
# DEEPFAKES: INTRODUCTION AND APPLICATIONS IN DIGITAL HEALTH

Digital Health Entrepreneurship

Angel Igareta *(angel@igareta.com)*

February 26, 2021

**Abstract**

Deep Learning algorithms are being used in several fields nowadays, such as Self-driving cars, Voice-Activated Assistants, or Digital Health. Software developed with these algorithms is getting easier for not experienced users to use and their features are constantly growing. Despite this being positive for contributing to previous sectors, deep learning-powered applications that could cause threats in politics, privacy, and security are being developed too. An example of this is the so-called *deepfakes*. This paper has the goal of introducing this type of algorithms and examples of what this technology is capable of in the field of Digital Health. The aim is to inform about the state-of-the-art applications on medical imagery and alert people about the possible harms this can provoke in the early future.

# 1 Introduction

Deepfake is a deep learning technique used to create convincing images, audio, and video hoaxes. The most common application is to swap a person in an existing image or video with another one, generating a fake synthetic media. The contrast with traditional techniques such as Photoshop is that deepfakes are powerful techniques for generating visual or audio content with high potential to deceive. Some of the popular areas where these algorithms are used are fake news, financial fraud, and celebrity pornographic videos. However, there are less popular but still critical ones in the field of digital health. According to the legal response, some countries have been introducing legislation to tackle deepfake content. For instance, in the United States, some states such as California, Texas, and New York have introduced *"The Malicious Deep Fake Prohibition Act"*, which will make a federal crime to create or distribute a deepfake when doing so will facilitate illegal conduct [4].

The first deepfake video was produced in 2017 where a popular face was swapped in pornographic content and published on Reddit. In 2018, the famous website banned this type of content classified as *fake porn*, which was followed by Twitter or Gfycat [7]. For instance, they can be used as part of social engineering scams through the branch of audio deepfakes, fooling people into thinking they are receiving instructions from a trusted source. In 2019, a UK energy firm CEO was scammed $\$243,000$ when he was ordered by a supposed chief executive from the parent's company to transfer the money to a Hungarian bank [5]. These algorithms could be particularly dangerous if applied to the field of medicine. For instance, Israeli researchers published a recent work where they generated CT and MRI fake scans (either injecting tumors or removing them from the original scan) to trick trained diagnosticians [9]. With the developed technique CT-GAN, they demonstrated the reach of these tools, and the causalities attackers could cause, such as to stop political candidates, perform an act of terrorism or even commit murder.

Despite all these dark applications of deepfake, there are many positive ones such as creating voices for those who have lost theirs, recording the same character in a film through

different ages, such as in the Netflix film *'The Irishman'* [15] or generating higher-resolution medical images to achieve better diagnosis [13].

The first section of the document has the aim of informing about the basic, as well as the state-of-the-art technologies used to generate deepfakes. In the second section, some applications of these technologies over medical imagery will be presented. Finally, there would be a personal opinion about the future of these technologies and the possible solutions.

# 2 Deepfake Introduction

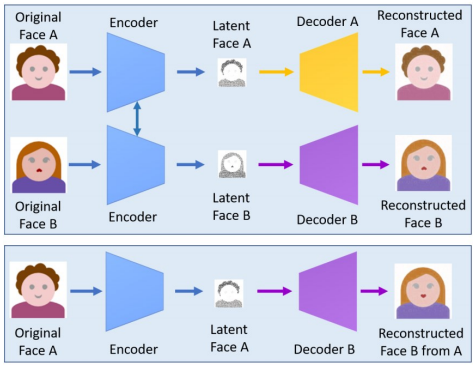
Since the creation of the first deepfake, the algorithms used to create them have improved significantly, at the same time as the algorithms used to detect them. This competition between the two sides has resulted in very robust deep learning techniques to generate fake content that is quite arduous to spot.

The first attempt of deepfake creation was FakeApp, developed using an autoencoder-decoder pairing structure. Autoencoder is an unsupervised artificial neural network that learns first how to efficiently compress data down to a selected size (encode step) and then reconstruct the data from the compressed representation to one that is as close as the input as possible (decode step). To measure how well the decoder is performing and how close the output is to the original input, the reconstruction loss is calculated, so the training involves using backpropagation to minimize this loss. Some of the most popular uses for these technologies are anomaly detection and image denoising [3].

