *	95.1%
[45]	2023	×	Interpretable Multitask Information Bottleneck Network *	93.0%
[46]	2023	×	Consistent Ordinal Representations	82.2%
[47]	2023	×	Multi-Task Learning + Attention *	91.0%
[48]	2021	×	Vision Transformer	74.0%
[49]	2020	×	CNN Ensemble Learning *	90.8%
[50]	2020	×	Hybrid Feature Set + Ensemble Classifier *	96.6%
[51]	2021	×	Machine Learning-Radiomics *	97.4%
[52]	2021	×	Deep Representations Scaling *	92.3%
[4]	2019	✓	CNN + DAGAN	94.0%
[53]	2021	✓	ResNet + DK-Guided Data Augmentation	81.1%
[54]	2022	✓	ResNet + Convolutional Autoencoder *	88.2%
[55]	2022	✓	Consistency Training + Vision Transformer + Adaptive Token Sampler	95.3%
Ours	–	✓	GSDA	**97.9%**

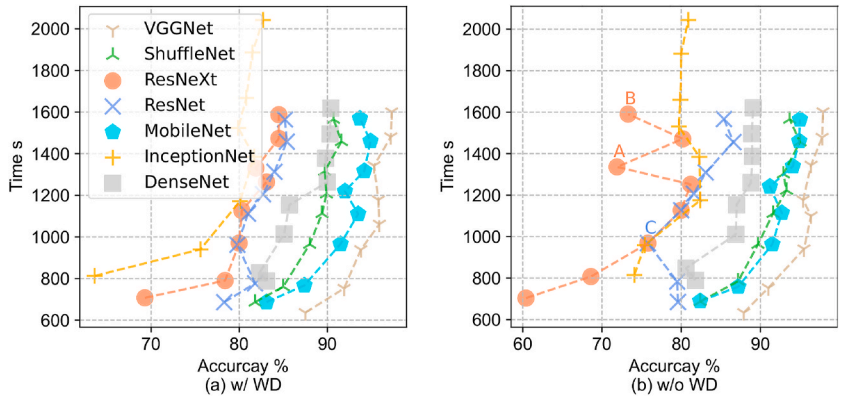


Fig. 8. Accuracy-time curve across different SGA-CNN pairs. The closer to the lower right corner, the better the overall performance. For each pair, the order of scatters corresponds to the increase of γ. Two subplots share the legend. (a) With WD, and (b) without WD.

reveals a significant drop in performance without SGA. For instance, the SGA-VGG pair shows a 15.6% decrease in accuracy regardless of whether the WD is implemented. Given the limited size of the dataset, these results align with our expectations and underscore the effectiveness of SGA. Regarding the SGA-Inception pairs, a tremendous accuracy drop of 16.9% and 16.6% are observed in scenarios with and without WD, respectively.

The dataset ablation study is conducted by comparing FID and IS across various TL experimental groups. From Table 4, it is found that TL from FFHQ performs best compared with TL from CelebA and TL from LSUN DOG, getting an FID of 62.92, 68.78, and 73.92 for

Table 3
Classification accuracy across different CNN models.

WD	Standard	VGGNet	ShuffleNet	ResNeXt	ResNet	MobileNet	InceptionNet	DenseNet
✓	$acc_{\gamma=0}$ ↑	**81.7%**	72.2%	66.5%	70.3%	74.7%	65.8%	69.0%
✓	$t_{\gamma=0}$ ↓	**262.7s**	290.8s	300.6s	291.0s	291.8s	327.4s	305.8s
×	$acc_{\gamma=0}$ ↑	**82.3%**	74.1%	69.6%	71.5%	73.4%	65.8%	74.7%
×	$t_{\gamma=0}$ ↓	**260.5s**	290.0s	300.3s	291.8s	293.2s	327.3s	306.4s

Table 4
Comparison of FID and IS across different TL experimental groups. In each group, the listed FID and IS are the optimal results calculated in the corresponding number of iterations.

