Medical Image Synthesis with Generative Adversarial Networks for Tissue Recognition

1st Qianqian Zhang
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, United States
qzhang40@binghamton.edu

2nd Haifeng Wang
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, United States
hwang58@binghamton.edu

3rd Hongya Lu
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, United States
hlu26@binghamton.edu

4th Daehan Won
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, United States
dhwon@binghamton.edu

5th Sang Won Yoon
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, United States
yoons@binghamton.edu

***Abstract*—This paper presents an adversarial learning-based approach to synthesize medical images for medical image tissue recognition. The performance of medical image recognition models highly depends on the representativeness and sufficiency of training samples. The high expense of collecting large amounts of practical medical images leads to a demand of synthesizing image samples. In this research, generative adversarial networks (GANs), which consist of a generative network and a discriminative network, are applied to develop a medical image synthesis model. Specifically, deep convolutional GANs (DCGANs), Wasserstein GANs (WGANs), and boundary equilibrium GANs (BEGANs) are implemented and compared to synthesize medical images in this research. Convolutional neural networks (CNNs) are applied in the GAN models, which can capture feature representations that describe a high level of image semantic information. Then synthetic images are generated by employing the generative network mapping from random noise. The effectiveness of the generative network is validated by a discriminative network, which is trained to detect the synthetic images from real images. Through a minimax two-player game, the generative and discriminative networks can train each other. The generated synthetic images are used to train a CNN classification model for tissue recognition. Through the experiments with the synthetic images, the tissue recognition accuracy achieves 98.83%, which reveals the effectiveness and applicability of synthesizing medical images through the GAN models.**

***Keywords*-generative adversarial networks; image synthesis; Wasserstein distance; tissue recognition**

I. Introduction

Tissue recognition is one of the most fundamental procedures for many medical imaging applications, especially for computer-aided diagnosis (CAD), which provides a second opinion by computer and assists the interpretation and diagnosis of diseases for radiologists [1]. Multiple medical images collected by various medical devices, such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), ultrasonography (US) and optical coherence tomography (OCT), can be used as the inputs of a CAD system. Through an interferometric imaging modality, OCT can visualize the micro-structure of tissues via near-infrared light interferometry. Because of the advantages of non-invasiveness, high resolution, and fast processing, OCT has draw a lot of attention in recent years and been used in different areas, such as thyroid tissue imaging and clinical ophthalmic imaging [2]–[4].

Over the decade, machine learning techniques have been developed rapidly and employed in CAD systems to provide accurate and fast diagnosis [5]. The remarkable success of artificial intelligence (AI) brings more progresses in the domain of medical image analysis. The capability of an effective AI model heavily relies on the learning from sufficient amount of training samples. For instance, in a convolutional neural network (CNN) training process, millions of parameters are optimized, which requires a large number of labeled samples. However, the OCT technique has not been widely used for different tissue imaging tests compared to other medical imaging technologies, such as CT and MRI. Moreover, the high cost of annotating collected images by radiologists aggravates the scarcity of training samples. The lack of annotated images leads to a demand of applying image synthesis techniques in CAD model development processes. Furthermore, thyroid cancer is the most common cancer of the human endocrine system. According to the National Cancer Institute, 56,870 new cases were reported in 2017 [4]. To diagnose thyroid cancer at an early stage, the first step is to detect thyroid tissues

2575-2634/18/$31.00 ©2018 IEEE
DOI 10.1109/ICHI.2018.00030

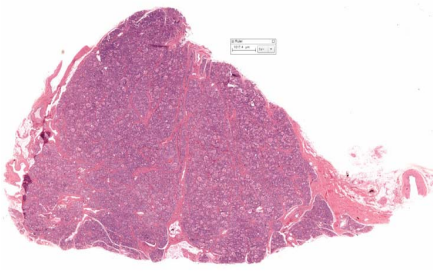
successfully from medical images.

Therefore, the objective of this research is to solve the dilemma of insufficient medical images and improve the capacity of recognizing human thyroid tissues, by means of synthesizing OCT images through the application of AI. Specifically, OCT images are synthesized by applying generative adversarial networks (GANs) [6], which consist of a generative network and a discriminative network, and can capture feature representations that describe a high level of image semantic information. Through a minimax two-player game, the generative and discriminative networks can train each other. The trained generative network is then applied to generate synthetic OCT images, which are mapped from random noise. Using the trained generative network, the synthetic images can be generated to recover the original image data distribution. Therefore, the diversity of image samples is increased, because the synthetic samples can cover diverse sample positions among the data distribution. Three GAN models, i.e., deep convolutional GANs (DCGANs) [7], Wasserstein GANs (WGANs) [8], and boundary equilibrium GANs (BEGANs) [9] are implemented and compared for synthesizing images. The effectiveness of the synthetic images generated by the three models are evaluated by the classification results of a CNN classification model regarding accuracy, sensitivity, specificity, and area under the curve (AUC). Through the implementation of synthetic images, an improvement of the CNN classification model's capacity to recognize tissues is achieved by DCGANs. The effectiveness of synthetic thyroid images by WGANs and BEGANs are also validated based on the thyroid classification results from the CNN classification model.