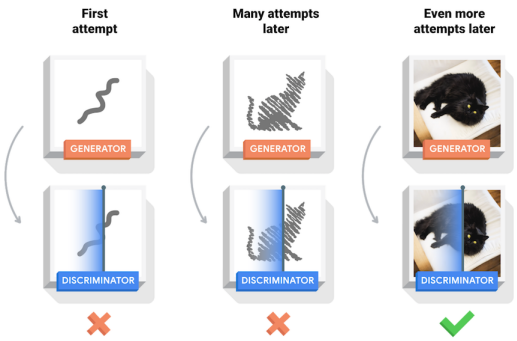
Furthermore, as FakeApp proved by allowing to swap faces of two individuals, these neural networks can also be used to generate deepfakes. In this application, the autoencoder intention is to extract facial features from an input face (as the algorithm compresses, only the essential features of the face will remain). Then, the decoder takes those features as input and reconstructs them to produce the original input face. As the goal is to swap two faces, it does not only use a single autoencoder-decoder structure but a pair. The first one will learn the facial expressions of the subject A and make use of the compressed features to reconstruct the subject A, the second one will perform likewise with subject B. On top of that, both structures possess the same encoder network. Subsequently, to swap faces, it only requires to swap the decoders, as they have learned how to reconstruct the subject they were trained with. As faces normally have similar features such as eyes, moth positions, eyebrows, the task is simple. In figure 1a, an example of these approaches is displayed, it is used to copy the facial expression of subject A into subject B.

A major improvement in the deepfake development, as well as in the overall Artificial Intelligence advances were the generative adversarial networks (GANs), related to the field Adversarial Learning. These are algorithmic architectures constituted by two neural networks, competing against each other, these two are called the generator and the discriminator. The generator is a neural network which aim is to produce new data instances, while

the discriminator evaluates them to see if they are real or fake, in other words, it determines if the received image belongs to an actual training set. If we apply this concept to deepfakes in medical imagery, our generator will be generating images that mimic the target object, for instance, a CT scan with a lung tumor. In the beginning, the generator will start by generating random noise, as it does not have any information about how a tumor is. That feedback is given by the discriminator, through its loss. Meanwhile, the discriminator is being trained with real CT scans where lung tumors appear from a training dataset and fake images previously generated by the generator. Hence, the generator progressively becomes better at creating CT scans that look real, while the discriminator becomes better at spotting which ones are fake. Our stop condition would be when the discriminator can no longer distinguish real CT scans from fakes. An example trained with images of cats is displayed in figure 1b.




(a) Example of autoencoding-decoder pairs structure to create a deepfake [10]




(b) Generative Neural Network example [1]

Figure 1

# 3 Deepfake Detection

With all these advances in the area of deepfakes, it has become easier to generate them and the input data size constraint has been lowered. Therefore, the detection tasks have become more important to recognize these fake synthetic media. Early techniques were based on obtaining handcrafted features from artifacts and inconsistencies of the fake videos. On the other hand, recent methods are based on deep learning to automatically extract these features.

The task of deepfake detection is a classification problem whose goal is to distinguish between tampered images and original ones. To accomplish that, most of the techniques require a large amount of input data, meaning a large amount of modified or original images.

This is a significant difficulty in some areas such as lung cancer deepfakes where there are not many synthesized content available to train techniques with.

In deepfake detection, two major categories can be distinguished: fake image detection and fake video detection. In the latter, apart from analyzing the features, it is possible to analyze the temporal features between frames based on deep learning recurrent classification models.

### 3.1 Fake Image Detection

One of the most dangerous techniques in Fake Image Detection is *face swapping*, as cyber-attackers could use this technique to penetrate in identification or authentication systems and gain illegitimate access. The application of GANs to enhance deepfakes has made challenging the job of detecting face swapped images, as they can imitate the pose and lighting conditions of the target image. One of the approaches with better generalization capability for any network was introduced by Xuan et al. [16], which applied an image preprocessing step, for instance, Gaussian blur, to remove low-level clues of GAN images. This forces the forensic classifier to learn deeper and meaningful features of the images.

