Deepfakes for Histopathology Images: Myth or Reality?

Nouf Alrasheed
University of Missouri-Kansas City
nalrasheed@mail.umkc.edu

Arun Zachariah
University of Missouri-Columbia
azachariah@mail.missouri.edu

Shivika Prasanna
University of Missouri-Columbia
spn8y@umsystem.edu

Deepthi Rao
University of Missouri-Columbia
raods@health.missouri.edu

Praveen Rao
University of Missouri-Columbia
praveen.rao@missouri.edu

2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) | 978-1-7281-8243-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/AIPR50011.2020.9425033

***Abstract*—Deepfakes have become a major public concern on the Internet as fake images and videos could be used to spread misleading information about a person or an organization. In this paper, we explore if deepfakes can be generated for histopathology images using advances in deep learning. This is because the field of digital pathology is gaining a lot of momentum since the Food and Drug Administration (FDA) approved a few digital pathology systems for primary diagnosis and consultation in the United States. Specifically, we investigate if state-of-the-art generative adversarial networks (GANs) can produce fake histopathology images that can trick an expert pathologist. For our investigation, we used whole slide images (WSIs) hosted by The Cancer Genome Atlas (TCGA). We selected 3 WSIs of colon cancer patients and produced 100,000 patches of 256×256 pixels in size. We trained three popular GANs to generate fake patches of the same size. We then constructed a set of images containing 30 real and 30 fake patches. An expert pathologist reviewed these images and marked them as either real or fake. We observed that the pathologist marked 10 fake patches as real and correctly identified 34 patches (as fake or real). Thirteen patches were incorrectly identified as fake. The pathologist was unsure of 3 fake patches. Interestingly, the fake patches that were correctly identified by the pathologist, had missing morphological features, abrupt background change, pleomorphism, and other incorrect artifacts. Our investigation shows that while certain parts of a histopathology image can be mimicked by existing GANs, the intricacies of the stained tissue and cells cannot be fully captured by them. Unlike radiology, where it is relatively easier to manipulate an image using a GAN, we argue that it is a harder challenge in digital pathology to generate an entire WSI that is fake.**

***Index Terms*—Deepfakes, generative adversarial networks, histopathology images, whole slide imaging, deep learning**

I. Introduction

Deepfakes are manipulated/fake images and videos generated with the help of deep learning techniques. Recently, deepfakes have become a significant public concern as the generated images or videos are hard to detect and could be used to spread misleading information about a person or an organization. The backbone of deepfake is a deep neural network and with proper post-processing, can generate content with a high level of realism. GANs [1] are the key drivers of deepfakes and use deep learning-based generative models containing an encoder and decoder neural networks.

In 2017, the FDA approved a whole slide imaging system of Philips for primary diagnosis [2]. Later in 2019, Leica Biosystems received FDA clearance for its digital pathology system [3]. These are remarkable advances for digital pathology to become mainstream in the US. Whole slide imaging is touted as a disruptive technology in digital pathology. WSIs are gigapixel images of glass slides scanned using digital scanners at near-optical resolution. WSIs enables pathologists to use computing devices to view and analyze histopathology slides. They can also enable pathologists to collaborate with others for consultation purposes as well as for teaching activities. Furthermore, deep learning techniques can be employed for automatic detection and analysis of cellular and morphological features in WSIs. This can enable improved and accurate diagnosis by pathologists for diseases such as cancer.

There has been growing interest in using GANs for real-world applications and medical imaging. GANs have shown their usefulness in synthetic image generation, denoising, reconstruction, and translation. While the benefits of GANs in medical imaging are many, one may wonder if they can be misused to generate fake images that can alter a pathologist's diagnosis. Recently, Mirsky et al. [4] developed a GAN architecture to tamper medical images in radiology. They showed that how an attacker could use deep learning techniques to add/remove evidence of medical conditions from 3D medical scans (e.g., MRI, CT scans). While a large body of work in digital pathology have focused on WSI image analysis [5], [6], storage techniques [7], [8], and use of GANs for synthetic histopathology images [9], none has explored if adversarial attacks (via deepfakes) can be injected in a pathologist's workflow when performing clinical diagnosis solely using WSIs.

