IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 9, 2021, accepted May 23, 2021, date of publication May 28, 2021, date of current version June 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3084597

TilGAN: GAN for Facilitating Tumor-Infiltrating Lymphocyte Pathology Image Synthesis With Improved Image Classification

MONJOY SAHA[1], XIAOYUAN GUO[2], AND ASHISH SHARMA[1]
[1]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322, USA
[2]Department of Computer Science, Emory University, Atlanta, GA 30332, USA

Corresponding author: Monjoy Saha (monjoybme@gmail.com)

This work was supported by the National Cancer Institute, National Institutes of Health, under Grant U24CA215109.

ABSTRACT Tumor-infiltrating lymphocytes (TILs) act as immune cells against cancer tissues. The manual assessment of TILs is usually erroneous, tedious, costly and subject to inter- and intraobserver variability. Machine learning approaches can solve these issues, but they require a large amount of labeled data for model training, which is expensive and not readily available. In this study, we present an efficient generative adversarial network, *TilGAN*, to generate high-quality synthetic pathology images followed by classification of TIL and non-TIL regions. Our proposed architecture is constructed with a generator network and a discriminator network. The novelty exists in the *TilGAN* architecture, loss functions, and evaluation techniques. Our *TilGAN*-generated images achieved a higher Inception score than the real images (2.90 vs. 2.32, respectively). They also achieved a lower kernel Inception distance (1.44) and a lower Fréchet Inception distance (0.312). It also passed the Turing test performed by experienced pathologists and clinicians. We further extended our evaluation studies and used almost one million synthetic data, generated by *TilGAN*, to train a classification model. Our proposed classification model achieved a 97.83% accuracy, a 97.37% F1-score, and a 97% area under the curve. Our extensive experiments and superior outcomes show the efficiency and effectiveness of our proposed *TilGAN* architecture. This architecture can also be used for other types of images for image synthesis.

INDEX TERMS Digital pathology, deep learning, generative adversarial network, lung cancer, artificial intelligence.

I. INTRODUCTION

Tumor infiltrating lymphocytes (TILs) play a significant role in cancer diagnosis and prognosis [1]. The presence of TILs in different cancer types (such as lung, colon, and breast cancer) signifies improved clinical outcomes and faster response to chemotherapy [2]. Recent evidence has emerged that the infiltration of antitumor type I lymphocytes can improve cancer prognosis [3]. TILs are a special white blood cell that shows a tendency to emigrate towards tumor cells from the bloodstream [4]. TILs comprise mainly T cells, B cells, mononuclear cells, and polymorphonuclear immune cells (such as neutrophils, eosinophils, and basophils) [5]. TILs normally float around tumor cells.

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

As per the World Health Organization and American Cancer Society, lung cancer is one of the most devastating cancers globally, accounting for almost 14% of new cancers in men and 13% of new cancers in women in the United States [6]. It has also been reported that in 2019, lung cancer caused approximately 228,150 new cases (116,440 men and 111,710 women) and 142,670 deaths (76,650 men and 66,020 women) in the United States [7]. For lung cancer prognosis, pathological image analysis is considered the primary and gold standard screening method. For this purpose, pathologists collect a small part of the tissue from the suspected tumor region. Next, the tissues are further processed and stained using different stains, including hematoxylin and eosin (H&E) [8], [9]. Lung cancer pathology images typically contain TILs, tumor cells, mitotic cells, stroma, etc. Under a microscope, TILs appear with round, deep bluish nuclei [10]. The details of the TIL and non-TIL regions are

shown in figure 1. Pathologists follow manual image analysis procedures to analyze the tissue regions. This procedure fully depends on the knowledge of the pathologists. Moreover, it is costly and time consuming.

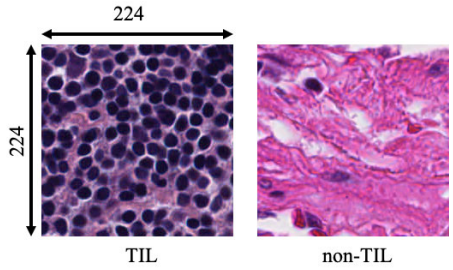


FIGURE 1. Original tumor-infiltrating lymphocyte (TIL) and non-tumor-infiltrating lymphocyte (non-TIL) patches.

Deep learning has shown promising results in image analysis, signal analysis, video analysis, and many more fields [9]. Currently, this method is one of the most popular machine learning approaches and is used to solve many complicated tasks, such as object classification, image segmentation, and risk prediction. However, it has a few disadvantages. The most significant one is that deep learning requires a large amount of data to meet satisfactory performance [11]. However, biomedical data are expensive and not readily available, as approval from the patients and institutional review board are required to use them. Biomedical data may also contain artifacts, noise, etc., which also reduce the total number of data points.

To solve the data availability problem, the authors of [12] proposed a generative adversarial network (GAN) for natural image synthesis. The GAN comprises mainly a generator network/model and a discriminator network/model. The generator network generates synthetic/fake data, which looks like real data, while the discriminator checks the quality of the synthetic data. Categorically, the generator network learns from latent space, and the discriminator differentiates the real and synthetic data distributions [13]. The generator attempts to fool the discriminator by increasing the generator loss [14]. In this study, we present an efficient generative adversarial network, *TilGAN*, to generate high-quality synthetic pathology data of TIL and non-TIL regions to mitigate data imbalances, to improve the classification accuracy, and finally to assist pathologists and clinicians in their decision-making processes. The main novelty exists in *TilGAN* architecture, loss functions, and evaluation techniques.

This manuscript has five sections. Section I introduces the work, and Section II discusses related work. In Section III, the materials and methods are discussed. Sections IV and V discuss the results and present the discussion and conclusion, respectively.

II. RELATED WORK

A GAN, an unsupervised method, is used to generate millions of synthetic data, which resembles the real dataset [15]. Traditional generative models follow the rules of explicit approximation inference and Markov fields, but GANs do not follow this rule. The generative network of GAN produces high-quality fake data to mislead the discriminator. The training process of the GAN ends when a Nash equilibrium from game theory is reached [16]. Hence, the GAN learning process is considered a minimum-maximum optimization problem.

Initially, a GAN was developed for natural image synthesis [12], but gradually, the default architecture was changed to improve the synthetic image quality and to solve other data processing issues, including color enhancement [34], image translation [35], [36], nuclei segmentation [37], [38], cell-level visual representation [39], and image classification [40]. Various researchers have proposed different cost functions for the generator and discriminator networks to improve the quality of synthetic images, such as relativistic GAN, hinge GAN [41], relativistic average GAN [42], and Wasserstein GAN [43]. The main difference between the standard GAN and the modified GANs is that the standard GAN tries to prove that the input data are real, whereas modified GANs measure the probability that generated data are less realistic than the real data (or vice versa). With a standard GAN, the discriminator squeezes the output into two ends, i.e., 0 or 1. Modified GANs measure the distances or differences between fake and real images [44]. When the discriminator reaches an optimum level, gradients vanish. Many new GAN architectures have been proposed for natural and biomedical image synthesis. Cycle-consistency GANs are one of the most common GAN architectures and was designed for biomedical image synthesis [45], image-to-image translation [46], etc.

The authors of [47] proposed a GAN architecture for stain transfer or stain normalization. Their architecture was trained with a multiobjective cost function to learn image-specific color transformations and dataset-specific staining properties. StainGAN [34] and InfoGAN [48] were also used for color normalization on WSIs in different studies. Pathology GAN was proposed for pathology image synthesis [49] with a Fréchet Inception distance of 16.65. This architecture was developed using BigGAN as the baseline architecture [50]. A GAN was also applied to radiology image synthesis and translation. The authors of [51] suggested an edge-aware GAN [51] for MRI image synthesis. Task-driven GAN was proposed in [52] for X-ray image synthesis. The authors of [53] proposed a GAN for computed tomography (CT) to magnetic resonance image (MRI) data synthesis and translation. A conditional GAN was used for PET image synthesis [54]. A deep convolutional GAN (DCGAN) was recommended for image synthesis and the detection of liver cancer on X-ray and CT images [55]. Table 1 refers to the characterization of existing GAN architectures. Here, we have summarized a few GAN architectures that are used for image synthesis, image translation, color normalization, etc.