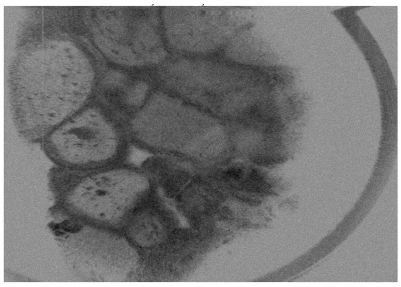
The structure of this paper is organized as follows: The background of OCT image tissue recognition problem and the related research about image synthesis are introduced in Section 2. Section 3 describes DCGANs, WGANs, BEGANs, and the thyroid recognition CNN model. The data description, synthetic images, and experimental results are shown and discussed in Section 4. The conclusion and future work are presented in Section 5.

II. Background

In this research, the objective of synthesizing images is to improve the capacity of the classification model to recognize human thyroid tissues from OCT images. The thyroid gland is at the front of the neck and consists of two lobes connected by an isthmus. The thyroid gland plays an important role in the control of metabolic rate and protein synthesis by producing hormones. Four main structures are included in thyroid: capsule, trabecula, parafollicular cell and follicle. Fig. 1a shows an example of human thyroid histology image. Follicle, with round or oval shapes ranging from 50 to 500 μm, is the most important component in thyroid by storing hormones. Because the high resolution of OCT images, the follicle structure information of a thyroid can be obtained, which is shown in Fig. 1b. The existence of the follicle is



(a) Human thyroid histology image



(b) Human thyroid OCT image

Fig. 1: Human thyroid images.

the main characteristic to represent thyroid, so it should be detected to recognize thyroid tissues [10].

Various image synthesis techniques have been developed to deal with the scarcity of training images. In previous research, a common approach to cope with few training samples is image transformation based on the available images. The image transformation strategies such as rotation, flipping, size scaling, translation, contrast enhancement, color transformation, and elastic distortions are selected manually [11]–[13]. Recently, inspired by the advance of deep learning techniques, efforts have been directed in synthesizing images from random inputs through the development of an adversarial learning-based image generator, which has various applications, such as remote sensing image classification [14]. Since the development of GANs in 2014 [6], numerous GANs related algorithms have been developed accordingly. For instance, WGANs introduce the measure of convergence, and get rid of mode collapse problems [8]. The WGAN model performs better than the DCGAN model in terms of stability and convergence. Applying an equilibrium enforcing method, BEGANs were proposed, which balance the generator and discriminator during the training process. Also, a trade-off between image diversity and visual quality is achieved using BEGANs [9]. However, GAN models have not been applied to improve the capacity of recognizing tissues. In this research, this gap is addressed by developing automatic image synthesis models through GAN models and applying to a human thyroid recognition problem.

III. METHODOLOGY

In this research, synthetic images generated by the GAN models are applied to train a thyroid recognition model. DCGANs, WGANs, and BEGANs are described in this section, which are used to synthesize training images. A thyroid recognition model is trained to classify thyroid images and non-thyroid images.

A. DCGANs

DCGANs use the concept of the typical GAN model, which consists of a generative network (G) and a discriminative network (D). As one of the state-of-the-art techniques for image processing, CNN is implemented in both G and D. The training data distribution can be captured and synthetic images can be generated mapping from random noise by G. Meanwhile, the effectiveness of the generative network is validated by D, which is trained to detect synthetic images from real images. Random noise z from distribution $p_g(z)$ is mapped to the synthetic data $G(z)$ by G. In addition, the output of D is single scalar $D(x)$, which represents the probability of image x comes from original sample distribution $p_r(x)$. The objective of G is to minimize the probability that synthetic images being recognized, which is represented as $\log(1 - D(G(z)))$. The objective of D is to maximize the probability of classifying the images correctly. Therefore, D and G constitute a two-player minimax game as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_g(z)}[\log(1 - D(G(z))] \tag{1}$$

where $\mathbb{E}$ is expected value. Both G and D are trained through learning iterations based on the competing objectives [6].

Fig. 2 shows the architecture of the DCGAN model used in this paper. Through a fully connected layer, 100 random numbers are used as the inputs of the generative network G and mapped to a 25,588 neuron layer. After the employment of a leaky rectified linear unit (leaky ReLU) activation function [15], the data is reshaped to $128 \times 14 \times 14$. Followed by an upsampling process, a filter process with 64 5×5 filters, and a leaky ReLU, the dimension of the third layer is $64 \times 28 \times 28$. Finally, a 56×56 synthetic image is obtained through an upsampling process, a convolution layer with a 5×5 filter, and a tanh activation function. The output of G, combined with the real training images, are used as the inputs of the discriminator network D. In this study, the training images are scaled to the size of 56×56. G is designed to generate synthetic images with the same size as the original training images to facilitate the discriminating process using D. The input image for D is convoluted by 64 5×5 filters. Leaky ReLU and dropout are used in this convolutional layer [16]. The second layer has the same architecture with 128 filters. Then, a fully connected layer with 25,088 neurons is used to export the classification result of the input image using sigmoid activation function. Suggested by Nie et al., polling layers are not involved in the developed GAN model [17]. Although pooling layer shows the advantage of enhancing the invariance of image distortions, the spatial resolution of feature maps the quality of synthetic images will be reduced using pooling.