Therefore, in this paper, we investigate if deepfakes of histopathology images can be generated using popular GANs that can trick an expert pathologist. The key contributions of our work are as follows:

- For our investigation, we used 3 WSIs of colon cancer patients obtained from TCGA [10]. We produced 100,000 patches of 256×256 pixels in size. We trained StyleGAN [11], StyleGAN2 [12], and PathologyGAN [13] to generate fake patches of the same size. We then constructed a set of images (containing 30 real and 30 fake patches) for evaluation by an expert pathologist.
- We observed that the pathologist marked 10 fake patches as real and correctly identified 34 patches (as fake or real). Thirteen patches were incorrectly identified as fake. Interestingly, the fake patches that were correctly identified by the pathologist, had missing morphological features, abrupt background change, pleomorphism, and other incorrect artifacts.
- Our investigation shows that while certain parts of a histopathology image can be mimicked by existing GANs, the intricacies of the stained tissue and cells cannot be fully captured by them. Therefore, more research is needed in better understanding how adversarial attacks can be accomplished by GANs in digital pathology.

The rest of the paper is organized as follows: Section II gives a background on GANs and provides motivation for our work. Section III presents our methodology including the dataset used and the types of GANs that were explored. Section IV discusses our findings based on an expert pathologist's evaluation of real and fake histopathology images. Finally, we conclude in Section V.

II. Background & Motivation

In this section, we provide a brief background on GANs and the motivation behind our work.

A. GANs

In 2014, Goodfellow *et al.* proposed GANs [1], which are deep learning-based generative models. A GAN uses two neural networks: a *generator* and a *discriminator*. The generator generates new data instances by learning from the training data while the discriminator decides if each of those instances is authentic or fake, and identifies if the generated images are real. These two neural networks contest against each other to produce new, synthetic instances of data that are very close to real images. Since the inception of GANs, various new GAN designs have been proposed to overcome their limitations and improve the quality of the images generated in different domains.

Radford *et al.* proposed deep convolutional generative adversarial networks (DCGANs) [14] by extending traditional GANs using deep convolutional neural networks for the generator and discriminator networks. The fully connected layers on top of convolutional features were eliminated, which enabled this network to scale to larger datasets. Batch normalization was used in both the generator and the discriminator. The model used the ReLU activation function in all the generator network layers, with an exception of the output layer, which used the tanh activation function. For the discriminator network layers, the model used the LeakyReLU activation function.

Brock *et al.* proposed BigGANs [15], which were introduced to bridge the fidelity and variety gap between the generated images and real-world images. The architectural changes done to the original GANs allowed BigGANs to scale. A *truncation trick*, which involved using different distributions for the generator's latent space while training and inferencing, allowed for an optimal trade-off between image quality (or fidelity) and image variety.

Karras *et al.* proposed StyleGAN [12] by extending the traditional GAN architecture by incorporating changes to the generator model including the use of a non-linear mapping network that mapped points in latent space to an intermediate latent space. Stochastic variation was introduced through noise added at each point in the generator model that enabled finer interpretation of the style of the generated image. Later, they proposed StyleGAN2 [11] to improve the quality of images generated by StyleGAN, by addressing erratic artifacts in the generated images. The root cause for these artifacts were attributed to the adaptive instance normalization. Hence, the generator normalization technique was redesigned and replaced with weight demodulation applied to the weights of each convolutional layers.

Recently, Quiros *et al.* proposed PathologyGAN [13] for generating synthetic histopathology image patches from breast cancer WSIs. PathologyGAN used BigGAN as its baseline architecture to build a latent space of key tissue features. In addition to it, it also incorporated advances from StyleGAN to optimize this latent space in order to identify features of cancerous tissues. The model also replaced *hinge loss* with Relativistic Average Discriminator as the GAN's objective was to enable faster convergence and help capture morphological structure of tissues accurately.