In this manuscript, we perform image synthesis with *TilGAN*, which is constructed using different baseline

TABLE 1. Characterization of existing GAN architectures.

Purpose	*GAN architectures*
Image synthesis	DCGAN [17], [18]; LostGAN [19]; StackGAN [20]; Standard GAN [21]; Cycle-consistency GAN [22]; Conditional GAN [23]
Image translation	MedGAN [24]; Cycle-consistency GAN [25]; Conditional GAN [26]
Image conversion	Context-aware GAN [27]; Cycle-consistency GAN [28]
Image enhancement and color normalization	Conditional GAN [29]; Cycle-consistency GAN [30]; Conditional GAN [31]
Style transferring	StyleGAN [32]; Conditional GAN [33]

architectures, such as Pathology GAN [49], BigGAN [50], a cycle-consistency GAN [56], and a relativistic average GAN [42].

Pathology images show important information, and small changes in the tissue characteristics may result in a wrong diagnosis and patient death. Therefore, it is a very challenging task to maintain the real image characteristics of synthetic images. Existing GAN architectures generate TIL and non-TIL patches, but our proposed network shows improved results. We targeted preserving real image features such as image appearance, chromatin information, stain colors, and tissue contents.

In summary, the novel technical contributions of this study can be summarized as follows:

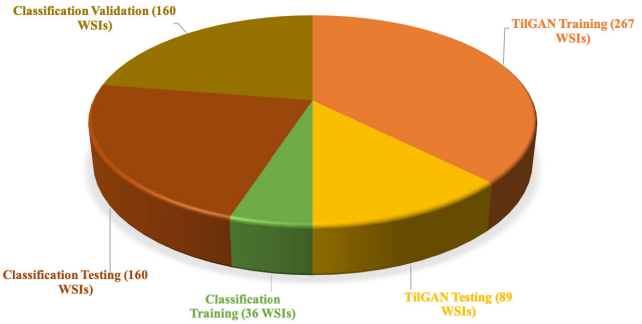
- The most important contribution of this study lies in the architecture of *TilGAN*. Due to its novel architecture, *TilGAN* generates millions of high-quality, clinically significant TIL patches.
- Second, we propose a modified version of the relativistic average cost function to preserve important pathological signatures.
- Third, to our knowledge, this is the first report to propose a GAN that specifically aims to generate TIL and non-TIL patches.
- Fourth, the generated synthetic images are used for classification model training.

The detailed method, along with the results, will be discussed in the subsequent sections.

III. MATERIALS AND METHODS

A. DATASET

In total, 712 H&E stained WSIs of lung cancer (356 adenocarcinomas and 356 squamous cell carcinomas) were collected from The Cancer Genome Atlas data repository (https://tcga-data.nci.nih.gov/tcga/). This is a public repository, and the data are freely available for research. For our study, the collected data were equally split into two sets, with zero overlaps. One half of the data, i.e., 356 WSIs (178 adenocarcinomas and 178 squamous cell carcinomas), was used for *TilGAN*, and the other half was used for classification purposes. Out of the 356 WSIs, we used 75% (267 WSIs) for training and 25% (89 WSIs) for testing of the *TilGAN* architecture. The ground truths were generated by experts using HistomicsTK (https://digitalslidearchive.github.io/HistomicsTK/). To train our classification model, we used one million high-quality synthetic images generated by *TilGAN* of size 224 × 224 pixels and 10% (i.e., 36 WSIs out of the remaining 356 WSIs) real labeled data. The rest of the 320 WSIs were split into testing (50%) and validation sets (50%) to evaluate the classification model. In the figure 2, the WSIs distribution chart has been shown.



N.B: For classification training, almost one million TilGAN generated synthetic data was used along with 36 WSIs.

FIGURE 2. Whole slide images (WSIs) distribution chart.

B. TilGAN METHODOLOGY

We developed the *TilGAN* architecture for the synthesis of TIL and non-TIL pathology images of size 224 × 224 pixels. This network was trained from scratch. We adopted a supervised learning strategy that uses hand-labeled images. There are many GAN architectures available for natural image synthesis, but few of them have been used for pathology image generation. Pathology images carry essential clinical features about cell nuclei, stroma, mitosis, lymphocytes, etc. Hence, image synthesis using pathology data requires special skills. Small changes in the visual appearance of nuclei, lymphocytes, etc., may change the clinical meaning. The workflow diagram of our proposed *TilGAN* architecture is shown in figure 3.

1) *TilGAN* ARCHITECTURE DETAILS

The *TilGAN* architecture comprised a generator network G_{NET} and a discriminator network D_{NET}. The input of the generator was randomly chosen from the annotated real TIL and non-TIL patches of size 224 × 224 pixels. The output of the generator was synthetic TIL and non-TIL patches. The generator was defined as a mapping function z to learn the generator's distribution over the data y. The discriminator, D_{NET}, showed the probability that y was more realistic than the generator's distribution. The generator network of *TilGAN* comprised six convolutional layers, five up-convolution layers, and two dense layers. The up-convolution, is obtained by a transposed convolution, operation increased the height and width of the feature maps by two. The discriminator of *TilGAN* was formed using six convolutional layers, six down-convolution layers, and two dense layers. The down-convolution, general convolution, operation decreased the height and width of the feature maps by two. This design helped the model learn the features from real images efficiently and effectively. Different learning rates were used

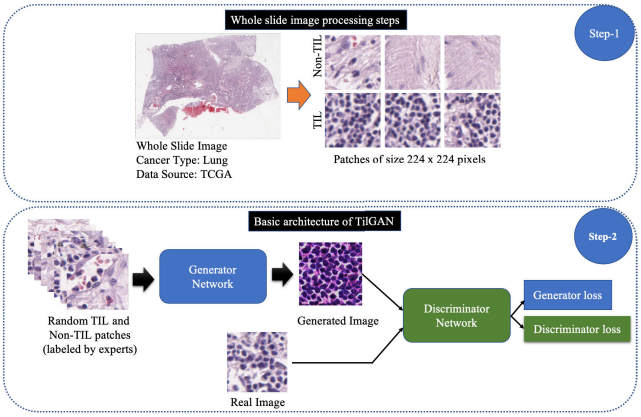


FIGURE 3. The workflow diagram of our proposed *TilGAN* architecture. At step 1, H&E-stained WSIs were processed to extract TIL and non-TIL patches of size 224 × 224 pixels. At step 2, the input of the generator network randomly selected real TIL and non-TIL patches, and the output of the generator was synthetic TIL and non-TIL patches of the same size. The inputs of the discriminator were real and synthetic TIL and non-TIL patches. The discriminator network was used to discriminate real and fake TIL and non-TIL cases. The training of the *TilGAN* model was performed twice, as there were two types of data: TIL and non-TIL data.

for the generator and discriminator networks. The overfitting issue was tracked by incorporating more data and varying the dropout layers. During training, we set the dropout value to 0.5. The rectified linear unit (ReLu) activation function was used after each convolution layer.

The detailed architecture of our proposed *TilGAN* model is shown in table 2.

TABLE 2. Proposed *TilGAN* architecture.

