2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-0358-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/BHI50953.2021.9508532

Improving Heart Transplant Rejection Classification Training using Progressive Generative Adversarial Networks

Ali Mirzazadeh[1], Arshawn Mohseni[2], Sahar Ibrahim[2], Felipe O. Giuste[2], Yuanda Zhu[3], Bahig M. Shehata[4], Shriprasad R. Deshpande[5], and May D. Wang[2]

***Abstract*—Cardiac allograft rejection is a life-threatening complication that can occur in patients following heart transplantation. Endomyocardial biopsies, the current gold-standard for monitoring rejection, require manual identification of samples by experts; however, this can be subjective, costly and time-consuming. Computer-aided diagnosis models can potentially provide an automated analysis that offers accurate and consistent detection of biopsy samples. Unfortunately, the lack of a large dataset of rejection pathology signs, due to time consuming clinical annotations, limits the classification performance of such conventional AI-based models. In this paper, we developed a generative adversarial network (GAN) that creates synthetic tissue tiles from heart transplant whole slide images (WSIs) to serve as data for training rejection classifier models. To generate synthetic rejection histology regions, we used inspirational image generation (IIG) with a single rejection reference image. Additionally, to demonstrate objective improvement in image classification using synthetic rejection regions, we use a pretrained VGG-19 classifier to differentiate between rejection versus nonrejection tiles. We greatly improved classification performance, achieving an increase in the Matthews correlation coefficient (MCC) from 0.411 to 0.790 when the training set was augmented with our synthetic rejection tiles. Our model was able to create visibly realistic rejection tissue tiles which was used to augment the rejection tile database and enhanced automated rejection detection.**

***Keywords*—Deep Learning, Digital Pathology, Generative Adversarial Network, Heart Transplant Rejection, Data Augmentation**

I. Introduction

Cardiac allograft rejection is a serious and often fatal complication that can affect heart transplant patients [2], [12]. Therefore, early detection of rejection can increase the success of heart transplantation by providing timely treatment to reduce complications and heart failure [2].

The presence of rejection is typically monitored through multiple post-transplant biopsies that undergo expert analysis. Specifically, this process requires a pathologist to manually examine and evaluate each whole slide image (WSI), making it costly, time-consuming and prone to human error [3], [12]. Automated classification can provide efficient and timely rejection detection to solve these challenges. However, current deep learning models have had limited success in assessing the presence of rejection regions. A significant issue is that medical imaging datasets are limited in the incidence of expert-annotated pathological findings, which in turn constrains the training of deep networks. We hypothesize that robust data augmentation with synthetic images, for classes with few instances, would therefore be beneficial towards optimizing deep learning-based rejection detection.

Generative adversarial network (GAN) is a deep-learning base method of data augmentation that creates realistic synthetic images. GANs generally consist of two networks, a generator and a discriminator. These networks compete with one another to improve synthetic image generation through iterations. Specifically, a generator converts random noise into synthetic images. Meanwhile, the discriminator network must classify images as real (from the original database) or fake (created from the generator). The result of this system is that the generator eventually learns to create realistic images in order to deceive the ever-improving discriminator.

Progressive GANs (PGAN) provide a major training speed and stability boost to their traditional GAN counterparts. In their paper, Karras et al. [9] demonstrated this novel method of GAN training, which begins by solely training the lowest resolution layer, and then gradually introducing new layers to the model architecture. Layers are added during training until the desired final image resolution is reached (Fig. 1). Inspirational image generation (IIG) is a method of GAN-based image generation that enables reference-based (or "inspired") image generation. In the system proposed by Roziere et al. [11], an additional optimizer is added to a pretrained GAN, which minimizes the distance of the generator input features from those extracted from a reference image with a known label. The resulting generated image is thus encouraged to share similarity with the reference image.

This research was supported a Wallace H. Coulter Distinguished Faculty Fellowship (M. D. Wang), a Petit Institute Faculty Fellowship (M. D. Wang), and by Microsoft Research.

[1]A. Mirzazadeh is with the College of Computing, Georgia Institute of Technology, Atlanta, GA 30322 USA (email: alimirz@gatech.edu).

[2]A. Mohseni, S. Ibrahim, F. O. Giuste, and M. D. Wang are with the Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30322 USA (email: amohseni6@gatech.edu; sibrahim32@gatech.edu; fgiuste@gatech.edu; maywang@bme.gatech.edu).

[3]Y. Zhu is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30322 USA (email: yzhu94@gatech.edu).