B. WGANs

Wasserstein distance, which is used to measure the minimum cost of transporting mass for transforming a distribution to the another distribution, is applied in the WGAN model. Specifically, Wasserstein-1, which is also known as earth-mover (EM) distance, is implemented in the loss function. The Wasserstein-1 distance is defined as follows:

$$W(P_r, P_g) = \inf_{\gamma\in\Pi(P_r,P_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||] \tag{2}$$

where $\Pi(P_r, P_g)$ is the set of all joint distributions $\gamma(x, y)$ with the marginal distributions of P_r and P_g. The use of Wasserstein-1 distance improves the stability of the training process significantly. To apply the Wasserstein-1 distance in the training process, (2) is calculated approximately according to the Kantorovich-Rubinstein duality and 1-Lipschitz functions [18]. As a result, the objective function of the WGAN model becomes to

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_r(x)}[f(x)] - \mathbb{E}_{z\sim p_{g(z)}}[f(G(z))] \tag{3}$$

The WGAN model is trained similar to the typical GAN model, excepting the $\log$ function is eliminated from WGANs, based on the property of Wasserstein-1 approximation [8].

To address the effectiveness of the Wasserstein-1 distance for the the WGAN model, a simple architecture, which is also implemented in the research of interpretable representation learning by information maximizing GAN (InfoGAN), is applied [19]. For the discriminator D, the training images are mapped to the first layer using 64 4×4 filters with the stride size of 2×2 and the leaky ReLU function. The second layer has the same shape but 128 filters. Following a batch normalization, a fully connected layer with 1,024 neurons and leaky ReLU is used to connect the output layer of the discriminator D. The generator G maps 100 random numbers to two fully connected layers with 1,024 neurons and 6,272 neurons using ReLU. Using an upsampling process, a filter process with 64 4×4 filters with the stride size of 2×2, a leaky ReLU and batch normalization are implemented. The synthetic images are generated using a 4×4 upsampling process.

C. BEGANs

Taking the advantages of Wasserstein distance, the BEGAN model utilizes the lower bound of the Wasserstein distance to train auto-encoders, instead of a probabilistic binary discriminator in the typical GAN model. In other words, the BEGAN model aims to match the distribution of the loss in an auto-encoder instead of the distribution of training samples. According to Jensens inequality, the lower bound of the Wasserstein distance for two distributions is the absolute difference of the means for the two distributions. As a result, the objective of the discriminator in BEGANs is to maximize the difference

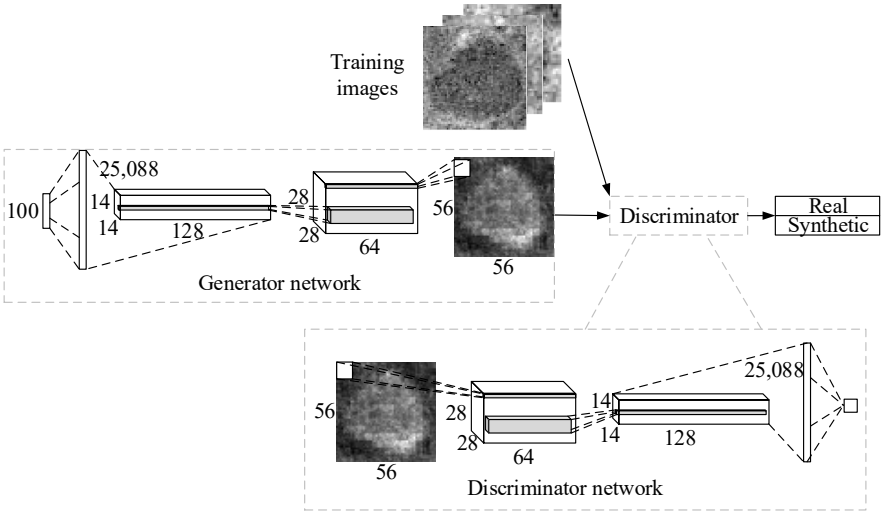


Fig. 2: The architecture of the DCGAN model. The synthetic images generated by G and training images are used as the input for D.

of the loss in the auto-encoders and the generator aims to minimize the difference of the losses. The objective function of BEGANs is shown as follows:

$$\min_G \max_D V(D,G) = \mathcal{L}(x;\theta_D) - \mathcal{L}(G(z);\theta_D) \tag{4}$$

where $\mathcal{L}(\cdot\,,\theta_D)$ is the loss function of the auto-encoder for the discriminator. A balance mechanism is applied to maintain the balance of the discriminator and the generator using a hyper-parameter $\gamma \in [0,1]$ [9], which is defined as follows:

$$\gamma = \frac{\mathbb{E}[\mathcal{L}(G(z);\theta_G)]}{\mathbb{E}[\mathcal{L}(x);\theta_D]} \tag{5}$$

The auto-encoder is involved in both the discriminator D and the generator G. Similar to the discriminator in WGANs, the input image is mapped to the first layer using 64 4×4 filters with the stride size of 2×2 and the leaky ReLU function. Then a batch normalization is implemented. Two fully connected layers with 32 neurons and 12,544 neurons are constructed using ReLU. The output of the two layers is mapped to a convolutional layer with the shape of $64 \times 14 \times 14$, and the output of the discriminator auto-encoder is generated through a 4×4 filter with the 2×2 stride and the sigmoid function. The architecture used for the generator is the same as the WGAN model.