B. Motivation

Recent advances in GANs have shown their usefulness in real-world applications and medical imaging for synthetic image generation, denoising, reconstruction, and translation. Sorin *et al.* [16] surveyed numerous applications of GANs in the field of radiology. A more recent work by Tschuchnig et al. [9] surveyed the potential of GANs in pathology. They looked at how GANs could be used to augment the various tasks in digital pathology. While there are many benefits to using GANs, one may wonder if deepfakes are a myth or reality in fields like radiology and digital pathology. Recently, Mirsky *et al.* proposed CT-GAN [4] and showed how an adversary can use a GAN to tamper a CT scan by adding or removing medical conditions. This opens up the possibility of malicious attacks in medical imaging systems and can lead to devastating consequences. We ask the following question: *can histopathology images be tampered by an adversary using GANs?* To the best of our knowledge, none of the prior work has attempted to answer the aforementioned question. As such attacks can seriously threaten the lives of patients, in this work, we explore if GANs could be used in an adversarial

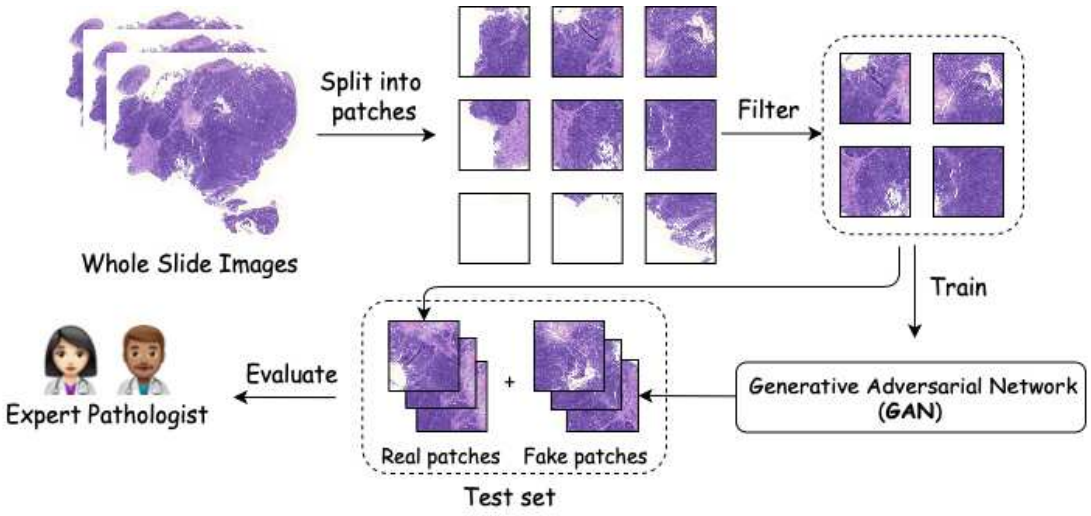


Fig. 1. Overall steps involved in constructing the dataset and generating fake histopathology patches

setting to produce fake histopathology images that could trick a pathologist. This will enable us to eventually develop robust schemes to detect tampering of WSIs once digital pathology becomes prevalent in US hospitals.

III. Methodology

In this section, we present the details of our methodology. We begin by describing the dataset used for the study. We then describe the implementation and experimental setup.

A. Dataset

One of the goals of our study was to use real datasets to train GANs so that the corresponding deepfakes may indeed look real. Therefore, we decided to use WSIs from TCGA, which is the result of a joint effort by the National Cancer Institute (NCI) and the National Human Genome Research Institute.

We illustrate our methodology in Figure 1. We prepared a dataset comprising of 100,000 patches by extracting image patches from 3 WSIs available via TCGA. As our expert pathologist had training in gastrointestinal pathology, we obtained the WSIs from The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) collection [17]. The files were in `.svs` format, which is a popular format used by WSI scanners such as Aperio. An SVS file contains a pyramid of image tiles. It contains multiple levels at different image resolutions based on the magnification used during scanning (e.g., 40X).

We used *py-wsi* [18], an open source Python package that uses OpenSlide [19], to enable easy and intuitive patch sampling. We extracted patches of size 256×256 pixels at level 17. Higher the level, higher is the image resolution. (All 3 WSIs had level 17.) We also provided an overlap value of 50 to py-wsi. This value indicates the number of pixels to be added to each side of a tile during extraction. We filtered out those patches that had a lot of white space; all the selected patches were of at least 100KB in size. A few examples of the extracted patches for training GANs are shown in Figure 2. We trained three GANs using these patches. The real and fake patches were then put into a test set, which was then evaluated by an expert pathologist.