TilGAN Generator Network						
No. of layers	***Layer type***	***Channels***	***Filter Size***	***Stride***	***Padding***	***Output Shape***
14	Generated Image	3	3	1	SAME	(?,224,224,3)
13	Convolution	32	2	1		(?,224,224,32)
12	Up-Convolution	32	1	2		(?,224,224,32)
11	Convolution	64	2	1		(?112,112,64)
10	Up-Convolution	64	1	2		(?,112,112,64)
9	Convolution	128	2	1		(?,56,56,128)
8	Up-Convolution	128	1	2		(?,56,56,128)
7	Convolution	256	2	1		(?,28,28,256)
6	Up-Convolution	256	1	2		(?,28,28,256)
5	Convolution	512	2	1		(?,14,14,512)
4	Up-Convolution	512	1	2		(?,14,14,512)
3	Convolution	256	2	1		(?,7,7,256)
2	Dens Layer	12544	-	-	-	(?,12544)
1	Dens Layer	1024	-	-	-	(?,1024)
TilGAN Discriminator Netwok						
No. of layers	***Layer type***	***Channels***	***Filter Size***	***Stride***	***Padding***	***Output Shape***
1	Convolution	3	2	1	SAME	(?,224,224,3)
2	Down-Convolution	16	4	2		(?,112,112,16)
3	Convolution	16	2	1		(?,112,112,16)
4	Down-Convolution	32	4	2		(?,56,56,32)
5	Convolution	32	2	1		(?,56,56,32)
6	Down-Convolution	64	4	2		(?,28,28,64)
7	Convolution	64	2	1		(?,28,28,64)
8	Down-Convolution	128	4	2		(?,14,14,128)
9	Convolution	128	2	1		(?,14,14,128)
10	Down-Convolution	256	4	2		(?,7,7,256)
11	Convolution	256	2	1		(?,7,7,256)
12	Down-Convolution	512	4	2		(?,4,4,512)
13	Dens Layer	1024	-	-	-	(?,1024)
14	Dens Layer	1	-	-	-	(?,1)

2) MODIFIED LOSS FUNCTION FOR THE *TilGAN* ARCHITECTURE

Pathology images possess distinct types of textural, color, and morphological features, which are linked with the patient's diagnosis and prognosis. Hence, it is essential to handle these types of data separately, unlike other nonclinical data. The existing loss functions of GAN architectures generated TIL and non-TIL patches, but the clinical features of those images were not consistent. Hence, we developed a modified version of the relativistic average loss function to solve these issues. We used the modified relativistic average cost function for both networks (generator and discriminator). The fundamental theory of the modified relativistic average loss function originates from the binary cross-entropy loss function [12] as follows:

$$Loss(\widehat{O}_d, O_d) = O_d \cdot \log \widehat{O}_d + (1 - O_d) \cdot \log(1 - \widehat{O}_d) \quad (1)$$

Here, O_d and $\widehat{O}_d$ denote the original and reconstructed data, respectively. When training the discriminator, the labeled data from the original data distribution $P_{image}(y)$ are $O_d = 1$ (when the data are real) and $\widehat{O}_d = D_{NET}(y)$. Substituting these values into equation 1, we obtain

$$\begin{aligned} Loss(D_{NET}(y), 1) &= 1 \cdot \log(D_{NET}(y)) + (1 - 1) \\ &\quad \cdot \log(1 - D_{NET}(y)) \\ \Rightarrow Loss(D_{NET}(y), 1) &= \log(D_{NET}(y)) \end{aligned} \quad (2)$$

When the data are fake, the values of O_d and $\widehat{O}_d$ will be 0 and $D_{NET}(G_{NET}(z))$, respectively. Substituting these values into equation 1, we obtain

$$\begin{aligned} Loss(D_{NET}(G_{NET}(z)), 0) &= 0 \cdot \log D_{NET}(G_{NET}(z)) \\ &\quad + (1 - 0) \cdot \log \\ &\quad \times (1 - D_{NET}(G_{NET}(z))) \\ \Rightarrow Loss(D_{NET}(G_{NET}(z)), 0) &= \log(1 - D_{NET} \\ &\quad (G_{NET}(z))) \end{aligned} \quad (3)$$

The main purpose of the discriminator D_{NET} is to distinguish real and fake images. Hence, equations 2 and 3 should be maximized. Next, the discriminator loss will be as follows:

$$Loss^{(D_{NET})} = \max[\log D_{NET}(y) + \log(1 - D_{NET}(G_{NET}(z)))] \quad (4)$$

The final generator G_{NET} loss will be

$$Loss^{(G_{NET})} = \min[\log D_{NET}(y) + \log(1 - D_{NET} (G_{NET}(z)))] \quad (5)$$

If we combine equations 4 and 5, we obtain equation 6. However, equation 6 is valid only for a single data point.

$$\begin{aligned} Loss = \min_{G_{NET}} \max_{D_{NET}} &[\log D_{NET}(y) \\ &+ \log(1 - D_{NET}(G_{NET}(z)))] \end{aligned} \quad (6)$$

To consider the entire large amount of data, equation 6 can be modified as below: [57]–[59],

$$\begin{aligned} \min_{G_{NET}} \max_{D_{NET}} & V(G_{NET}, D_{NET}) \\ &= E_{y \sim P_{image}(y)}[\log D_{NET}(y)] + E_{z \sim P_z(z)} \\ &\quad \times [\log(1 - D_{NET}(G_{NET}(z)))] \end{aligned} \quad (7)$$

The loss functions of a standard GAN can be classified into saturating and non-saturating loss functions [42]. Equation 7 is an example of a non-saturating loss function. In the case of saturating loss, the equation for the discriminator will be

$$Loss^{(D_{NET})} = -E_{y\sim P_{image}(y)}[\log D_{NET}(y)] - E_{z\sim P_z(z)} [\log(1 - D_{NET}(G_{NET}(z)))] \quad (8)$$

In the standard GAN, $D_{NET}(y)$ has been represented as a $D_{NET}(y) = sigmoid(C(y))$ [60], [61]. Here, $C(y)$ determines the possibility of having real or fake data. Hence, it is also known as a critic or non-transformed discriminator output. If the value of $C(y)$ is negative, then the input data are fake, and vice versa. After substituting the value of $D_{NET}(y)$ into equation 8, we obtain

$$Loss^{(D_{NET})} = -E_{y\sim P_{image}(y)}[\log(sigmoid(C(y))] - E_{z\sim P_z(z)}[\log(1 - D_{NET}(G_{NET}(z)))] \quad (9)$$

With a relativistic standard GAN, we compute the distance, which depends on the real and fake data distribution. Hence, $D_{NET}(y)$ will change to $D_{NET}(y_r, y_f) = sigmoid(D_{NET}(y_r) - D_{NET}(y_f))$. Here, r and f indicate real and fake data, respectively. Now, the discriminator loss will be:

$$Loss^{(D_{NET})} = -E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[\log(sigmoid(D_{NET}(y_r) - D_{NET}(y_f)))] \quad (10)$$

and the generator loss will be:

$$Loss^{(G_{NET})} = -E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[\log(sigmoid(D_{NET}(y_f) - D_{NET}(y_r)))] \quad (11)$$

Here, $D_{NET}(y_r) = D_{NET}(y_f) = 0.5$ has been set as an optimal point [61]. Equation 7 can also be generalized as

$$Loss^{(D_{NET})} = E_{y_r\sim\mathbb{R}}[f_1(D_{NET}(y_r))] + E_{z\sim\mathbb{R}_z} [f_2 D_{NET}(G_{NET}(z)))] \quad (12)$$

$$Loss^{(G_{NET})} = E_{y_r\sim\mathbb{R}}[g_1(D_{NET}(y_r))] + E_{z\sim\mathbb{R}_z} [g_2 D_{NET}(G_{NET}(z)))] \quad (13)$$

Here, functions f and g map a scalar input to another scalar. The corresponding relativistic cost function will be as follows:

$$Loss^{(D_{NET})} = E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[f_1(D_{NET}(y_r) - D_{NET}(y_f))] + E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[f_2(D_{NET}(y_f) - D_{NET}(y_r))] \quad (14)$$

$$Loss^{(G_{NET})} = E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[g_1(D_{NET}(y_r) - D_{NET}(y_f))] + + E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[g_2(D_{NET}(y_f) - D_{NET}(y_r))] \quad (15)$$