D. Thyroid Recognition Model and Validation Method

The synthetic images generated using the GAN models are applied on a human thyroid recognition problem by a CNN classification model. Two convolutional layers are constructed with 5×5 filters and leaky ReLU functions. 32 and 128 filters are used in the two layers, respectively. A max-pooling process is applied with the stride size of 2×2. The output of the two layers is fed forward to a fully connected network with 1,024 neurons and ReLU activation function. The classification result is generated through another fully connected network with two neurons and sigmoid activation function.

Five-fold cross validation is applied to validate the recognition results. The images in the dataset are divided into five partitions. For each validation, four of the partitions are used as training images and the remaining partition is used as testing images. The training images are used to train the GAN model. In other words, in each validation process, a GAN model is trained based on the training images. Synthetic images are generated based on the trained generator G. Then the synthetic images and the original training images are combined to train the CNN classification model. The classification results are evaluated by classification accuracy, sensitivity, specificity, and AUC using the testing images.

IV. Experimental Results and Analysis

The effectiveness and applicability of the image synthesis models are illustrated by applying to a human thyroid recognition problem using OCT images. From the OCT thyroid scans, region of interests (ROIs) are cropped and labeled manually. 291 thyroid images and 362 non-thyroid images are collected. The collected OCT images are scaled to the size of 56×56 pixels to cope with the variety of image sizes caused by the image cropping process. To compare the DCGAN, WGAN, and BEGAN models, a comparable study is conducted by means of classification results from a CNN model, which is trained by the synthetic and the original images. The networks are trained using the Adam optimizer with a learning rate of 0.002 and $\beta_1 = 0.5$ [20]. 200 and 40 epochs are used to train the GAN and CNN classification models, respectively.

The batch size used is 64. $p = 0.3$ and $\alpha = 0.2$ are used for dropout and leaky ReLU functions in the DCGAN model. The critic iteration for the WGAN model is $n = 1$. $\gamma = 0.75$ and $\lambda = 0.001$ are used in the BEGAN model.

To achieve the best thyroid recognition performance, different image synthesis strategies are also applied and compared. In terms of the synthetic image labels, the first strategy is termed as "Thyroid", in which the label for the synthetic images is "Thyroid". The second strategy is termed as "All", which synthesizes both thyroid and non-thyroid images. In addition, 200, 400 and 600 synthetic images are employed in the two strategies. For examples, if the number of synthetic images is 200 for "Thyroid", 200 thyroid images are generated. If the strategy is "All", 200 synthetic thyroid images and 200 non-thyroid images are generated. The synthetic images and original images are combined to train the CNN classification model. The experiments are replicated five times. All the GAN models and the CNN classification model are implemented on a 64 bit Windows 8 system with Intel i7 processor (3.40 GHz) and 16 GB random-access memory (RAM), using the Keras and TensorFlow libraries on Python 3.6.

The original OCT images are showed in Fig. 3, which plots 25 thyroid samples and 25 non-thyroid images. Also, using DCGANs, WGANs, and BEGANs, synthetic images are generated and plotted in Fig. 4. It is clear that all the three models can capture most of the general visual characteristics of the original images for both thyroid and non-thyroid images. In terms of synthetic images, those samples generated by the DCGAN model show the same patterns as the original images for both thyroid and non-thyroid images. However, the quality of synthetic images is reduced compared to the original images because of the loss of information in the training process. WGANs can generate more realistic images compared to DCGANs and BEGANs. The images generated by BEGANs have higher variation compared to the images generated by the other two models. In addition, it is noticed that there are noises existing in the images generated from WGANs for both thyroid and non-thyroid images.

A. Classification Performance Evaluation

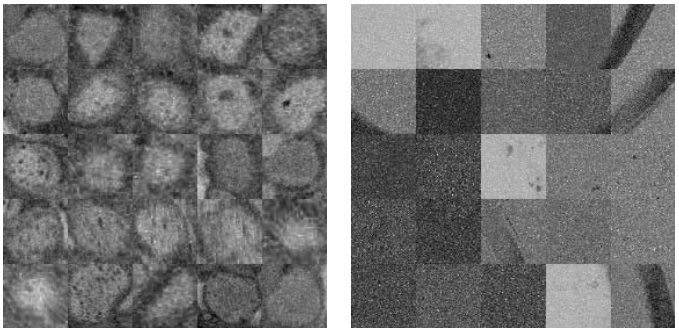
This research aims to improve the classification model's capacity of recognizing thyroid tissues through the implementation of synthetic images. Therefore, the classification accuracy, sensitivity, specificity, and AUC are used to evaluate the effectiveness of the image synthesis models. The classification results obtained using different image synthesis strategies are summarized in Table I, which shows the average and standard deviation (SD) of classification performance measures. Compared to the classification results using the original images, all the image synthesis models and strategies can improve the classification results significantly. In addition, the SD of the classification results regarding accuracy, sensitivity, specificity and AUC are reduced significantly by implementing the synthetic images. In other words, the increase of training samples improves the robustness of the classification model. The DCGAN model achieves the highest accuracy as 98.83% by incorporating both 600 thyroid and 600 non-thyroid images. Compared to the classification results using the original images, DCGAN-based synthesis approach helps the CNN classification model obtain up to 3% higher accuracy. In addition, the more synthetic images are used, the higher accuracies obtained. To visualize the classification performance, the classification accuracies, sensitivities, specificities using DCGANs, WGANs, and BEGANs are plotted through boxplots in Fig. 5-7. The average accuracies of different strategies are plotted using the symbol of "*".