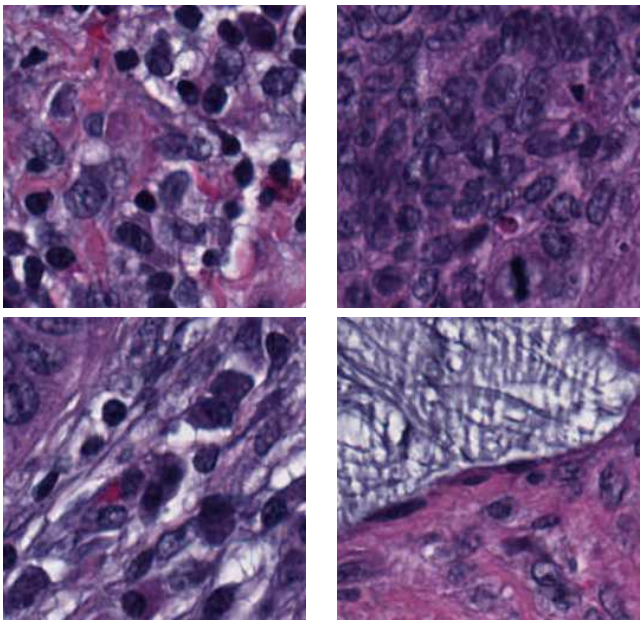


Fig. 2. Examples of patches used to train GANs

B. Experimental Setup

We conducted all our experiments on CloudLab [20], an experimental tested for cloud computing. We chose machines in the Wisconsin data center. Each machine had 2 Intel Xeon Silver 4114 10-core CPUs (2.20 GHz), 192 GB of RAM, 480 GB SSD storage, 1 TB disk drive, and ran Ubuntu 18.04. It also had an NVIDIA PCI P100 GPU with 12 GB of GPU memory. The source code for StyleGAN, StyleGAN2 and PathologyGAN were obtained from the GitHub sites published

by the authors of these GANs. We used a single machine to train each GAN on 100,000 patches using the default parameters specified by the authors. For StyleGAN, the default minibatch size was 4 and the initial resolution at the beginning of the training was set to 8. For StyleGAN2, the default minibatch size was 32 and the initial resolution was set to 8. The output image patch size for StyleGAN and StyleGAN2 was 256×256. For PathologyGAN the default number of epochs and default batch size were 45 and 64, respectively. The output patch size was 224×224.

C. Evaluation Strategy

To conduct a fair evaluation, we prepared a test set of 60 patches, wherein 30 were real and the remaining 30 were fake. We did inform our expert pathologist that half of the images in the test set were real. This way our expert pathologist would not have a biased assessment. We asked the pathologist to mark each patch in the test set as "real", "fake", or "unsure". We also requested our pathologist to provide a morphological reason when she flagged a patch as fake. To prepare the test set, we first chose a real patch. Then we picked a fake patch that appeared to be similar to the real one as a non-expert. Note that among the 30 fake patches, we selected 10 fake patches from StyleGAN, 10 from StyleGAN2, and 10 from PathologyGAN. We randomly ordered the 60 patches in the test set. Figure 3 show samples of fake and real patches used in the test set.

IV. Results

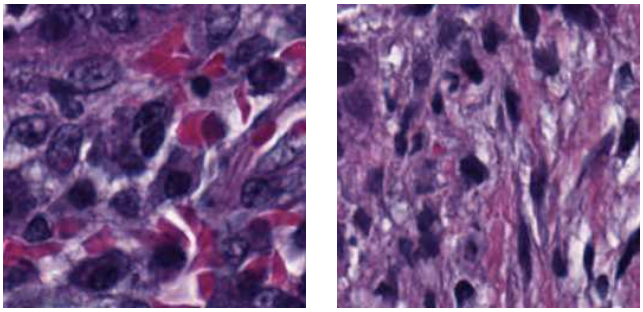
In this section, we present the results of the evaluation on the test set by the expert pathologist.