### 3.2 Fake Video Detection

Most image detection methods are not suitable for video detection because of the strong degradation of each frame after video compression. Besides, videos possess temporal features that vary among the set of frames and that can be used for the detection. Recurrent Convolutional Networks (RCNs) could be applied to exploit these temporal discrepancies across frames. For instance, a study used the physiological signal of eye blinking to expose deepfakes, based on the evidence that a person in deepfakes has much less frequent blinking than in real videos [8]. To extract this information from the tampered videos a CNN based on long short term memory (LTSM) was used. On top of that, recent research funded by Google and DARPA (research wing of the Pentagon) presented a digital forensic technique able to detect world leaders or celebrities deepfakes. The method uses machine learning to analyze specific individual's *style of speech and movement*, what researchers reference as a "soft biometric signature" [2].

## 4 Deepfake Applications in Digital Health

In medicine, GANs can have a very positive impact as they could be used to generate synthetic healthcare data for Artificial Intelligence techniques to, for instance, improve disease diagnosis. However, there can also be negative uses of deepfakes with the aim of condition doctors' opinions or completely control them by altering diagnosis.

Contribution of GANs in medical image study can be classified into 7 different categories: de-noising, reconstruction, segmentation, registration, detection, classification, and synthesis. They are better visually distinguished in figure 2. Attending to the distribution of papers from each category in the literature, the most popular topics are about segmentation and synthesis of medical images [6]. In this section, the latest deepfakes applications from these categories will be presented.

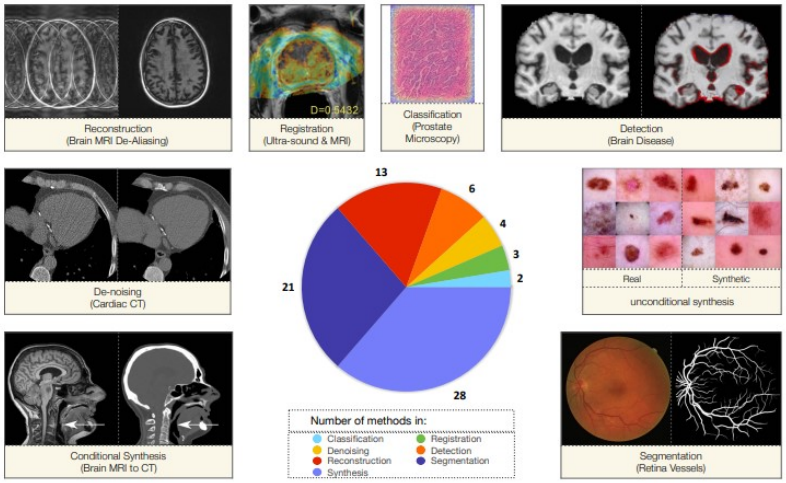



Figure 2: Distribution of papers and visual examples of GANS applications in Medicine [6]

## 4.1 Medical Image Synthesis

One major problem related to medical imagery datasets is that there is little data to train with, causing the Machine Learning techniques to not generalize properly. To tackle this, recent research has shown that DCGANs (Deep Convolutional Generative Adversarial Networks) can be used to synthesize highly deceiving: small patches of prostate lesions, retinal images, or lung cancer nodules [6]. This does not only help to gain deeper insights into the distribution of that data and their nature but also to generate data that could be used to augment the training data in deep learning techniques.

Regarding the supervised medical image synthesis (where the GANs are not only trained with raw noise but with previous data that helps with the synthesis goal) there have been significant advances in the generation of CT (Computed Tomography) scans from MRI (Magnetic Resonance Imaging). The reason for this is because in many clinical conditions CT scans are required to make decisions, however, CT scans utilize X-rays to produce the images, while MRI uses powerful magnetic fields and radiofrequency, which is less harmful to the patient. For instance, recent research proposes a cascaded generative adversarial network to tackle this [17]. They use deep supervision discrimination alongside feature descriptions of a

pre-trained VGG16 mode to achieve feasible image-to-image translation results. This helps discriminate real and synthetic CT images and provide gradient updates to the actual GAN.

## 4.2 Medical Image Segmentation

For many applications such as detection and classification in medical images, the segmentation of objects and organs is crucial. The traditional way of manual segmentation is highly time-consuming and tiresome, hence causing this branch to be one of the most active in deep-learning research in medicine. However, pixel-based segmentation does not work efficiently over this task, due to the complex anatomical structures. The introduction of GANs has granted a different learning approach to address this problem, with promising results.