From the classification results, although the quality of the synthetic images from DCGANs is not as good as those from WGANs and BEGANs, the improvement of classification results is dramatic by the DCGANs image synthesis. Moreover, sensitivity, specificity, and AUC are all increased using the synthetic images generated by DCGANs. Therefore, the overall classification ability increases because of the implementation of the synthetic images. Comparing the classification results of "Thyroid" and "All" with the DCGAN model, "All" strategy can generate similar or higher accuracies compared to "Thyroid" strategy. This is because the implementation of synthetic non-thyroid images improves the classifier's ability to recognize the non-thyroid images.

Regarding the WGAN model, the similar results are obtained with a dramatic improvement of the classification results. However, using different number of synthetic images, the classification results from WGANs have less variation than the results from DCGANs and BEGANs, which imply that the impact on the classification results caused by the WGAN model is more stable than the other two image synthesis models. Moreover, comparing "Thyroid" and "All" strategies, it is noticed that "Thyroid" strategy outperforms "All" strategy. In other words, the synthesis of the non-thyroid images has a negative impact on the training of the classification model. This can be also explained by the sensitivity and specificity, which are decreased because of the introduction of the synthetic non-thyroid images. Therefore, the synthetic non-thyroid images have limited ability to improve the classification model's tissue recognition capacity due to the influence of noise generated by the WGAN model. Using the WGAN model, the highest average classification accuracy obtained is 98.54% by synthesizing 600 thyroid images, which is consistent with the conclusion that more synthetic images being used, the more the classification capacity can be achieved.

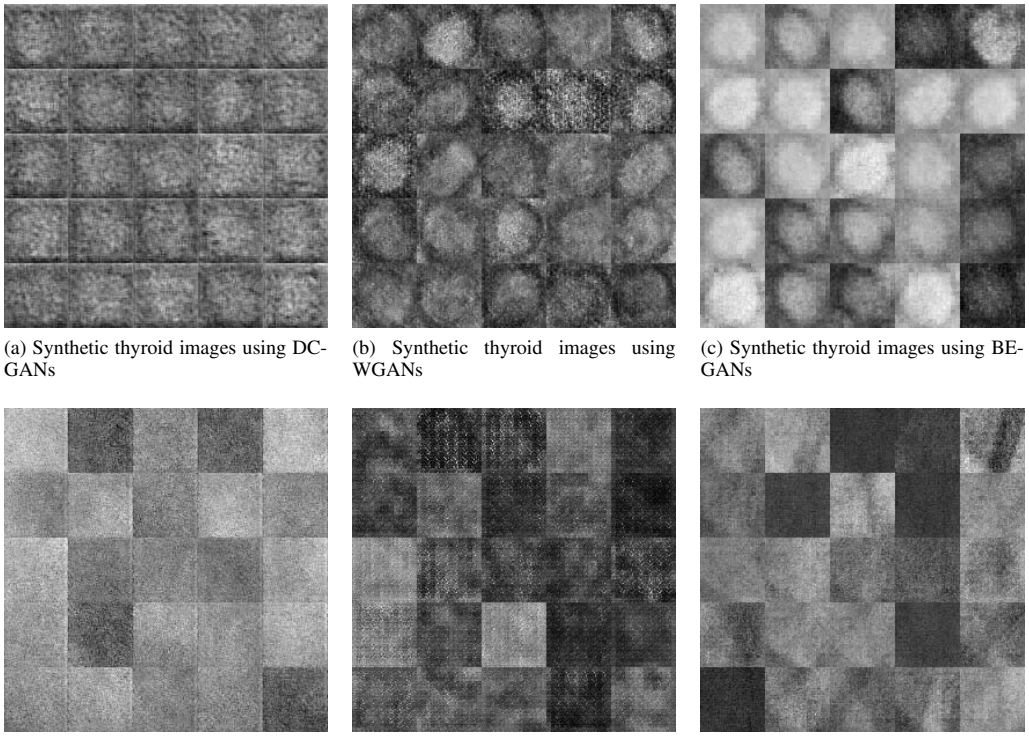
The classification accuracies with the synthetic images generated by the BEGAN model are lower than the accuracies from WGANs. In addition, "Thyroid" strategy outperforms "All" strategy for BEGANs, which is similar to WGANs. Using the BEGAN model, the highest average classification accuracy obtained is 98.48% by synthesizing 600 thyroid images.