TABLE I
SUMMARY OF THE PATHOLOGIST'S EVALUATION

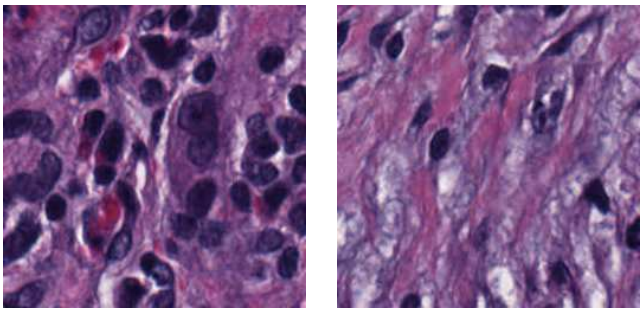
		Pathologist's evaluation		
		Real	Fake	Unsure
Ground truth	Real	17	13	-
	Fake	10	17	3

Table I shows the summary of the pathologist's evaluation for the test set containing both real and fake patches. As observed, out of 60 patches in the test set, 10 fake patches were marked as real by the pathologist and indeed tricked an expert's eye. Some examples of these fake patches are shown in Figure 4. However, the pathologist correctly identified 34 out of 60 patches (as fake or real). Note that three fake patches were marked as unsure by the pathologist.

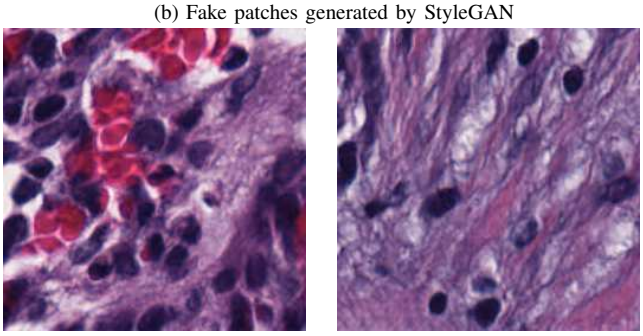
We also asked the pathologist to provide morphological reasons as to why a particular patch was marked as fake. This would give us useful insights from a clinical perspective and provide a better understanding of possible adversarial attacks on histopathology images.