[4]B. Shehata is with the Department of Pathology, Wayne State University School of Medicine, Detroit, MI 48201 USA (e-mail: bshehata@dmc.org).

[5]S. R. Deshpande is with the Pediatric Cardiology, Children's National Health System, Washington, DC 20010 USA (email: sdeshpan@childrensnational.org).

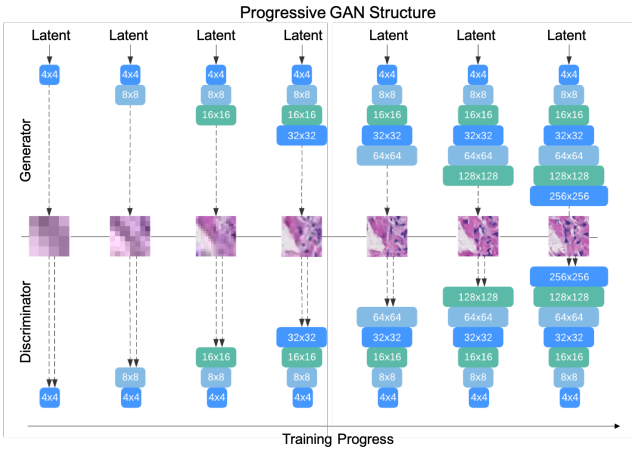

Fig. 1. Breakdown of the original GAN and Progressive GAN utilized. Each layer for the progressive GAN is much deeper which allows for more feature importance to be extracted and generated per iteration when compared to the initial GAN pipeline that was being used. This allows increases the training speed. The final output of the progressive GAN was a 256 x 256 image.

II. Background

A. Heart Transplant Rejection

Once the host recognizes allograft histocompatibility antigens as foreign, T-cells and macrophages infiltrate the allograft myocardium, inducing myocyte damage and necrosis [13]. This inflammatory infiltrate as well as macrophage deterioration can be observed through biopsy samples and therefore, was a key indicator for developing rejection tiles in our GAN.

Machine learning models have been developed to aid in detection of rejection regions within whole slide image biopsies [2], [12], [17]. While demonstrating moderate success, these models were inherently limited in performance by the minute amount of rejection tile data available, as clinically labeled, pathology data from such conditions is quite rare.

B. Generative Adversarial Networks

In recent literature, we found that GANs had been implemented for a multitude of applications [6], [7], [10], [14], [15] within the medical field due to their potential to improve classification performance as well as their significance as an anonymization tool. Shin et al. [6] trained a GAN to generate synthetic abnormal MRI images and, although similar to our goal, their model utilized labels as an input for the discriminator which was not feasible for our imbalanced dataset. Likewise, the conditional GAN (cGAN) generated by Yuan et al. [14] was trained on labeled images to guarantee relevant features were implemented for their classification objective. Related models generated for classification also focused on feature-based filtering employed with the cGAN, with labels being inputted for training [15]. The GAN developed by Liu et al. [10] improved prediction performance of a distinct mutation found in tumors, through data augmentation. Their model did not include inputting labels into the discriminator, as observed in previous models. Instead, two GANs were utilized, one network being trained on mutant image patches and the other trained on wild-type patches, in order to produce 256x256 images as a result of their joint implementation. These approaches are not feasible for heart transplant rejection because there is too few rejection tiles for the GAN to properly train as its own class.

III. Methods

A. Dataset

We developed our model utilizing images from two datasets: the DNA Based Transplant Rejection (DTRT) [16] and Children's Hospital of Atlanta (CHOA) [2], [3], [12], [17] datasets. The latter consisted of digital high-resolution WSIs of endomyocardial biopsies, with hematoxylin and eosin (H&E) staining, obtained from pediatric heart transplant patients from CHOA. A certified expert annotated areas of rejection in both datasets providing 8 manual labels to identify these tiles during patch generation.

To ensure there was no data leakage when evaluating the classifier, three WSIs that had not been utilized in the training of the GAN were segmented and 1029 tiles were set aside which would consist of the nonrejection tiles for the test set (Fig. 2). For the rejection label, there were seven WSI that contained manual rejection annotations were segmented for a total of eight rejection regions, and any tiles that overlapped with the rejection area were set aside. From here we ensured that there was approximately 50% overlap between the tile and the annotated rejection region. Next there was a training testing split of the rejection and nonrejection tissues while ensuring there was no subjected overlap between the two groups.