The comparison of classification accuracies are plotted in Fig. 8. If the same number of synthetic images are generated using "Thyroid" strategy, both WGANs and BEGANs can obtain higher classification accuracies compared to DCGANs. The improvements of classification accuracies demonstrate the



(a) Original thyroid OCT images (b) Original non-thyroid OCT images

Fig. 3: Original human thyroid OCT images.



(a) Synthetic thyroid images using DCGANs (b) Synthetic thyroid images using WGANs (c) Synthetic thyroid images using BEGANs

(d) Synthetic non-thyroid images using DCGANs (e) Synthetic non-thyroid images using WGANs (f) Synthetic non-thyroid images using BEGANs

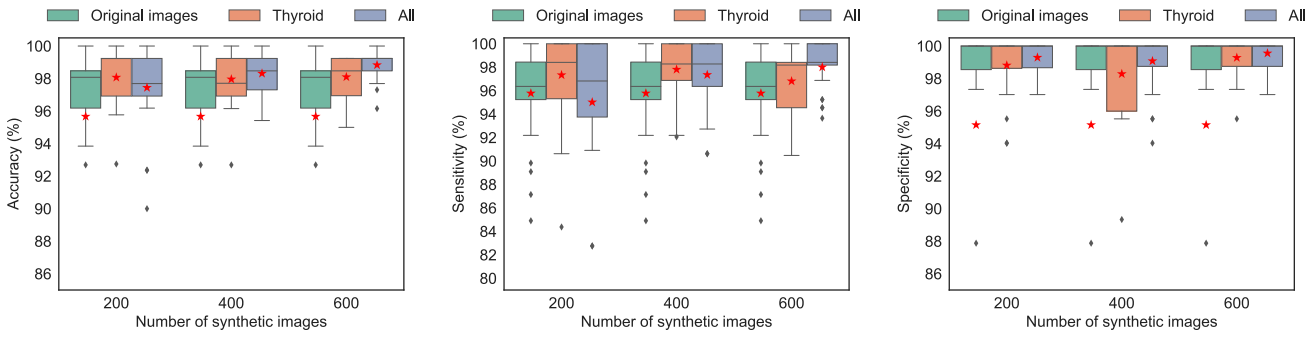
Fig. 4: Synthetic human thyroid OCT images.

advantage of WGANs and BEGANs to synthesize thyroid images compared DCGANs. Moreover, the WGAN model shows a better stability to improve the classification results for different strategies regarding the number of the images to be synthesized. For instance, if 200 synthetic thyroid images are synthesized, the classification accuracy obtained by WGANs is 98,19%, which is higher than 97.96% generated by BEGANs. Additionally, if the image synthesis strategy used is "All", WGANs and BEGANs do not achieve the same results as DCGANs. Consequently, WGANs and BEGANs have the ability to capture the distribution of thyroid images. In contrast, DCGANs show the higher capacity to learn the distribution of non-thyroid images.

In summary, the classification results through different image synthesis strategies manifest the effectiveness of the synthetic images, which can be used to solve the dilemma of insufficient samples for tissue recognition. Also, it is indicated that more synthetic images are desired to generate better tissue recognition results. Comparing DCGANs, WGANs and BEGANs through thyroid classification accuracy from five-fold cross validation, the thyroid images generated from WGANs and BEGANs improve the classification results more than those images generated by DCGANs. However, the DCGAN model shows a higher capacity to improve the thyroid

TABLE I: Image classification performance

Image synthesis model	Image synthesis strategy	Number of synthetic images	Accuracy (SD)	Sensitivity (SD)	Specificity (SD)	AUC (SD)
Original images	-	0	95.67% (9.38)	95.78% (4.15)	95.15% (19.03)	99.46% (0.72)
DCGANs	Thyroid	200	98.07% (1.88)	97.34% (3.87)	98.81% (1.84)	99.54% (0.59)
		400	97.96% (1.62)	97.81% (2.51)	98.29% (2.60)	99.55% (0.60)
		600	98.10% (1.40)	96.81% (2.90)	99.29% (1.10)	99.59% (0.51)
	All	200	97.44% (2.50)	95.02% (6.09)	99.29% (1.20)	99.62% (0.46)
		400	98.31% (1.31)	97.35% (2.66)	99.08% (1.67)	99.65% (0.49)
		600	98.83% (0.96)	98.01% (1.88)	99.56% (0.77)	99.65% (0.49)
WGANs	Thyroid	200	98.19% (1.20)	97.46% (2.32)	98.94% (1.43)	99.92% (0.09)
		400	98.56% (1.16)	98.20% (2.10)	98.98% (1.32)	99.92% (0.11)
		600	98.54% (1.21)	97.98% (2.25)	99.09% (1.15)	99.91% (0.12)
	All	200	98.02% (1.06)	97.42% (2.41)	98.41% (1.36)	99.86% (0.17)
		400	98.04% (1.32)	96.99% (2.45)	98.86% (1.28)	99.78% (0.35)
		600	98.27% (1.23)	97.61% (2.16)	98.80% (1.18)	99.85% (0.22)
BEGANs	Thyroid	200	97.96% (1.74)	97.37% (2.34)	98.56% (2.51)	99.93% (0.09)
		400	98.47% (1.16)	97.92% (2.51)	98.93% (1.26)	99.93% (0.11)
		600	98.48% (1.19)	97.77% (2.11)	99.07% (0.94)	99.93% (0.08)
	All	200	97.79% (1.48)	96.60% (3.02)	98.61% (1.47)	99.80% (0.35)
		400	97.87% (1.26)	97.08% (2.40)	98.44% (1.51)	99.80% (0.25)
		600	97.98% (1.07)	97.35% (2.11)	98.55% (1.55)	99.84% (0.23)