Group	Subset	TL	FID ↓	Iterations	IS ↑	Iterations
1	Benign	CelebA	69.24	200	**3.58**	2200
2	Benign	LSUN DOG	102.95	600	2.92	4000
3	Benign	FFHQ	**62.92**	1000	**3.58**	2200
4	Malignant	CelebA	72.55	400	**2.84**	2000
5	Malignant	LSUN DOG	89.68	600	2.25	600
6	Malignant	FFHQ	**68.78**	3200	2.82	400
7	Normal	CelebA	79.69	600	2.19	1200
8	Normal	LSUN DOG	91.63	200	2.01	1400
9	Normal	FFHQ	**73.92**	2200	**2.24**	2400

three subsets respectively. However, for the malignant subset, TL from FFHQ obtains a lower IS compared with TL from CelebA. This conflicting outcome highlights some limitations of IS. It is worth noting that despite its higher diversity, LSUN DOG performs the worst in our experiments. This observation contrasts with the conclusion that the success of the TL technique likely hinges more on dataset diversity than on the similarity between subjects [11]. We speculate that this conclusion might be influenced by the close relationship between dogs and cats.

5. Conclusions

We introduced the GSDA, a novel method aimed at enhancing the classification accuracy of medical US images under small data limits. Experimental results on the BUSI dataset underscore the effectiveness and robustness of GSDA in image classification. Given its commendable performance, GSDA has the promising potential to serve as a supplementary diagnostic instrument. However, it is imperative to acknowledge certain limitations. The SGA is trained independently on distinct subsets to mitigate mutual interference for performance consideration. However, when there are numerous subsets, this approach may become impractical due to computational resource constraints. In such scenarios, the SGA can be conditionally trained by feeding class labels alongside the images. Using the trained SGA, images from various subsets can then be synthesized. Furthermore, a potential challenge of GSDA is the complexity introduced by separately training the SGA and CNN. To mitigate this, the two stages can be trained synchronously, leveraging the D of SGA for classification. While this training method allows for synchronous training of SGA and CNN, it poses challenges when comparing performances across diverse CNN models. Integrating CNN into SGA demands significant computational resources, given that the computational cost of SGA surpasses that of CNN by multiple orders of magnitude. The avenues for future research can be categorized into three primary domains. Firstly, while the GSDA is designed for 2D medical image classification, there is potential to extend the method to image segmentation and 3D imaging. Secondly, the GSDA presently sets the size of the extended dataset through comprehensive experimentation. Exploring more efficient methods to determine this size could curtail computational costs. Lastly, in light of the rapid advancements in the vision transformer [56], integrating CNN and the vision transformer appears promising. Such integrations can potentially enhance model performance by effectively capturing both local and global features, as discussed in [57].

Author contribution statement

Zhaoshan Liu: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Qiujie Lv: Analyzed and interpreted the data; Wrote the paper.

Chau Hung Lee: Analyzed and interpreted the data.

Lei Shen: Conceived and designed the experiments. Wrote the paper.

Funding statement

This work was supported by Tan Tock Seng Hospital (Grant number: A-8001334-00-00).

Data availability statement

Data will be made available on request.