### 4.2.1 Brain Tumor Segmentation

AI Researchers from the giant NVIDIA, MGH  BWH Center for Clinical Data Science, and the Mayo Clinic [11] used GANs to generate synthetic abnormal MRI images with brain tumors and to illustrate how these synthetic images affect tumor segmentation tasks. First, they retrieved two public datasets of brain MRI images. Through the use of GANs, they were able to generate new pictures where the tumors could be moved to different areas of the brain or make them smaller or larger. They could even put some tumors into an otherwise healthy brain. By leveraging these synthetic images as a form of data augmentation, they achieved to improve performance on tumor segmentation.

## 4.3 Retinal Vessel Segmentation

By analyzing retinal vessel networks, ophthalmologists can detect early signs of the increased systematic vascular burden from diabetes or hypertension. To aid this task, an automatic vessel segmentation method has been broadly researched. In a paper presented by Vuno Inc, they introduce a method that generates the precise map of retinal vessels the specialists need using generative adversarial learning [12]. This technique overcomes some problems of already existing solutions that tend to miss fine vessels or allow false positives at terminal branches. Finally, they achieve state-of-the-art performance in DRIVE (Digital Retinal Images for Vessel Extraction) and STARE (STructured Analysis of the Retina Dataset) datasets.

## 4.4 Generation of High-Resolution Medical Images

Generative Adversarial Networks could also be used to improve cancer diagnostics, due to their ability to synthesize highly realistic images. One of the difficulties to accomplish this

goal is the computational power creating these images take, making it unfeasible in hospitals. A research made in the university of Lübeck proposes a novel multi-scale GAN technique to generate high-resolution 2D and 3D images requiring much less computational power [13]. The idea is to create them by first learning from a low-resolution version of the images and then produce pieces of sequentially growing resolutions. This could be applied to produce better predictions in other GAN applications such as the ones previously presented.

## 4.5 Dangers of Deepfakes on Medical Images

As well as the positive advantages GANs offer in medical applications, they can be used to provoke the opposite effect. One problem in many medical buildings is the lack of computer security and the outdated software being used, letting a potential window for attackers to break in and handle patients data. Israeli researchers have presented a flaw in CT and MRI systems that allowed them to inject realistic images of cancerous growths or remove tumors from the scan entirely, convincing trained diagnosticians no disease was present were there surely was [9]. An intruder could perform these attacks to stop a political candidate, commit insurance fraud, perform an act of terrorism, or even commit murder.

CT scans are used to diagnose cancer, appendicitis, trauma, and infectious diseases. In 2016 there were approximately 79 million CT scans performed in the United States [9]. These scans consist of 3D images created by taking multiple 2D scans of the axial plane along the body part being scanned. Nowadays, they are managed through a picture archiving and communication system (PACS), which is composed of a central server that handles the scans communication (receive, store, and retrieve). These scans are sent and stored using DICOM format.

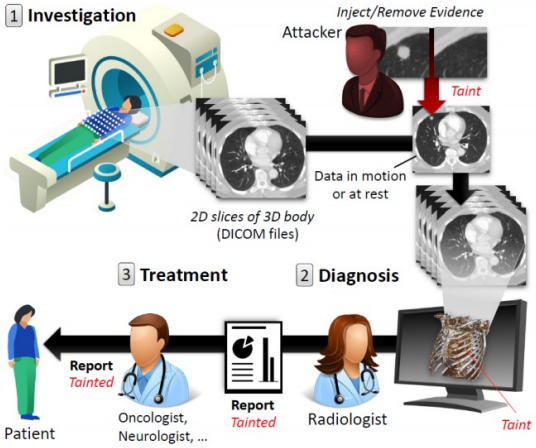



Figure 3: Example of tampering medical imagery in each stage between the investigation and diagnosis [9]

The threat of these systems is that an attacker with access to medical systems can alter their content to cause a misdiagnosis. A search in Shodan.io shows that 2,645 DICOM servers are exposed to the Internet, which can facilitate access to attackers. Even if they were not directly connected to the Internet they are indirectly linked to the internal networks, which make them unsafe for social engineering attacks or insiders. A potential act from the attacker would be to add or remove evidence of some medical condition, illustrated in figure 3 with a lung cancer diagnosis.