(a) Classification accuracies of DCGANs (b) Classification sensitivities of DCGANs (c) Classification specificities of DCGANs

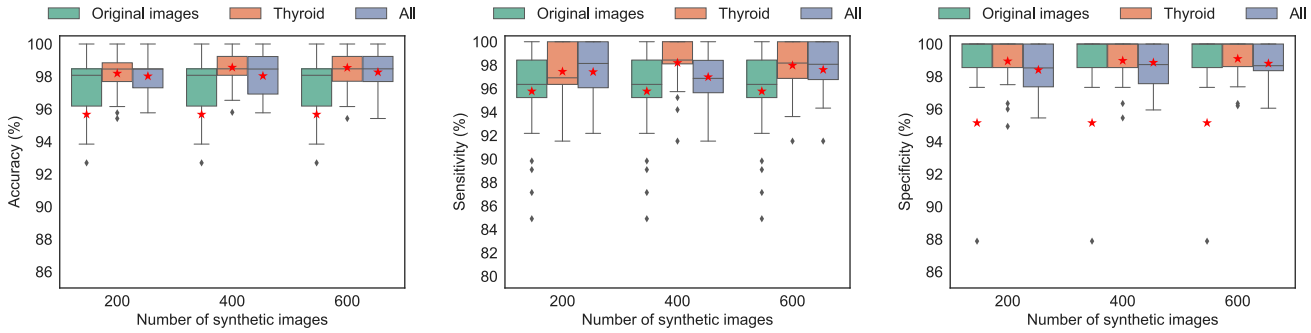
Fig. 5: Classification performance comparison of DCGANs using different image synthesis strategies.

recognition model by implementing "All" strategy. Overall, the DCGAN model increases the average classification accuracy to 98.83%. Regarding "Thyroid" strategy, the highest average classification accuracy achieved is 98.54% using WGANs.

V. Conclusion and Future Work

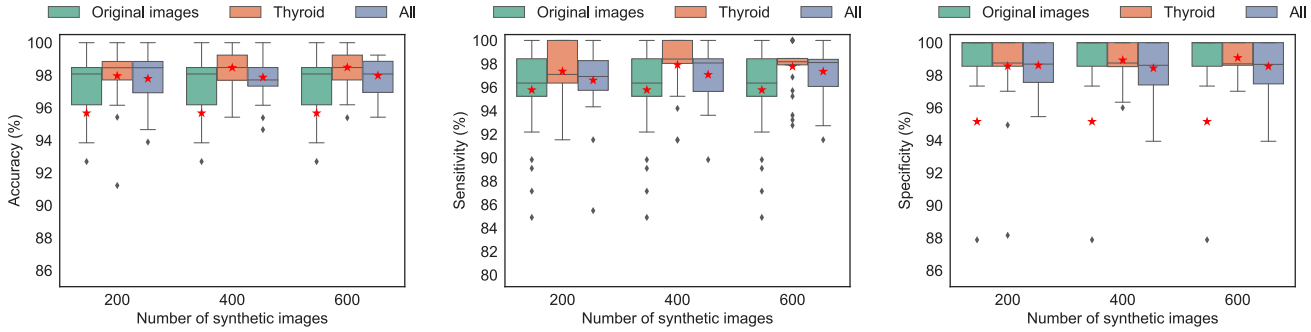
To synthesize medical images for a medical image tissue recognition problem, three GAN models, i.e., DCGANs, WGANs, and BEGANs, are considered in this study. A GAN model consists a generative network and a discriminative network. Through a minimax two-player game, the two networks train each other iteratively. The image synthesis models solve the dilemma of insufficient medical image sample problem for tissue recognition. A CNN classification model is developed and trained using the synthetic images from GANs. This research firstly applies GANs to a thyroid recognition problem. Using different image synthesis strategies and multiple classification measures, an improvement regarding the capacity to recognize thyroid images is achieved. The effectiveness and applicability of synthesizing medical images through the image synthesis models are also demonstrated. It is concluded that WGANs and BEGANs outperform DCGANs to synthesize thyroid images, while the images synthesized by DCGANs have higher impact than WGANs and BEGANs on tissue recognition.