B. Patch Generation

Preprocessing of the WSIs was integral to maximizing the performance of our GAN. Tissue tiles were generated at 40X magnification with pixel dimensions of 256x256. HistoQC [8] was attempted to be used for quality control and normalization, but due to differences in the required file input type and the format of the WSI, this was not feasible. We therefore used manual quality control alongside automated tissue masking (Otsu's method [5]). The tissue mask was generated from a low-resolution thumbnail (10X smaller in both spacial dimensions than the original WSI) and applied to the WSI. Tiles with greater than 80% tissue were generated.

After each WSI was segmented and patch generated, the patches were manually examined for quality to remove blurred tiles or other artifacts (eg. scratches, external stains, dirt, etc.).

Seven of the WSI had been manually labeled with regions

Classifier Training and Testing Data Split

	Original				Augmentation			
	Training		Testing		Training		Testing	
	Normal	Rejection	Normal	Rejection	Normal	Rejection	Normal	Rejection
Tile Count	900	18	129	8	900	18 + 47 Synthetic	129	8
Patient Count	3	5	3	1	3	5	3	1

Fig. 2. Breakdown of classifier training data. A VGG-19 classifier was trained on 18 rejection and 900 nonrejection tiles from 5 and 3 subjects respectively. The testing set consisted of 8 rejection tiles from 1 patient and 129 nonrejection tiles from 3 patients. Meanwhile, the augmented training set is 65 tiles from the 5 patients plus synthetic rejection tiles from our GAN.

of rejection. These annotations provided gave us the specific WSI annotated as well as the coordinate regions of the specific rejection regions. To create specific rejection tiles that could be used for IIG, the same patch generation approach as above was implemented to extract tiles from annotated regions.

C. Progressive Generative Adversarial Network

Our progressive GAN structure consists of seven convolutional layers to generate synthetic tile images (Fig. 1). Note that the Discriminator and Generator have mirrored structures, with the key difference being input dimension of the Generator (1x1x512 latent vector), while the output of the Discriminator is a 2-class decision layer for binary classification of synthetic vs. nonsynthetic images. To avoid model collapse and excessive hyperparameter optimization, our model uses WGANGP as its loss function [4], as its been shown to increase the stability of GAN networks.

The model begins with a single, 4x4 layer. After 96,000 epochs of training, the next layer (8x8 resolution) is smoothly added. The new layer is treated as a residual block as in [9], beginning with an alpha value weight of 0 and linearly increasing to 1, while the direct output from the previous layer is weighed as 1 – alpha, therefore slowly fading to 0. The result is a smooth introduction of the new layer to the GAN. This process is repeated for each layer, doubling the resolution of the output image until the desired resolution is reached.

D. Inspirational Image Generation

Pathology images, due to their rarity and small size, are difficult to use for training conventional GAN networks. To overcome this challenge, our progressive GAN was trained on both normal and nonrejection tissue. We then utilized IIG as a method to feed in our limited number of rejection tiles and generate a multitude of similar, rejection-like tiles. Fig. 3 captures the structure of IIG. After completing training of the progressive GAN, its layers are frozen, and we add an optimizer and feature extractor (Fig. 3). VGG-19 was used as the feature extractor, and the loss was calculated as in [11]. The random input latent vector was optimized using gradient descent over 3000 epochs to produce an output image most similar to the reference image.

Discrete one plus one (DOPO) [1] was used as the model optimizer as it produced the most reliably realistic looking synthetic imaging compared to others tested (CMA, DE, OnePlusOne, PortfolioDiscreteOnePlusOne, PSO, TwoPointsDE).

E. Classifier

To evaluate the performance of synthetic image augmentation on rejection detection using our GAN system, we trained three VGG-19 classifiers to differentiate between rejection and nonrejection tissue tiles. The first classifier was trained purely on the real dataset with no data augmentation, while the second classifier used the real dataset with standard data augmentation techniques (random flipping, rotating, resizing, and cropping). The third classifier had real rejection training tiles augmented

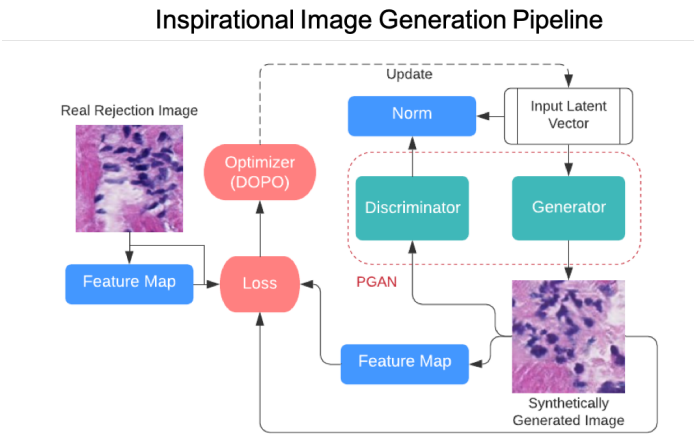

Fig. 3. Pipeline of IIG used to create rejection tiles. IIG would take in a reference rejection image and optimize the latent vector into the pretrained GAN to produce similar, yet synthetic images