From equations 14 and 15 above, we can say that $f_1(D_{NET}(y_r) - D_{NET}(y_f)) = f_2 - (D_{NET}(y_f) - D_{NET}(y_r))$. Moreover, in the case of non-saturating loss, $f_2(D_{NET}(y_f) - D_{NET}(y_r)) = g_1(D_{NET}(y_r) - D_{NET}(y_f))$, and $g_2(D_{NET}(y_f) - D_{NET}(y_r)) = f_1(D_{NET}(y_r) - D_{NET}(y_f))$. Based on the above properties, we can further simplify equations 14 and 15 as:

$$Loss^{(D_{NET})} = E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[f_1(D_{NET}(y_r) - D_{NET}(y_f))] \quad (16)$$

$$Loss^{(G_{NET})} = E_{(y_r,y_f)\sim(\mathbb{R},\mathbb{N})}[f_1(D_{NET}(y_f) - D_{NET}(y_r))] \quad (17)$$

The generic cost functions of the relativistic average GAN for a generator and discriminator can be computed as:

$$Loss^{(D_{NET})} = E_{y_r\sim\mathbb{R}}[f_1(D_{NET}(y_r) - E_{y_f\sim\mathbb{N}}(D_{NET}(y_f)))] + E_{y_f\sim\mathbb{N}}[f_2(D_{NET}(y_f) - E_{y_r\sim\mathbb{R}}(D_{NET}(y_r)))] \quad (18)$$

$$Loss^{(G_{NET})} = E_{y_r\sim\mathbb{R}}[g_1(D_{NET}(y_r) - E_{y_f\sim\mathbb{N}}(D_{NET}(y_f)))] + E_{y_f\sim\mathbb{N}}[g_2(D_{NET}(y_f) - E_{y_r\sim\mathbb{R}}(D_{NET}(y_r)))] \quad (19)$$

In our data, TILs appear round and dark purple with deep bluish nuclei. We aimed to maintain the stain color, tissue contents, morphology, and textural details in our generated images. The relativistic average GAN cost functions did not generate output as expected. Hence, we tweaked the relativistic average GAN as follows:

$$Loss^{(D_{NET})}_{modified} = E_{y_r\sim\mathbb{R}} [sum\left\{f_1(D_{NET}(y_r)) - E_{y_f\sim\mathbb{N}}(D_{NET}(y_f))\right\}] + E_{y_f\sim\mathbb{N}} [sum\left\{f_2(D_{NET}(y_f)) - E_{y_r\sim\mathbb{R}}(D_{NET}(y_r))\right\}] + V(r,f) \quad (20)$$

$$Loss^{(G_{NET})}_{modified} = E_{y_r\sim\mathbb{R}} [sum\left\{g_1(D_{NET}(y_r)) - E_{y_f\sim\mathbb{N}}(D_{NET}(y_f))\right\}] + E_{y_f\sim\mathbb{N}} [sum\left\{g_2(D_{NET}(y_f)) - E_{y_r\sim\mathbb{R}}(D_{NET}(y_r))\right\}] \quad (21)$$

Here, $V(r,f)$ is computed using $\sqrt{\{r * (1-\epsilon) + f * \epsilon\}}$ where $\epsilon = e^{-4}$. Figure 4 shows the results of the relativistic average GAN loss and our proposed loss functions. The results of figure 4(b) are much smoother than the figure 4(a). Moreover, in 4(b) nuclei are easily identifiable. From the results, it is clear that our loss function generates much better result. The training and validation loss graph of *TilGAN* is shown in figure 11.

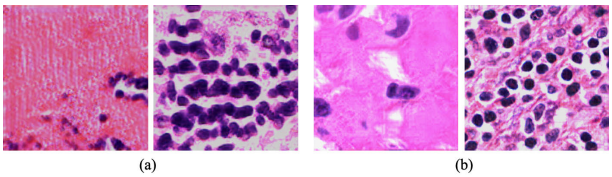


FIGURE 4. (a) Results of relativistic average GAN loss functions; (b) results of proposed loss functions.

3) TilGAN TRAINING AND TESTING PROCEDURE

The *TilGAN* architecture was trained on 267 WSIs and tested on 89 WSIs. For the training of the *TilGAN* model, we did not perform data augmentation because it would generate additional noise with poor-quality images. Hence, our suggestion is to use as many real, high-quality, hand-labeled images as the input of the generator. We set the *TilGAN* model batch size to 100 and the learning rates of G_{NET} and D_{NET} as to 1e-4 and 1e-5, respectively. The initial weights were standardized to a mean of zero with 0.02 as a standard deviation. We used the Adam optimizer with adaptive momentum. The values of $\beta 1$ and $\beta 2$ were set to 0.5 and 0.99, respectively. We used the TensorFlow framework for the development of *TilGAN*.

C. CLASSIFICATION ARCHITECTURE DETAILS

Classification was performed to verify whether our synthetic images are efficient for discriminating real TIL and non-TIL patches. Our classification architecture, developed using Keras with the TensorFlow backend, was designed using six convolution layers, ReLU, two max-pooling layers, four dense layers, one flattened layer, one dropout layer, and one batch normalization layer. The details of our classification architecture are depicted in table 3.

TABLE 3. Classification architecture.

No. of layers	*Layer type*	*Channels*	*Filter Size*	*Output Shape*
1	Input	3	-	(?,224,224,3)
2	Convolution	180	3	(?,222,222,180)
3	Convolution	160	3	(?,220,220,160)
4	Max-pooling	160	5	(?,44,44,160)
5	Convolution	160	3	(?,42,42,160)
6	Convolution	140	3	(?,40,40,140)
7	Convolution	100	3	(?,38,38,100)
8	Convolution	50	3	(?,36,36,50)
9	Max-pooling	50	5	(?,7,7,50)
10	Flatten	-	-	(?,2450)
11	Dense	-	-	(?,180)
12	Dense	-	-	(?,100)
13	Dense	-	-	(?,50)
14	Dropout	-	-	(?,50)
15	Batch Normalization	-	-	(?,50)
16	Dense	-	-	(?,1)

1) CLASSIFICATION MODEL TRAINING, TESTING, AND VALIDATION PROCEDURE

Figure 5 shows the classification model workflow. For the classification, we used one million high-quality synthetic image patches of size 224×224 pixels, which were generated by *TilGAN*. We added only 36 WSIs out of 356 WSIs with *TilGAN*-generated images for better classification performance. The rest of the WSIs were split into testing and validation sets to evaluate the classification model. For our classification algorithm, we used a sigmoid classifier and an Adam optimizer. The training parameters were as follows: learning rate as 0.0001, epoch as 50, dropout ratio as 0.5, and loss function as binary cross-entropy. We used rectified linear unit after each convolution layer.

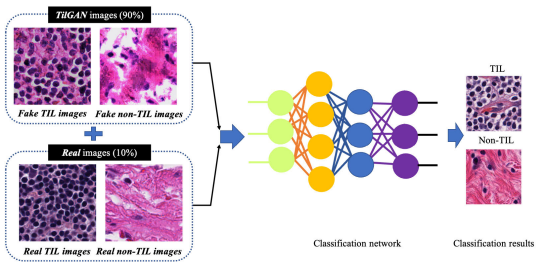


FIGURE 5. Workflow diagram of the classification model.

D. EVALUATION METRICS

1) EVALUATION METRICS FOR TilGAN-GENERATED IMAGES

For the quantitative evaluation of the images generated by our proposed *TilGAN* model, we used the Inception score (IS) [15], kernel Inception distance (KID) [62], and Fréchet Inception distance (FID) [63]. All the scores were calculated using a pretrained Inception-v3 network [59], [64]. We calculated the IS as follows [59], [65]:

$$IS = exp(\mathbb{E}_{y \sim p_{image}} KL(p(x|y) \parallel p(x))) \tag{22}$$

The marginal class distribution can be evaluated as:

$$p(x) = \int_y p(x|y) p_{image}(y) \tag{23}$$

Here, $y \sim p_{image}$ means that y is an image set of p_{image}. $p(x|y)$ represents the conditional class distribution. KL means KL divergence.