In order to deploy and test the attack, the research team performed a penetration test and was successfully able to infiltrate and install a Raspberry Pi server into a hospital network. They identified that the data being sent over the internal network was in plaintext, hence not encrypted in any form, as well as the usernames and passwords of the 27 staff members and doctors. Once inside they could alter the scans from the CT scan before reaching radiologists, with a technique called CT-GAN.

To evaluate the attack, they concentrated on injecting and removing lung cancer from CT scans. To test the attack results, the team selected three radiologists with different years of experience (from 2 to 7) and perform them open and blinded tests. First, they tested them with unmodified scans for the presence or absence of cancer, where all performed correctly. The second test was with modified scans, where the success rate in cancer injection was above 99 percent, and the removal success rate was 95,8 percent. Finally, the open tests reduced the success rate to 90 and 70 percent respectively. This demonstrates how three expert radiologists were highly susceptible to the attack.

# 5 Conclusions

After the latest example of how harmful GANs can be the rationale would be to research detection techniques to spot these synthetic images. However, the current state in deepfakes is complicated. Every time a new deepfake technology is released, users will try to develop a technology to detect the images generated with that technology, and then hackers would try to bypass that detection and ideate a new approach for generating them, and this will continue with a virus/anti-virus dynamic. On top of that, despite achieving high accuracy in deepfake detection, for instance, 99%, sometimes that 1% could still cause much harm depending on the deepfakes amount.

Another approach for tackling this issue would be to use software that can automatically watermark original images or implement blockchain technology to verify content from trusted sources [14]. This could be very beneficial especially in medical imagery where the content exposed is crucial. However, the likely future is that none of these approaches would help 'solve' this issue, it is an endless competition that can have damaging results in terms of privacy and safety. Everyone needs to be aware of these technologies, not fully trusting their observations and verifying each source they read or watch. Almost everything can be faked.

# References


[1] Deep convolutional generative adversarial network : Tensorflow core.

[2] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019), pp. 38–45.

[3] Badr, W. Auto-encoder: What is it? and what is it used for? (part 1), Jul 2019.

[4] Brown, N. I. Congress wants to solve deepfakes by 2020. that should worry us., Jul 2019.

[5] Damiani, J. A voice deepfake was used to scam a ceo out of $243,000, Sep 2019.

[6] Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., and Mukhopadhyay, A. Gans for medical image analysis. *arXiv preprint arXiv:1809.06222* (2018).

[7] Kelion, L. Reddit bans deepfake porn videos, Feb 2018.

[8] Li, Y., Chang, M.-C., and Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (2018), IEEE, pp. 1–7.

[9] Mirsky, Y., Mahler, T., Shelef, I., and Elovici, Y. Ct-gan: Malicious tampering of 3d medical imagery using deep learning. In *28th {USENIX} Security Symposium ({USENIX} Security 19)* (2019), pp. 461–478.

[10] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., and Nahavandi, S. Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573* (2019).

[11] Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., and Michalski, M. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging* (2018), Springer, pp. 1–11.

[12] Son, J., Park, S. J., and Jung, K.-H. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv preprint arXiv:1706.09318* (2017).

[13] Uzunova, H., Ehrhardt, J., Jacob, F., Frydrychowicz, A., and Handels, H. Multi-scale gans for memory-efficient generation of high resolution medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), Springer, pp. 112–120.

[14] Vincent, J. Deepfake detection algorithms will never be enough, Jun 2019.

[15] Wood, C. A deepfake artist's attempt to make robert de niro look younger in 'the irishman' is being hailed as superior to netflix's cgi, Jan 2020.

[16] Xuan, X., Peng, B., Wang, W., and Dong, J. On the generalization of gan image forensics. In *Chinese Conference on Biometric Recognition* (2019), Springer, pp. 134–141.

[17] Zhao, M., Wang, L., Chen, J., Nie, D., Cong, Y., Ahmad, S., Ho, A., Yuan, P., Fung, S. H., Deng, H. H., et al. Craniomaxillofacial bony structures segmentation from mri with deep-supervision adversarial learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), Springer, pp. 720–727.