with synthetic rejection tiles from our GAN + IIG pipeline in addition to standard augmentation techniques. Hyperparameters were then optimized for both models including the training-validation split, learning rate, and number of epochs.

IV. Results

To assess the performance of the GAN in creating rejection tiles as realistic training data, the aforementioned classifiers were used and confusion matrices generated (Fig. 4). The classifier with no data augmentation achieved an MCC score of 0.411. Once augmented with standard augmentation techniques and retrained, the classifier MCC score rose to 0.557 (Fig. 4b). We then augmented the rejection training set with our synthetic rejection tiles and retrained the classifier, achieving a final MCC score of 0.790 (Fig. 4c).

A heatmap was generated to illustrate classifier results across an annotated WSI. (Fig. 5). The rejection regions are being correctly identified by the classifier indicating high model sensitivity for rejection. It is also important to note that other tiles from the WSI are also being flagged as being potential rejections, indicating potentially low sensitivity for rejection detection.

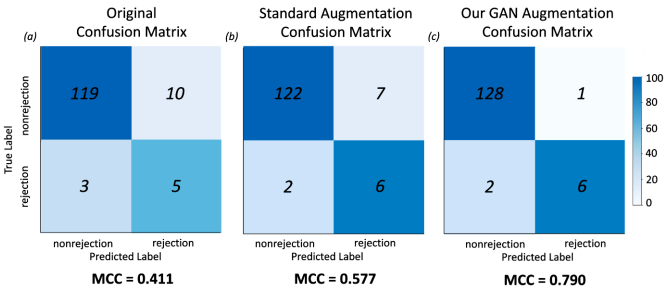

Fig. 4. Confusion matrices and MCC scores of the VGG-19 classifier on the testing set when trained on the (a) original dataset consisting of only real tiles (b) original dataset plus standard augmentation techniques (c) original dataset plus our synthetic rejection tiles. The raw values are shown while the background color has been normalized to the percentage.

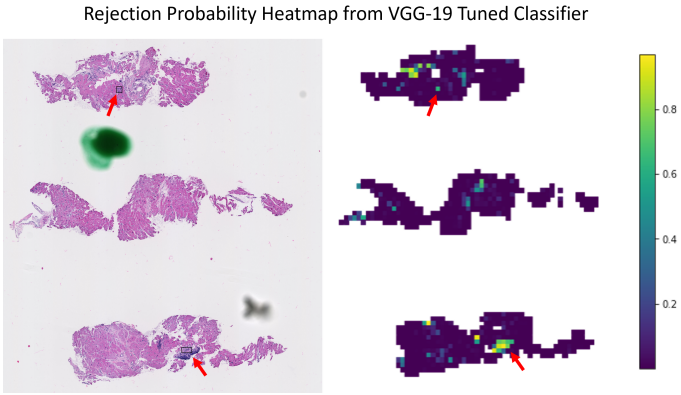

Fig. 5. Rejection WSI with its corresponding heatmap. Each tile classified by the augmented data-trained VGG-19 classifier. There are two definitive areas of rejection that are annotated on the WSI and the classifier was able to successfully identify both of them. The classifier also identified other tiles that it deemed to be rejection regions.

V. Discussion

The synthetic rejection tiles created by the GAN were able to improve the training of a classifier differentiating between rejection and nonrejection tissue tiles. These tiles resulted in an MCC score improvement from 0.411 to 0.790. Since the annotated regions contain many cells, not all of which show rejection signs, our classifier is biased towards high sensitivity for rejection detection. For future work we would like to either change the methodology so that the rejection tiles are created with additional rejection annotation examples to limit this effect.

On the heat map, we noticed that there were regions that the classifier had labeled as potential rejection. Since some of the rejection tiles that were fed into the classifier for training might have contained nonrejection signal, we would expect to see more tiles from the WSIs that might be labeled as rejection even if they were not given a manual annotation. This is acceptable in our scenario as the main goal would be to signal potential regions of interest for the clinician to examine. With the emphasis of sensitivity over specificity, we would rather falsely flag potential regions of rejection than not identify true rejection regions.