2) EVALUATION METRICS FOR THE CLASSIFICATION MODEL

The classification performances have been measured by classification accuracy, precision, recall, and F1-score [8], [66], [67]. We also computed the confusion matrix and area under the receiver operating characteristic curve.

IV. RESULTS AND DISCUSSION

A. RESULTS OF TilGAN

We evaluated the quality of our proposed *TilGAN*-generated fake images through a clinical evaluation by our experts. They independently classified each image as real or fake from sets of almost 1000 images. A subset of all the real and *TilGAN*-generated fake images are shown in figures 6 and 7, respectively. Over 96% of the *TilGAN*-generated fake images were classified as real images, and all the real images were classified as real. Less than 4% of the *TilGAN*-generated fake images were classified as fake. From this experiment, it is obvious that even for an expert, it is difficult to distinguish *TilGAN*-generated fake images from a mixture of fake and

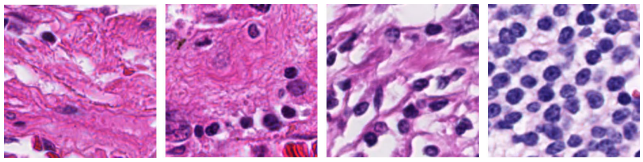


FIGURE 6. Real images of size 224 × 224 pixels.

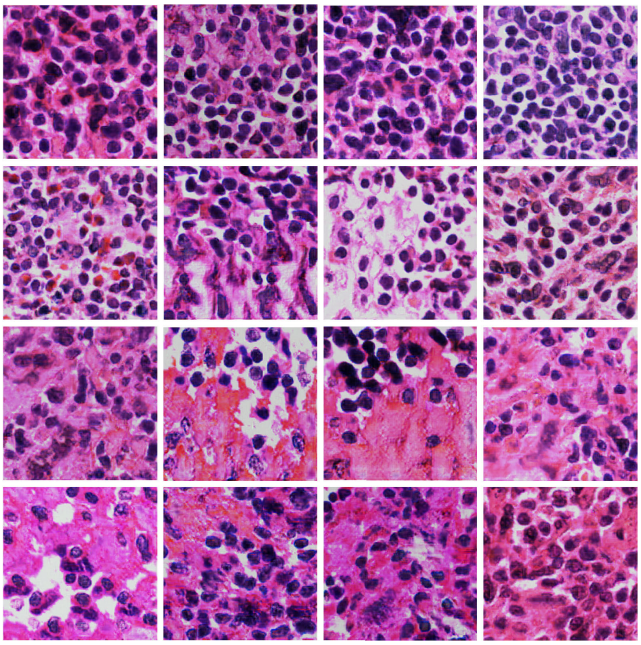


FIGURE 7. ***TilGAN*-generated images of size 224 × 224 pixels.**

real data. This finding means that the *TilGAN* architecture generates high-quality images and maintains the proper stain color with a significant amount of tissue content based on the tumor stage.

Moreover, we also evaluated the quality and diversity of the *TilGAN*-generated fake images by the most popular quantitative evaluation metrics for GANs, i.e., the Inception score, Fréchet Inception distance, and kernel Inception distance. The Inception score was used to evaluate the quality and diversity of the fake images. A high Inception score indicates that the generated fake image contains high-density and clear objects for all classes. However, this scoring technique has a few disadvantages. One of the main disadvantages is that it does not use the statistics of real data. To overcome this issue, we used the Fréchet Inception distance and kernel Inception distance. The Fréchet Inception distance has been used to calculate the distance between Inception feature vectors for fake and real images. The value of the Fréchet Inception distance changes with the image diversity, as it is robust to noise. If a dataset contains many diverse images, the Fréchet Inception distance will be low or closer to zero. On the other hand, if the image diversity between the real and synthetic images decreases, the Fréchet Inception distance will be high. The kernel Inception distance has been used to evaluate the similarity between real and fake images [62]. If the kernel Inception distance is low, the real and fake images are very similar to each other, or it is very hard to distinguish them from a mixture of real and fake images.

The results of the Inception score, Fréchet Inception distance, and kernel Inception distance are shown in table 4. We calculated the Inception score on both real and *TilGAN*-generated fake images because it only uses one kind of image

TABLE 4. Performance comparison between pathology GAN and the proposed method (*TilGAN*) using evaluation metrics (IS, KID, and FID).

Types of data	*IS*	*KID*	*FID*
Real Data	2.32±0.02	-	-
Pathology GAN [49]	2.75±0.012	4.66±0.48	1.07±0.023
TilGAN	***2.90±0.04***	***1.44±0.025***	***0.312±0.001***

at a time. We achieved an Inception score of 2.32±0.02 (mean ± standard deviation) for the real images and an Inception score of 2.90±0.04 (mean ± standard deviation) for the fake images. This finding indicates that the *TilGAN*-generated fake images contain high-density tissues and clear objects and or more diverse. For the Fréchet Inception distance and kernel Inception distance measurements, we used the outputs of the last hidden layer, i.e., the pooling layer, of the same pretrained Inception-v3 network [64]. The Fréchet Inception distance is 0.312, which is very close to zero, and the kernel Inception distance is 1.44±0.025. These two values are lower than the Inception scores of the real and fake data. Undoubtedly, the *TilGAN* generates a more diverse and high-quality dataset, which is almost similar to the real images. The real and fake data distribution is shown using the t-stochastic neighbor embedding (t-SNE) plot in figure 8. This plot gives a good understanding of the visual and color similarities of the generated synthetic images with the real images. The TilGAN generates wide varieties of fake non-TIL patches, which also include some white patches. Hence, some green dots are away from the dense population.

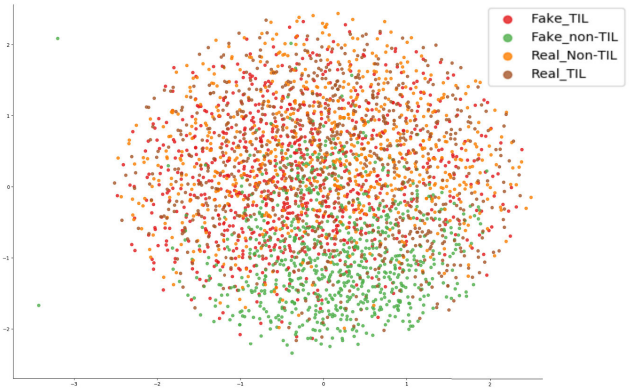


FIGURE 8. t-SNE graph.

B. TIL AND NON-TIL IMAGE CLASSIFICATION RESULTS

In the previous sections, we have shown the results of the qualitative and quantitative analysis of *TilGAN*. In this section, we will show the performance of our classification model, where 90% (i.e., one million) *TilGAN*-generated images and 10% (i.e., 36 WSIs) real hand-labeled images were used for model training for distinguishing real TIL and non-TIL patches. Testing and validation of our trained model was performed using only real images with zero overlap.