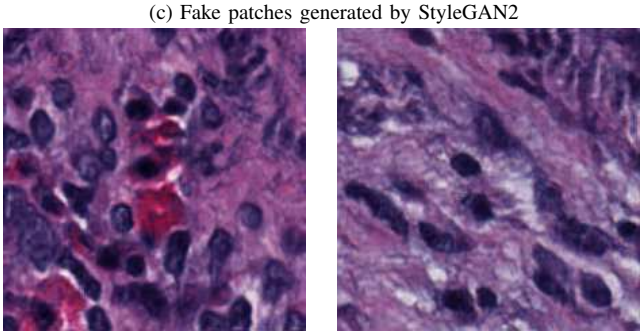
(a) Real Patches



(b) Fake patches generated by StyleGAN



(c) Fake patches generated by StyleGAN2

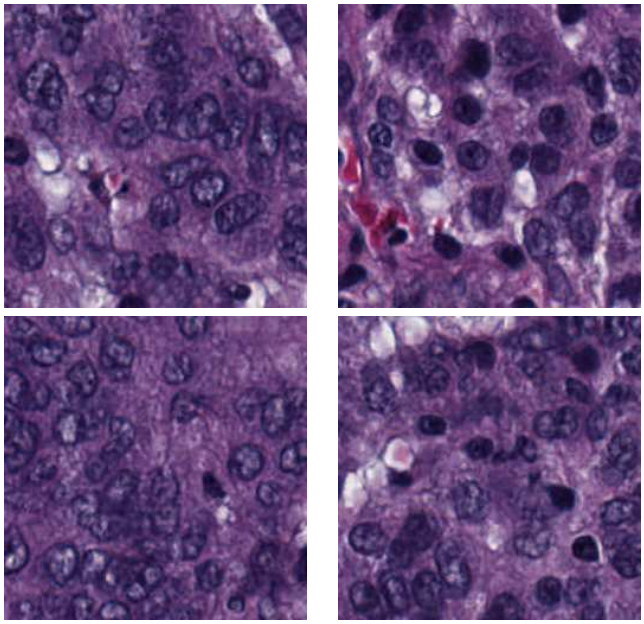


(d) Fake patches generated by PathologyGAN

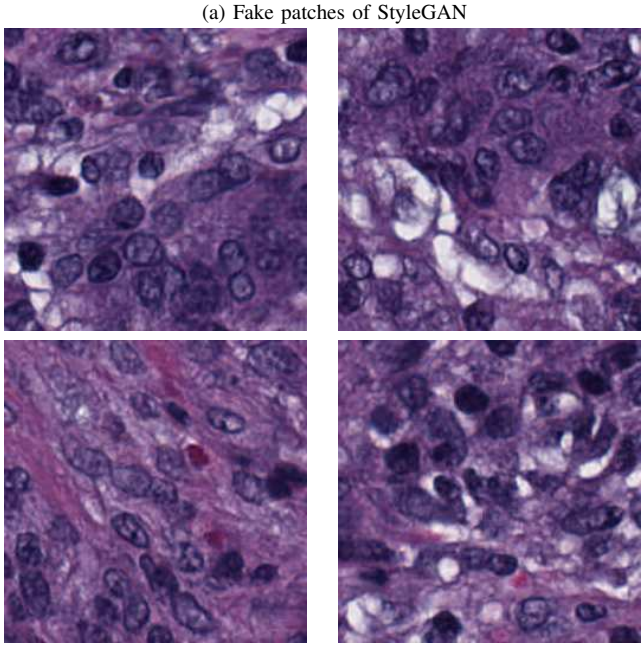
Fig. 3. Samples of real patches and fake patches (generated by GANs) used in the test set

Figure 5 illustrates the scenario where there were size variation of cells in fake patches produced by PathologyGAN. (These are shown as red ovals in the figure.) This means that GANs must ensure that the variation of the cell sizes are not very obvious in a generated patch to avoid being detected as fake.

Figure 6 illustrates the scenario where there were unusual



(a) Fake patches of StyleGAN

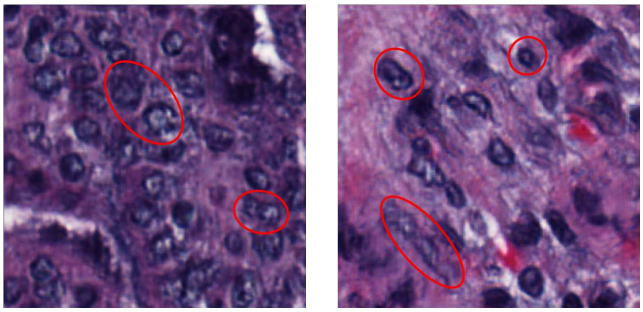


(b) Fake patches of StyleGAN2

Fig. 4. Examples of fake patches that were identified as real by the pathologist

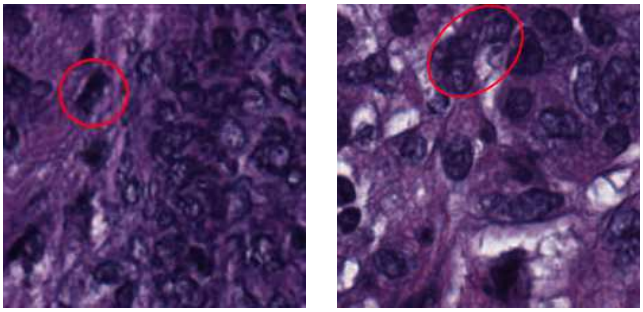
nuclear lobulations in the fake patch generated by PathologyGAN. Furthermore, the nuclear shapes were unusual in the fake patch generated by StyleGAN. (These are shown as red ovals in the figure.) Thus, a GAN must consider the nuclear shapes and lobulations in its architecture to generate fake patches that can evade a pathologist's keen eye.

Figure 7 illustrates the scenario where the presence of naked nuclei within the extracellular mucinous matrix in the fake patches was obvious to the pathologist. These patches were generated by PathologyGAN and StyleGAN2. (These are shown as red ovals in the figure.) Therfore, GANs must avoid



(a) PathologyGAN (b) PathologyGAN

Fig. 5. Size variation of cells in fake patches



(a) PathologyGAN (b) StyleGAN

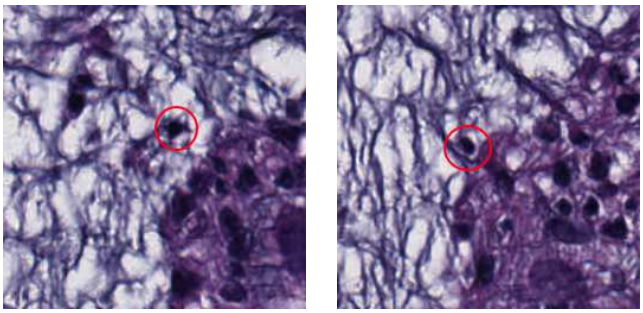
Fig. 6. Unusual nuclear lobulation and unusual nuclear shapes in fake patches

generating naked nuclei within the extracellular mucinous matrix to generate fake patches that look real.