Additional information

No additional information is available for this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Aimen Zlitni, Sanjiv S. Gambhir, Molecular imaging agents for ultrasound, Curr. Opin. Chem. Biol. 45 (2018) 113–120, 10.1016/j. cbpa.2018.03.017.
[2] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, Deep learning for computer vision: a brief review, Comput. Intell. Neurosci. 2018 (2018) 1–13, https://doi.org/10.1155/2018/7068349.
[3] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE (1998) 2278–2324, https://doi.org/10.1109/5.726791.
[4] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, Fahmy Aly, Deep learning approaches for data augmentation and classification of breast masses using ultrasound images, Int. J. Adv. Comput. Sci. Appl. 10 (1–11) (2019), https://doi.org/10.14569/IJACSA.2019.0100579.
[5] Dan Hendrycks, Norman Mu, D. Ekin, Cubuk, Barret Zoph, Justin Gilmer, Balaji Lakshminarayanan, Augmix: A Simple Data Processing Method to Improve Robustness and Uncertainty, arXiv preprint, 2019, https://doi.org/10.48550/arXiv.1912.02781.
[6] Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, Furao Shen, Image Data Augmentation for Deep Learning: A Survey, arXiv preprint, 2022, https://doi.org/10.48550/arXiv.2204.08610.
[7] Zhaoshan Liu, Qiujie Lv, Yifan Li, Ziduo Yang, Lei Shen, Medaugment: universal automatic data augmentation plug-in for medical image analysis, arXiv preprint arXiv:2306.17466, 2023.
[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde- Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[9] Ali Madani, Mehdi Moradi, Alexandros Karargyris, Tanveer Syeda-Mahmood, Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation, in: Proceedings of the IEEE International Symposium on Biomedical Imaging, 2018, pp. 1038–1042, https://doi.org/10.1109/ISBI.2018.8363749.
[10] Ting Pang, Jeannie Hsiu Ding Wong, Wei Lin Ng, Chee Seng Chan, Semi-supervised gan-based radiomics model for data augmentation in breast ultrasound mass classification, Comput. Methods Progr. Biomed. 203 (2021), 106018, https://doi.org/10.1016/j.cmpb.2021.106018.
[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, Timo Aila, Training generative adversarial networks with limited data, in: Proceedings of the Conference on Neural Information Processing Systems, 2020, pp. 1–37.
[12] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, Improved techniques for training gans, in: Proceedings of the Conference on Neural Information Processing Systems, 2016. **1–10**.
[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Proceedings of the Conference on Neural Information Processing Systems, 2017, pp. 1–38.
[14] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, Fahmy Aly, Dataset of breast ultrasound images, Data Brief 28 (2020), 104863, https://doi.org/10.1016/j.dib.2019.104863.
[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[16] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[17] Junho Kim, Minjae Kim, Hyeonwoo Kang, Kwanghee Lee, U-gat-it: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-To-Image Translation, arXiv Preprint, 2019, https://doi.org/10.48550/arXiv.1907.10830.
[18] Mehdi Mirza, Simon Osindero, Conditional Generative Adversarial Nets, arXiv preprint, 2014.
[19] Alec Radford, Luke Metz, Soumith Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, arXiv preprint, 2016.
[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville, Improved training of wasserstein gans, in: Proceedings of the Conference on Neural Information Processing Systems, 2017, pp. 5769–5779.
[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo 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.
[22] Martin Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein generative adversarial networks, Proc. Int. Conf. Mach. Learn. 70 (2017) 214–223.
[23] Jimmy Lei Ba, Jamie Ryan Kiros, E. Geoffrey, Hinton. Layer Normalization, arXiv preprint, 2016.
[24] Ibrar Amin, Saima Hassan, Jafreezal Jaafar, Semi-supervised learning for limited medical data using generative adversarial network and transfer learning, in: Proceedings of the International Conference on Computational Intelligence, IEEE, 2020, pp. 5–10, https://doi.org/10.1109/ICCI51257.2020.9247724.
[25] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, Hayit Greenspan, Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification, Neurocomputing 321 (2018) 321–331, https://doi.org/10.1016/j.neucom.2018.09.013.
[26] Tero Karras, Samuli Laine, Timo 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.
[27] Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, Progressive Growing of Gans for Improved Quality, Stability, and Variation, arXiv preprint, 2017, https://doi.org/10.48550/arXiv.1710.10196.
[28] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, Jianxiong Xiao, Lsun: Construction of a Large-Scale Image Dataset Using Deep Learning with Humans in the Loop, arXiv preprint, 2015, https://doi.org/10.48550/arXiv.1506.03365.
[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[30] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, Ferenc Huszaŕ, Amortised Map Inference for Image Super-resolution, arXiv preprint, 2016, https://doi.org/10.48550/arXiv.1610.04490.
[31] Terrance DeVries, Graham W. Taylor, Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv Preprint*, 2017, https://doi.org/10.48550/arXiv.1708.04552.