The classification model was run for up to 50 epochs, and the model started converging after 41 epochs. The training and validation losses of our classification model are shown in figure 12. A subset of the *TilGAN*-generated synthetic TIL images and non-TIL images are shown in figures 9 and 10,

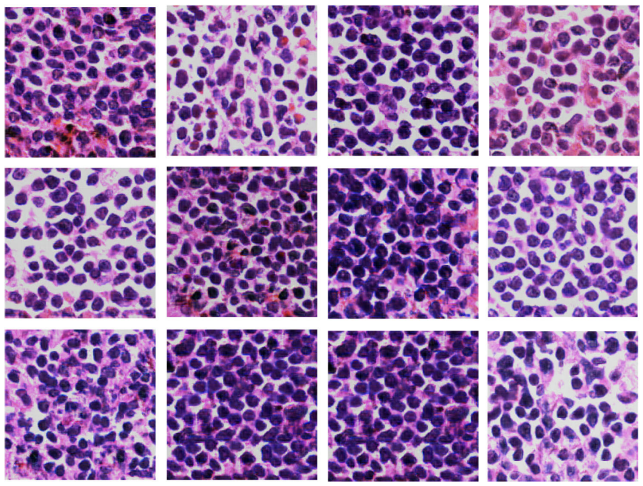


FIGURE 9. ***TilGAN*-generated tumor-infiltrating lymphocyte images of size 224 × 224.**

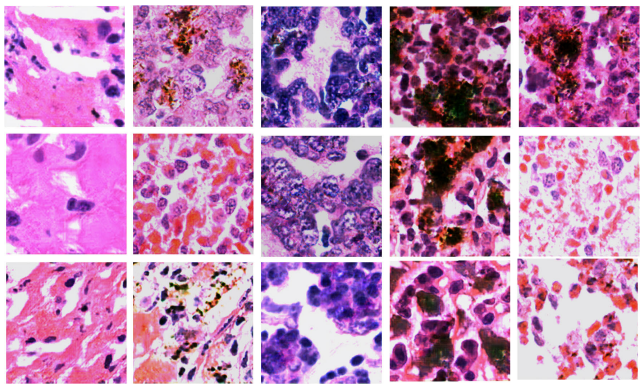


FIGURE 10. ***TilGAN*-generated non-tumor-infiltrating lymphocyte images of size 224 × 224.**

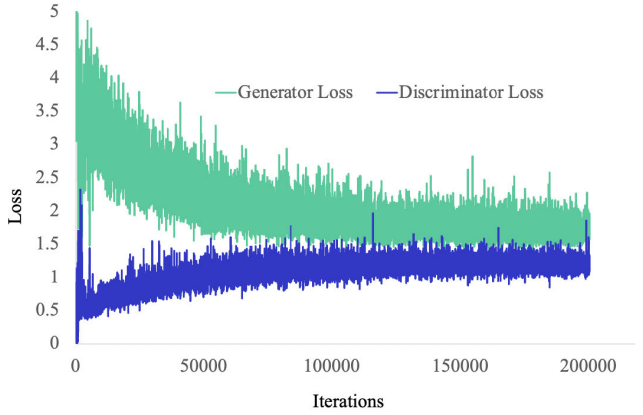


FIGURE 11. **Training and validation loss for the *TilGAN* model.**

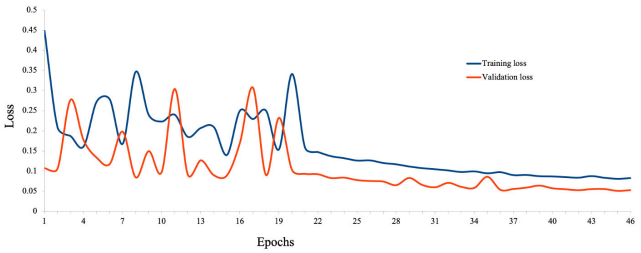


FIGURE 12. **Training and validation loss for the TIL and non-TIL classification model.**

respectively. Figure 15 shows an accuracy plot of the real, fake, and combined outcomes. In this plot, when only fake images were used, we observed the unstable behavior of the model's performance from epochs 16 to 21. However, this behavior is normal for any classification model. This finding indicates that our proposed model is learning, and based on the quality of the batch images, the learning performance varies. We also noticed that after 36 epochs, the accuracy plots of the real, fake, and combined outcomes converged. However, their accuracy levels are different.

The accuracy for the real images is comparatively lower than that of the proposed model (where 90% fake and 10% real images were used). The main reason for this behavior is that we only used a minimal number of real hand-labeled data (i.e., 10%) for model training. Significant changes in the accuracies were not observed when only fake and combined (90% fake and 10% real) images were used for model training.

The training scheme was repeated ten times with different data as per the Monte Carlo cross-validation criteria. Each time, the training and testing dataset was split randomly, but the same principle applies. We achieved an average classification accuracy of 97.83%, an F1-score of 97.37%, a precision of 98.34%, and a recall of 96.49%. Table 5 shows the 10-fold cross-validation results. We also computed the confusion matrix on 18,400 image patches (8750 TIL and 9650 non-TIL patches) of size 224 × 224 pixels. The confusion matrix is shown in figure 16. Figure 17 shows a receiver operating characteristic curve, which has an area under the curve of 97%.

TABLE 5. **Ten-fold cross-validation results.**

Folds	*Accuracy*	*F1-score*	*Precision*	*Recall*
Fold-1	0.98	0.98	0.98	0.98
Fold-2	0.98	0.98	0.98	0.97
Fold-3	0.98	0.98	0.99	0.96
Fold-4	0.97	0.97	0.97	0.97
Fold-5	0.98	0.97	0.99	0.96
Fold-6	0.98	0.98	0.98	0.98
Fold-7	0.97	0.96	0.99	0.93
Fold-8	0.98	0.97	0.99	0.96
Fold-9	0.97	0.97	0.97	0.96
Fold-10	0.98	0.97	0.98	0.97
Average	***0.98***	***0.97***	***0.98***	***0.96***

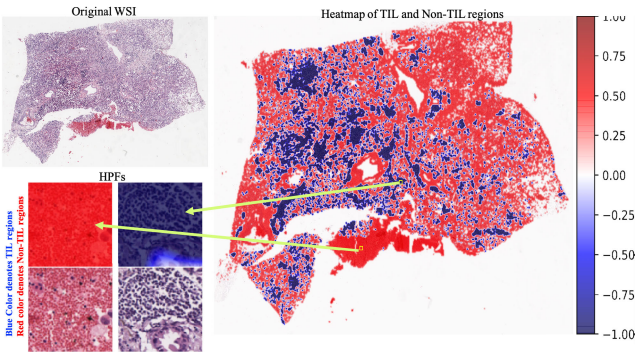


FIGURE 13. TIL classification results on a single whole-slide image.

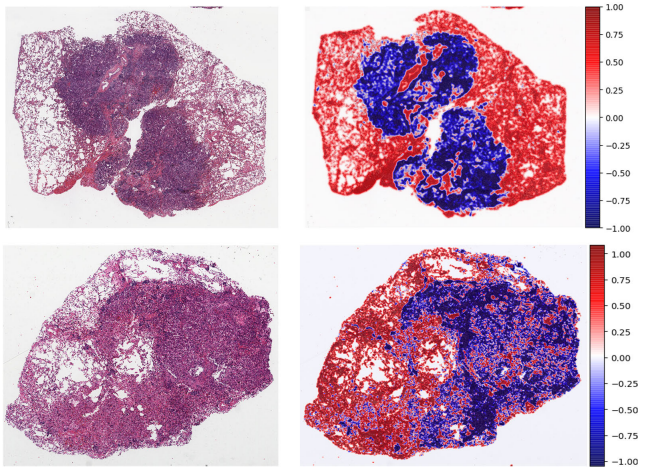


FIGURE 14. TIL classification results.

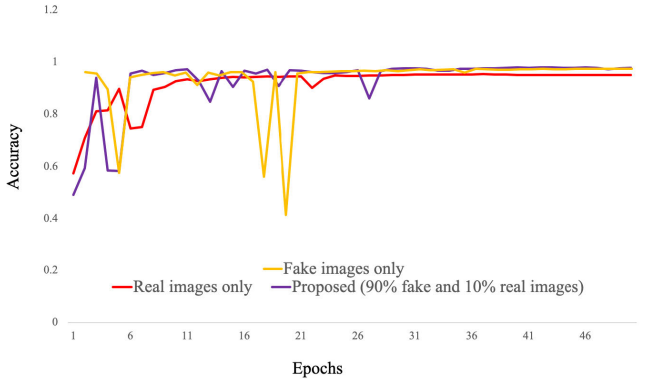


FIGURE 15. Accuracy curves.

From the above results, we can say that our classification model accurately classifies the real TIL and non-TIL patches.