Figure 8 illustrates the scenario where unusual red blood cell (RBC) shapes were detected in the fake patches by the pathologist. (These are shown as red ovals in the figure.) This observation was made for patches generated by both PathologyGAN and StyleGAN2. We believe detailed bounding boxes marking RBC boundaries would facilitate GANs to more accurately learn the RBC shapes.

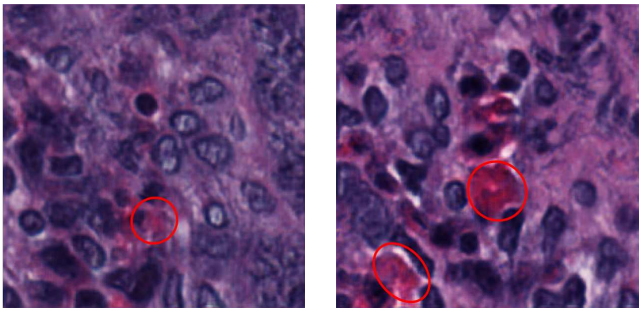
In another scenario, the patch generated by PathologyGAN shown in Figure 9 was identified as fake due to irregular or poorly delineated extracellular space. This limitation of PathologyGAN could be resolved by training over a larger labeled dataset.

Figure 10 illustrates the scenario where there were am-



(a) PathologyGAN (b) StyleGAN2

Fig. 7. Naked nuclei within the extracellular mucinous matrix



(a) PathologyGAN (b) StyleGAN2

Fig. 8. Unusual RBC shape

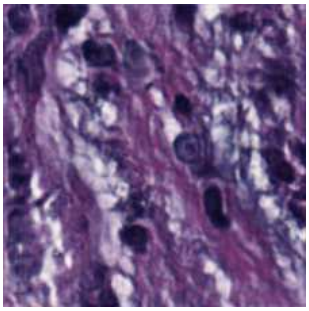


Fig. 9. Poor delineation of extracellular space in the patch generated by PathologyGAN

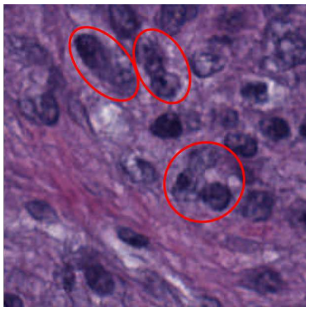


Fig. 10. The separation between the nuclei was not well defined in the patch generated by StyleGAN

biguous separation between the nuclei. (These are shown as red ovals in the figure.) A GAN should consider the nuclei morphological features to generate fake patches that contain well-defined separation among nuclei in order to evade a pathologist.

TABLE II
COMPARISON BETWEEN GANS

Model	Identified as Real	Fake	Unsure
StyleGAN	4	6	-
StyleGAN2	6	3	1
PathologyGAN	0	8	2

Table II breaks down the pathologist's evaluation based on the GAN model that generated the corresponding patches. Of the three models, StyleGAN2 showcased the best results, tricking the pathologist 6 out of 10 times, followed by StyleGAN, with 4 of the patches being incorrectly identified as real. While none of the patches generated by PathologyGAN were identified as real, 2 of them were marked as unsure by the pathologist.

V. CONCLUSION

In this paper, we investigated if a GAN could be used to generate deepfakes for histopathology images. We used 3 GANs namely, StyleGAN, StyleGAN2 and PathologyGAN, to generate fake patches by training them over 100,000 real patches from 3 WSIs of colon cancer patients. An expert pathologist then evaluated these generated patches. While certain parts of a histopathology image can be mimicked by a GAN and trick the pathologist, the intricacies of the stained tissue and cells cannot be fully captured by existing GANs. Thus, it was easy for the pathologist to observe numerous incorrect artifacts related to morphological features of cells in the fake patches. Hence, through our study we conclude that it is more challenging to generate deepfakes in digital pathology than in radiology. While GANs in their current state cannot completely trick a pathologist, we believe that with a much larger labeled dataset and careful feature engineering, it could be possible to accurately capture the morphological features in a histopathology specimen. Only time will tell when an adversarial attack on histopathology images will become a reality. The instructions for training StyleGAN, StyleGAN2, and PathologyGAN and the datasets used are available on GitHub [21].