VI. Conclusion

Our goal was to augment a severely limited training dataset for deep learning classification models for heart transplant rejection detection using a novel GAN approach. We developed a progressive GAN that generates realistic synthetic myocardial tissue images. IIG was then used to create synthetic myocardial rejection tiles from single rejection tissue tiles. We show that data augmentation with these synthetic images significantly improved rejection classification performance. Heart transplant rejection data augmentation through a progressive GAN may eventually serve as a clinical support tool that results in improved heart transplant patient outcomes.

References

[1] J. Cai and G. Thierauf, “Evolution strategies for solving discrete optimization problems,” Advances in Engineering Software, vol. 25, no. 2-3, pp. 177–183, 1996.

[2] A. E. Dooley, L. Tong, S. R. Deshpande, and M. D. Wang, ”Prediction of heart transplant rejection using histopathological whole-slide imaging.” In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 251-254, 2018.

[3] F. Giuste, M. Venkatesan, C. Zhao, L. Tong, Y. Zhu, S. R. Deshpande, and M. D. Wang, “Automated classification of acute rejection from endomyocardial biopsies,” In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 1–9, 2020.

[4] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of Wasserstein GANs,” 2017.

[5] D. A. Gutman, M. Khalilia, S. Lee, M. Nalisnik, Z. Mullen, J. Beezley, D. R. Chittajallu, D. Manthey, and L. A. D. Cooper, “The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research,” Cancer Research, vol. 77, no. 21, Nov. 2017.

[6] S. Hoo-Chang, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. Andriole, and M. Michalski, “Medical image synthesis for data augmentation and anonymization using generative adversarial networks,” In International workshop on Simulation and Synthesis in Medical Imaging, pp. 1–11, 2018.

[7] L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta, and J. H. Saltz, “Unsupervised histopathology image synthesis,” 2017.

[8] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, “HistoQC: An open-source quality control tool for digital pathology slides,” JCO Clinical Cancer Informatics, no. 3, pp. 1–7, 2019.

[9] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” 2017.

[10] S. Liu, Z. Shah, A. Sav, C. Russo, S. Berkovsky, Y. Qian, E. Coiera, and A. Di Ieva, “Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning,” Scientific Reports, vol. 10, no. 1, May 2020.

[11] B. Roziere, M. Riviere, O. Teytaud, J. Rapin, Y. LeCun, and C. Couprie, “Inspirational adversarial image generation,” IEEE Transactions on Image Processing, vol. 30, pp. 4036–4045, 2021.

[12] L. Tong, R. Hoffman, S. R. Deshpande, and M. D. Wang, “Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout,” 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 3715–3719, 2017.

[13] E. Toro-Trujillo, E. Garcia, A. A. Garcia-Peña, O. M. Muñoz-Velandia, and A. Mariño, “Factors related to the acute cellular rejection during the first year after heart transplant,” Transplantation Proceedings, vol. 50, no. 10, pp. 3715–3719, Dec. 2018.

[14] Y. Xue, J. Ye, Q. Zhou, L. R. Long, S. Antani, Z. Xue, C. Cornwell, R. Zaino, K. C. Cheng, and X. Huang, “Selective synthetic augmentation with HistoGAN for improved histopathology image classification,” Medical Image Analysis, vol. 67, p. 101816, Jan. 2021.

[15] Y. Xue, Q. Zhou, J. Ye, L. R. Long, S. Antani, C. Cornwell, Z. Xue, and X. Huang, “Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification,” Lecture Notes in Computer Science, pp. 387–396, 2019.

[16] S. D. Zangwill, S. J. Kindel, J. N. Schroder, D. P. Bichell, S. R. Deshpande, M. A. Wigger, M. E. Richmond, K. R. Knecht, N. A. Gaglianello, E. Pahl, P. M. Simpson, W. T. Mahle, A. T. Mitchell, and M. E. Mitchell, “Increase in total cell-free DNA correlates with death in adult and pediatric heart transplant recipients: DNA based transplant rejection test (DTRT)-a prospective blinded multicenter NIH/NHLBI funded clinical study,” The Journal of Heart and Lung Transplantation, vol. 38, no. 4, pp. S50–S51, Apr. 2019.

[17] Y. Zhu, L. Tong, S. R. Deshpande, and M. D. Wang, “Improved prediction on heart transplant rejection using convolutional autoencoder and multiple instance learning on whole-slide imaging,” In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 1–9, 2019.