The classification results on the whole-slide pathology images are shown in figures 13 and 14 using a heat map. For heat map generation, the first whole-slide images were tiled into 224 × 224 pixels. Next, the probability score was calculated for each tile using our trained classification model. The probability score determines the probability of having TILs or non-TILs in a specific tile. In our experiment, 0 indicates a

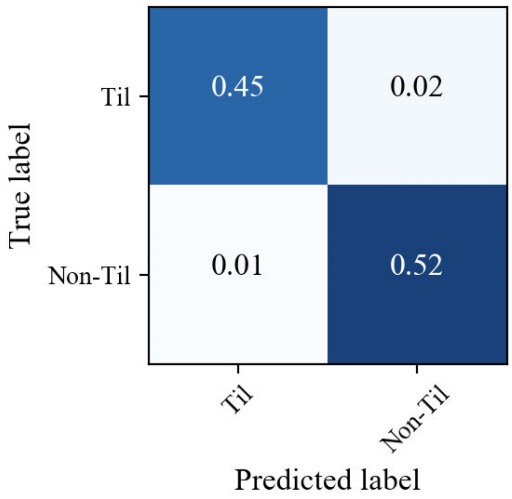


FIGURE 16. Confusion matrix between the predicted label and true label.

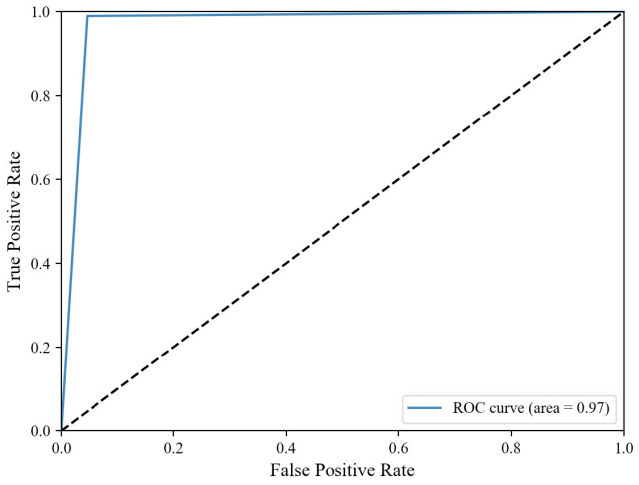


FIGURE 17. Receiver operating characteristic curve.

TIL tile, and 1 indicates a non-TIL tile. When the probability score was close to zero, then the tile was considered a TIL tile, and when the probability score was close to 1, then the tile was considered a non-TIL tile. In figure 13, red and blue regions from the heat map were separately highlighted and matched with the original WSI. The red and blue regions represent the non-TIL and TIL regions, respectively, of the original WSI. In figure 14, we show the heat map representation of the classification scores for two other whole-slide images.

V. CONCLUSION

In this study, we proposed the *TilGAN* model for improving the quality of synthetic pathology images. Our proposed architecture differs from existing GANs mainly because of architecture and loss functions. *TilGAN* does not contain an attention layer, similar to the Pathology GAN architecture [49]. In the *TilGAN* model, the numbers of layers in the generator and discriminator are different. Because of

these properties, we can maintain the quality and quantity of specific types of target objects; in our case, it is TIL in each tile. The generated TIL patches are mostly covered by lymphocytes rather than stroma or other artifacts. Similarly, non-TIL patches are mostly covered by stroma and other artifacts rather than lymphocytes. This phenomenon is obviously a good sign of our architecture. This finding shows that we are not using up our resources on generating low-quality images. Another interesting point is our loss functions, which maintain the features of TIL morphology, texture, and color of real images in the synthetic images. Hence, image normalization, enrichment, and translation are not essential as in the approach proposed by [46]. We properly verified the quality of our synthetic images by the Inception score, kernel Inception distance, and Fréchet Inception distance metrics, which showed promising results. We plotted the t-SNE graph using real and synthetic images, which showed a strong correlation between the real and synthetic images. Next, the generated synthetic images were physically verified by experts. They faced difficulties in distinguishing the real and synthetic images from the mixture of real and synthetic images. The use of one million synthetic images for training the classification model was an additional evaluation measure for the *TilGAN* model. Here, we showed that the *TilGAN*-generated images can efficiently classify real TIL and non-TIL patches with improved accuracy. From the various image verification methods, we proved the usefulness and effectiveness of our proposed *TilGAN* architecture. Therefore, we can say that our approach performs better in generating TIL and non-TIL images than other methods. In the future, this architecture can be used to generate radiology and other non-clinical data.

SOFTWARE AND HARDWARE

Our model was trained using TensorFlow (v 1.14.0) on NVIDIA DGX-1 servers equipped with eight NVIDIA V100 GPUs. As additional software, we used OpenSlide and Python.

DATA AVAILABILITY

In total, 712 H&E stained WSIs of lung cancer (356 adenocarcinomas and 356 squamous cell carcinomas) were collected from The Cancer Genome Atlas data repository (https://tcga-data.nci.nih.gov/tcga/). This is a public repository, and the data are freely available for research.

CODES

The relevant codes will be available for public upon acceptance of this manuscript.

REFERENCES

[1] S. Hendry, R. Salgado, T. Gevaert, P. A. Russell, T. John, B. Thapa, M. Christie, K. Van De Vijver, M. V. Estrada, P. I. Gonzalez-Ericsson, and M. Sanders, "Assessing tumor infiltrating lymphocytes in solid tumors: A practical review for pathologists," *Adv. Anatomic Pathol.*, vol. 24, no. 6, p. 311, 2017.

[2] S. Abousamra, L. Hou, R. Gupta, C. Chen, D. Samaras, T. Kurc, R. Batiste, T. Zhao, S. Kenneth, and J. Saltz, "Learning from thresholds: Fully automated classification of tumor infiltrating lymphocytes for multiple cancer types," 2019, *arXiv:1907.03960*. [Online]. Available: http://arxiv.org/abs/1907.03960

[3] S. E. Stanton and M. L. Disis, "Clinical significance of tumor-infiltrating lymphocytes in breast cancer," *J. ImmunoTherapy Cancer*, vol. 4, no. 1, p. 59, Dec. 2016.

[4] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran, and J. H. Saltz, "Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images," *Pattern Recognit.*, vol. 86, pp. 188–200, Feb. 2019.

[5] L. M. Coussens, L. Zitvogel, and A. K. Palucka, "Neutralizing tumor-promoting chronic inflammation: A magic bullet?" *Science*, vol. 339, no. 6117, pp. 286–291, Jan. 2013.

[6] P. M. de Groot, C. C. Wu, B. W. Carter, and R. F. Munden, "The epidemiology of lung cancer," *Transl. Lung Cancer Res.*, vol. 7, no. 3, p. 220, 2018.

[7] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA, Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[8] M. Saha and C. Chakraborty, "Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2189–2200, May 2018.

[9] M. Saha, C. Chakraborty, and D. Racoceanu, "Efficient deep learning model for mitosis detection using breast histopathology images," *Computerized Med. Imag. Graph.*, vol. 64, pp. 29–40, Mar. 2018.

[10] *Lymphocytes and Microscopy Staining, Observations, Discussion*. Accessed: Apr. 11, 2020. [Online]. Available: https://www.microscopemaster.com/lymphocytes.html

[11] W. Pedrycz and S.-M. Chen, *Deep Learning: Concepts Architectures*. Springer, 2020.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[13] A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, "Medical image computing and computer assisted intervention—MICCAI 2018," in *Proc. MICCAI, Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 11073, Granada, Spain: Springer, Sep. 2018, p. 534.

[14] H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 117–129, Nov. 2017.

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[16] Z. Zhong, J. Li, D. A. Clausi, and A. Wong, "Generative adversarial networks and conditional random fields for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3318–3329, Jul. 2020.

[17] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[18] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 990–994.

[19] W. Sun and T. Wu, "Image synthesis from reconfigurable layout and style," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10531–10540.