ACKNOWLEDGMENTS

This work was partially funded by the National Science Foundation under Grant No. 1747751. The first author (N. A.) would like to thank the University of Tabuk, Saudi Arabia for sponsoring her scholarship. We thank the CloudLab team for the infrastructure support. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[2] "Philips receives FDA clearance to market philips IntelliSite pathology solution for primary diagnostic use in the US," https://www.philips.com/a-w/about/news/archive/standard/news/press/2017/20170413-philips-receives-fda-clearance-to-market-philips-intellisite-pathology-solution-for-primary-diagnostic-use-in-the-us.html.

[3] "Leica biosystems receives FDA 510(k) clearance to market a digital pathology system for primary diagnosis," https://www.prnewswire.com/news-releases/leica-biosystems-receives-fda-510k-clearance-to-market-a-digital-pathology-system-for-primary-diagnosis-300857825.html.

[4] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 461–478.

[5] Y. Liu, T. Kohlberger, M. Norouzi, G. E. Dahl, J. L. Smith, A. Mohtashamian, N. Olson, L. H. Peng, J. D. Hipp, and M. C. Stumpe, "Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists," *Archives of Pathology & Laboratory Medicine*, vol. 143, no. 7, pp. 859–868, 2019.

[6] D. F. Steiner, K. Nagpal, R. Sayres, D. J. Foote, B. D. Wedin, A. Pearce, C. J. Cai, S. R. Winter, M. Symonds, L. Yatziv, A. Kapishnikov, T. Brown, I. Flament-Auvigne, F. Tan, M. C. Stumpe, P.-P. Jiang, Y. Liu, P.-H. C. Chen, G. S. Corrado, M. Terry, and C. H. Mermel, "Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies," *JAMA Network Open*, vol. 3, no. 11, pp. e2 023 267–e2 023 267, 2020.

[7] D. E. L. Barron, D. V. K. Yarlagadda, P. Rao, O. Tawfik, and D. Rao, "Scalable Storage of Whole Slide Images and Fast Retrieval of Tiles Using Apache Spark," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, 2018, pp. 291–296.

[8] D. E. L. Barron, P. Rao, D. Rao, O. Tawfik, and A. Zachariah, "Large-Scale Storage of Whole Slide Images and Fast Retrieval of Tiles using DRAM," in *Big Data II: Learning, Analytics, and Applications*, vol. 11395, 2020, pp. 45 – 50.

[9] M. E. Tschuchnig, G. J. Oostingh, and M. Gadermayr, "Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential," *arXiv preprint arXiv:2004.14936*, 2020.

[10] "The Cancer Genome Atlas Program," https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[12] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[13] A. C. Quiros, R. Murray-Smith, and K. Yuan, "PathologyGAN: Learning Deep Representations of Cancer Tissue," *arXiv preprint arXiv:1907.02644*, 2019.

[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv preprint arXiv:1511.06434*, 2015.

[15] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[16] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) - A Systematic Review," *Academic Radiology*, pp. 1175–1185, 2020.

[17] "TCGA-COAD," https://portal.gdc.cancer.gov/projects/TCGA-COAD.

[18] R. Stone, "py-wsi." [Online]. Available: https://github.com/ysbecca/py-wsi

[19] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "Openslide: A Vendor-Neutral Software Foundation for Digital Pathology," *Journal of Pathology Informatics*, vol. 4, 2013.

[20] D. Duplyakin, R. Ricci, A. Maricq, G. Wong, J. Duerig, E. Eide, L. Stoller, M. Hibler, D. Johnson, K. Webb, A. Akella, K. Wang, G. Ricart, L. Landweber, C. Elliott, M. Zink, E. Cecchet, S. Kar, and P. Mishra, "The Design and Operation of CloudLab," in *Proceedings of the USENIX Annual Technical Conference*, 2019, pp. 1–14.

[21] "Deepfakes for Histopathology Images," https://github.com/MU-Data-Science/Deepfakes-Histopathology.