Regardless of remarkable results through the study, there are still some limitations. Firstly, the improvements of the thyroid recognition model caused by the synthetic non-thyroid images are still limited because of the heterogeneity of non-thyroid images. So the future of this model can be extended to handle the heterogeneity of different images. In addition to the thyroid recognition problem, further evaluation of the



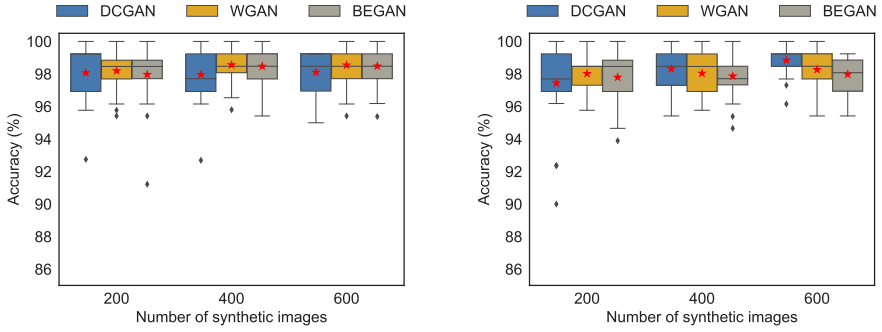
(a) Classification accuracies of WGANs (b) Classification sensitivities of WGANs (c) Classification specificities of WGANs

Fig. 6: Classification performance comparison of WGANs using different image synthesis strategies.



(a) Classification accuracies of BEGANs (b) Classification sensitivities of BEGANs (c) Classification specificities of BEGANs

Fig. 7: Classification performance comparison of BEGANs using different image synthesis strategies



(a) Classification accuracies of "Thyroid" strategy (b) Classification accuracies of "All" strategy

Fig. 8: Comparison of classification performance using different image synthesis models.

GAN models can be done through the application on standard datasets and the involvement of more performance metrics. Also, the implementation and comparison of more GAN models are desired. Moreover, this research recognizes thyroid images based on follicles. A GAN model can be developed to synthesize the whole thyroid structure instead of the manually cropped ROI images.

References

[1] D. Kunio, "Current status and future potential of computer-aided diagnosis in medical imaging," The British Journal of Radiology, vol. 78, pp. 3-19, 2005.

[2] R.G. Mirza, M.W. Johnson, and L.M. Jampol, "Optical coherence tomography use in evaluation of the vitreoretinal interface: a review," Survey of Ophthalmology, vol. 52, no. 4, pp. 397-421, 2007.

[3] L. Pantanowitz, P.L. Hsiung, T.H. Ko, K. Schneider, P.R. Herz, J.G. Fujimoto, S. Raza, and J.L. Connolly, "Highresolution imaging of the thyroid gland using optical coherence tomography," Head & Neck, vol. 26, no. 5, pp. 425-434, 2004.

[4] S.J. Erickson-Bhatt, K.J. Mesa, M. Marjanovic, E.J. Chaney, A. Ahmad, P.C. Huang, Z.G. Liu, K. Cunningham, and S.A. Boppart, "Intraoperative optical coherence tomography of the human thyroid: Feasibility for surgical assessment," Translational Research, vol. 195, pp. 13-24, 2018.

[5] K. Doi, "Computer-aided diagnosis in medical imaging: historical

review, current status and future potential," Computerized Medical Imaging and Graphics, vol. 31, no. 4, pp. 198-211, 2007.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Proceedings of the Advances in Neural Information Processing Systems, pp. 2672-2680, 2014.
[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.
[8] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," Proceedings of the International Conference on Machine Learning, pp. 214-223, 2017.
[9] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: boundary equilibrium generative adversarial networks," arXiv Preprint arXiv:1703.10717, 2017.
[10] H. Wang, and S.W. Yoon,"A machine learning model for medical image recognition using texture-based features," Proceedings of the IIE Annual Conference, Institute of Industrial and Systems Engineers (IISE), pp. 1655-1660, 2017.
[11] M. Drozdzal, G. Chartrand, E. Vorontsov, L. Di Jorio, A. Tang, A. Romero, Y. Bengio, C. Pal, and S. Kadoury, "Learning normalized inputs for iterative estimation in medical image segmentation," arXiv Preprint arXiv:1702.05174, 2017.
[12] M.W. Lafarge, J.P. Pluim, K.A. Eppenhof, P. Moeskops, and M. Veta, "Domain-adversarial neural networks to address the appearance variability of histopathology images," Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, Cham, pp. 83-91, 2017.
[13] P.Y. Simard, D. Steinkraus, and J.C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," Proceedings of the ICDAR, vol. 31, pp. 958-962, 2003.
[14] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: unsupervised representation learning for remote sensing image classification," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 11, pp. 2092-2096, 2017.
[15] A.L. Maas, A.Y. Hannun, and A.Y. Ng, "Rectifier nonlinearities improve neural network acoustic models" Proceedings of the Proc. ICML, vol. 30, no. 1, 2013.
[16] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014.
[17] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, pp. 417-425, 2017.
[18] C. Villani, 2008, "Optimal transport: old and new," Springer Science and Business Media, pp. 338, 2008.
[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: interpretable representation learning by information maximizing generative adversarial nets," Proceedings of the Advances in Neural Information Processing Systems, pp. 2172-2180, 2016.
[20] D. Kingma, and J. Ba, "Adam: a method for stochastic optimization," arXiv Preprint arXiv:1412.6980, 2014.