[20] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[21] L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta, and J. H. Saltz, "Unsupervised histopathology image synthesis," 2017, *arXiv:1712.05021*. [Online]. Available: http://arxiv.org/abs/1712.05021

[22] D. Ravì, A. B. Szczotka, S. P. Pereira, and T. Vercauteren, "Adversarial training with cycle consistency for unsupervised super-resolution in endomicroscopy," *Med. Image Anal.*, vol. 53, pp. 123–131, Apr. 2019.

[23] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Cukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2375–2388, Oct. 2019.

[24] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "MedGAN: Medical image translation using GANs," *Computerized Med. Imag. Graph.*, vol. 79, Jan. 2020, Art. no. 101684.

[25] J. Song, K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Learning to sketch with shortcut cycle consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 801–810.

[26] E. A. Burlingame, A. A. Margolin, J. W. Gray, and Y. H. Chang, "SHIFT: Speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks," *Proc. SPIE*, vol. 10581, Mar. 2018, Art. no. 1058105.

[27] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Springer, 2017, pp. 417–425.

[28] C. You, W. Cong, M. W. Vannier, P. K. Saha, E. A. Hoffman, G. Wang, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, and Z. Zhang, "CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE)," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 188–203, Jan. 2020.

[29] N. Bayramoglu, M. Kaakinen, L. Eklund, and J. Heikkila, "Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 64–71.

[30] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1602–1612, Apr. 2019.

[31] J. Wang, Y. Zhao, J. H. Noble, and B. M. Dawant, "Conditional generative adversarial networks for metal artifact reduction in CT images of the ear," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Springer, 2018, pp. 3–11.

[32] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[33] H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using GAN for histopathological images," 2017, *arXiv:1710.08543*. [Online]. Available: http://arxiv.org/abs/1710.08543

[34] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Stainan: Stain style transfer for digital histological images," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 953–956.

[35] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2021.

[36] C.-T. Lin, S.-W. Huang, Y.-Y. Wu, and S.-H. Lai, "GAN-based day-to-night image style transfer for nighttime vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 951–963, Feb. 2021.

[37] F. Mahmood, D. Borders, R. J. Chen, G. N. Mckay, K. J. Salimian, A. Baras, and N. J. Durr, "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3257–3267, Nov. 2020.

[38] F. Mahmood, R. Chen, D. Borders, G. N. McKay, K. Salimian, A. Baras, and N. J. Durr, "Adversarial U-Net with spectral normalization for histopathology image segmentation using synthetic data," *Proc. SPIE*, vol. 10956, Mar. 2019, Art. no. 109560N.

[39] B. Hu, Y. Tang, E. I.-C. Chang, Y. Fan, M. Lai, and Y. Xu, "Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1316–1328, May 2019.

[40] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[41] D. Tran, R. Ranganath, and D. Blei, "Hierarchical implicit models and likelihood-free variational inference," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5523–5533.

[42] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*. [Online]. Available: http://arxiv.org/abs/1807.00734

[43] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[44] *GAN—RSGAN & RaGAN (a New Generation of Cost Function*. Accessed: Apr. 11, 2020. [Online]. Available: shorturl.at/kwMS4

[45] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.

[46] J. Wei, A. Suriawinata, L. Vaickus, B. Ren, X. Liu, J. Wei, and S. Hassanpour, "Generative image translation for data augmentation in colorectal histopathology images," 2019, *arXiv:1910.05827*. [Online]. Available: http://arxiv.org/abs/1910.05827

[47] A. Bentaieb and G. Hamarneh, "Adversarial stain transfer for histopathology image analysis," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 792–802, Mar. 2018.

[48] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[49] A. C. Quiros, R. Murray-Smith, and K. Yuan, "PathologyGAN: Learning deep representations of cancer tissue," 2019, *arXiv:1907.02644*. [Online]. Available: http://arxiv.org/abs/1907.02644

[50] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*. [Online]. Available: http://arxiv.org/abs/1809.11096

[51] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1750–1762, Jul. 2019.

[52] Y. Zhang, S. Miao, T. Mansi, and R. Liao, "Unsupervised X-ray image segmentation with task driven generative adversarial networks," *Med. Image Anal.*, vol. 62, May 2020, Art. no. 101664.

[53] J. H. Lee, I. H. Han, D. H. Kim, S. Yu, I. S. Lee, Y. S. Song, S. Joo, C.-B. Jin, and H. Kim, "Spine computed tomography to magnetic resonance image synthesis using generative adversarial networks : A preliminary study," *J. Korean Neurosurgical Soc.*, vol. 63, no. 3, pp. 386–396, May 2020.

[54] H. Kang, J.-S. Park, K. Cho, and D.-Y. Kang, "Visual and quantitative evaluation of amyloid brain PET image synthesis with generative adversarial network," *Appl. Sci.*, vol. 10, no. 7, p. 2628, Apr. 2020.

[55] K. Doman, T. Konishi, and Y. Mekada, "Lesion image synthesis using DCGANs for metastatic liver cancer detection," in *Deep Learning in Medical Image Analysis*. Springer, 2020, pp. 95–106.

[56] L. Gupta, B. M. Klinkhammer, P. Boor, D. Merhof, and M. Gadermayr, "GAN-based image enrichment in digital pathology boosts segmentation accuracy," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Springer, 2019, pp. 631–639.

[57] R. Atienza, *Advanced Deep Learning With Keras: Apply Deep Learning Techniques, Autoencoders, GANs, Variational Autoencoders, Deep Reinforcement Learning, Policy Gradients, and More*. Birmingham, U.K.: Packt Publishing Ltd, 2018.

[58] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," 2019, *arXiv:1906.01529*. [Online]. Available: http://arxiv.org/abs/1906.01529

[59] J.-Y. Baek, Y.-S. Yoo, and S.-H. Bae, "Adversarial learning with knowledge of image classification for improving GANs," *IEEE Access*, vol. 7, pp. 56591–56605, 2019.

[60] S. Martin Arjovsky and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 214–223.

[61] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*. [Online]. Available: http://arxiv.org/abs/1701.04862

[62] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," 2018, *arXiv:1801.01401*. [Online]. Available: http://arxiv.org/abs/1801.01401

[63] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2017, pp. 6626–6637.

[64] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[65] S. Barratt and R. Sharma, "A note on the inception score," 2018, *arXiv:1801.01973*. [Online]. Available: http://arxiv.org/abs/1801.01973
[66] M. Giordano, M. Natale, M. Cornaz, A. Ruffino, D. Bonino, and E. M. Bucci, "IMole, a Web based image retrieval system from biomedical literature," *Electrophoresis*, vol. 34, no. 13, pp. 1965–1968, Jul. 2013.
[67] M. Saha, I. Arun, R. Ahmed, S. Chatterjee, and C. Chakraborty, "HscoreNet: A deep network for estrogen and progesterone scoring using breast IHC images," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107200.

MONJOY SAHA received the Ph.D. degree in machine learning and medical image analysis from the Indian Institute of Technology Kharagpur, India, in 2018. He is currently working as a Research Scientist with the School of Medicine, Emory University, Atlanta, GA, USA. His research interests include AI/machine learning, biomedical data (image and signals) analysis, natural language processing, survival modeling, risk prediction, and computer-aided diagnostic algorithm developments for cancer prognosis.

XIAOYUAN GUO is currently pursuing the Ph.D. degree with the Department of Computer Science, Emory University. Her expertise is in computer vision, medical image processing, and deep learning. Her works focus on AI-aided medical image segmentation and open-world out-of-distribution detection.

ASHISH SHARMA received the Ph.D. degree in computer science from the University of Southern California, Los Angeles, CA, USA, in 2005. From 2006 to 2009, he served as a Postdoctoral Fellow, a Research Scientist, and a Research Assistant Professor with The Ohio State University, Columbus, OH, USA. He is currently an Associate Professor with Emory University, Atlanta, GA, USA. His research interests include AI/machine learning, biomedical data analysis, and software systems.