[32] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015, pp. 1–14.
[33] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun, Shufflenet: an extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856, https://doi.org/10.1109/CVPR.2018.00716.
[34] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, Kaiming He, Aggregated residual transformations for deep neural networks, 7, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500, https://doi.org/10.1109/CVPR.2017.634.
[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[36] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam, Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324, https://doi.org/10.1109/ICCV.2019.00140.
[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826, https://doi.org/10.1109/CVPR.2016.308.
[38] Gao Huang, Zhuang Liu, Laurens van der Maaten, Q. Kilian, Weinberger. Densely connected convolutional networks, 7, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708, https://doi.org/10.1109/cvpr.2017.243.
[39] Yeşim Eroğlu, Muhammed Yildirim, Ahmet Çinar, Convolutional neural networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mrmr, Comput. Biol. Med. 133 (2021), 104407, https://doi.org/10.1016/j.compbiomed.2021.104407.
[40] Arijit Das, Srinibas Rana, Exploring residual networks for breast cancer detection from ultrasound images, in: Proceedings of the International Conference on Computing, Communication and Networking Technologies, 2021, pp. 1–6, https://doi.org/10.1109/ICCCNT51525.2021.9580160.
[41] Priyanka Khanna, Mridu Sahu, Bikesh Kumar Singh, Improving the classification performance of breast ultrasound image using deep learning and optimization algorithm, in: Proceedings of the IEEE International Conference on Technology, Research, and Innovation for Betterment of Society, IEEE, 2021, pp. 1–6, https://doi.org/10.1109/TRIBES52498.2021.9751677.
[42] Rakesh Chandra Joshi, Divyanshu Singh, Vaibhav Tiwari, Malay Kishore Dutta, An efficient deep neural network based abnormality detection and multi-class breast tumor classification, Multimed. Tool. Appl. 81 (10) (2022) 13691–13711, https://doi.org/10.1007/s11042-021-11240-0.
[43] Hossam Magdy Balaha, Mohamed Saif, Ahmed Tamer, Ehab H. Abdelhay, Hybrid deep learning and genetic algorithms approach (hmb-dlgaha) for the early ultrasound diagnoses of breast cancer, Neural Comput. Appl. 34 (11) (2022) 8671–8695, https://doi.org/10.1007/s00521-021-06851-5.
[44] Adyasha Sahu, Pradeep Kumar Das, Sukadev Meher, High accuracy hybrid cnn classifiers for breast cancer detection using mammogram and ultrasound datasets, Biomed. Signal Process Control 80 (2023), 104292, https://doi.org/10.1016/j.bspc.2022.104292.
[45] Junxia Wang, Yuanjie Zheng, Jun Ma, Xinmeng Li, Chongjing Wang, James Gee, Haipeng Wang, Wenhui Huang, Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation, Med. Image Anal. 83 (2023), 102687.
[46] Yiming Lei, Zilong Li, Yangyang Li, Junping Zhang, Hongming Shan, Core: Learning Consistent Ordinal Representations for Image Ordinal Estimation, arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2301.06122.
[47] Meng Xu, Kuan Huang, Xiaojun Qi, A regional-attentive multi-task learning framework for breast ultrasound image segmentation and classification, IEEE Access (2023), https://doi.org/10.1109/ACCESS.2023.3236693.
[48] Behnaz Gheflati, Hassan Rivaz, Vision Transformer for Classification of Breast Ultrasound Images, arXiv preprint, 2021.
[49] Woo Kyung Moon, Yan-Wei Lee, Hao-Hsiang Ke, Su Hyun Lee, Chiun-Sheng Huang, Ruey-Feng Chang, Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks, Comput. Methods Progr. Biomed. 190 (2020), 105361, https://doi.org/10.1016/j.cmpb.2020.105361.
[50] Tariq Sadad, Ayyaz Hussain, Asim Munir, Muhammad Habib, Sajid Ali Khan, Shariq Hussain, Shunkun Yang, Mohammed Alawairdhi, Identification of breast malignancy by marker-controlled watershed transformation and hybrid feature set for healthcare, Appl. Sci. 10 (6) (2020) 1900, https://doi.org/10.3390/app10061900.
[51] Arnab K. Mishra, Pinki Roy, Sivaji Bandyopadhyay, Sujit K. Das, Breast ultrasound tumour classification: a machine learning—radiomics based approach, Expet Syst. 38 (7) (2021), e12713, https://doi.org/10.1111/exsy.12713.
[52] Michal Byra, Breast mass classification with transfer learning based on scaling of deep representations, Biomed. Signal Process Control 69 (2021), 102828, https://doi.org/10.1016/j.bspc.2021.102828.
[53] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Qingfeng Li, Yong Wang, Shaojie Tang, Dk-consistency: a domain knowledge guided consistency regularization method for semi-supervised breast cancer diagnosis, Proc. Int. Conf. Bioinf. Biomed. (2021) 3435–3442, https://doi.org/10.1109/BIBM52615.2021.9669494.
[54] Mingue Song, Yanggon Kim, Deep representation for the classification of ultrasound breast tumors, in: Proceedings of the International Conference on Ubiquitous Information Management and Communication, 2022, pp. 1–6, https://doi.org/10.1109/IMCOM53663.2022.9721796.
[55] Wei Wang, Ran Jiang, Ning Cui, Qian Li, Feng Yuan, Zhifeng Xiao, Semi-supervised vision transformer with adaptive token sampling for breast cancer classification, Front. Pharmacol. 13 (2022), 929755, https://doi.org/10.3389/fphar.2022.929755.
[56] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv Preprint, 2020, https://doi.org/10.48550/arXiv.2010.11929.
[57] Zhaoshan Liu, Qiujie Lv, Ziduo Yang, Yifan Li, Chau Hung Lee, Lei Shen, Recent progress in transformer-based medical image analysis, Comput. Biol. Med. (2023), 107268.