Applied Computer Science 2025, vol. 21, no. 2, pp. 139–153
https://doi.org/10.35784/acs_7054

Submitted: 2025-01-11 | Revised: 2025-02-19 | Accepted: 2025-03-02



Keywords: deepfakes, medical imaging, deep learning, artificial intelligence (AI) in healthcare, Generative Adversarial Networks (GANs)

*Pradeepan P* [1], *Gladston RAJ S* [2], *Juby GEORGE* [3*]
[1] University of Kerala, India, pradeepanp@cdit.org
[2] Government College Kariavattom, India, gladston@rediffmail.com
[3] Marian College Kuttikanam, India, jubygeorgek@rediffmail.com
[*] Corresponding author: pradeepanp@cdit.org

# A comprehensive review of deepfakes in medical imaging: Ethical concerns, detection techniques and future directions

**Abstract**

*Deep fakes pose a significant threat to medical imaging. These deep fakes appear very similar to real diagnostic scans and are often difficult to distinguish from real medical images. This paper discusses how deepfakes are created and highlights their potential for research and education, as well as risks such as misdiagnosis and data manipulation. We also review various deepfake detection techniques, ranging from traditional image forensics to advanced deep learning models, and highlight the strengths and weaknesses of these approaches for detecting sophisticated deepfakes. We also discuss the ethical issues of deepfakes in healthcare, such as patient privacy, data security, informed consent, algorithmic bias, and the potential loss of trust in medical systems. In addition, we present an experimental study that evaluates how well different deep learning models detect deepfakes in a lung CT scan dataset, demonstrating both the potential and limitations of current detection methods. Finally, we outline future research directions, including real-time detection, explicable AI, enhanced cybersecurity, and strengthened ethical guidelines. This review is a valuable resource for researchers, clinicians, and policymakers interested in exploring AI medical imaging and ethics in the age of deepfakes.*

## 1. INTRODUCTION

Deepfakes are artificial images or videos created using advanced AI algorithms. In healthcare, these images can mimic real medical images such as X-rays, MRIs, and CT scans (Westerlund, 2019). These images are generated using methods such as generative adversarial networks (GANs), variational auto encoders (VAEs), and diffusion models (Goodfellow et al., 2020). Although deepfakes can be helpful in research and education (Frid-Adar et al., 2018), they pose significant risks. These risks can compromise patient safety, affect diagnostic accuracy, and damage trust in the medical system. As deepfake technologies continue to improve, it is critical to address the ethical, security, and regulatory issues surrounding them. This review explores the relationship between the ethics of deepfake technology and cybersecurity in medical imaging. It also examines the gap between the rapid development of deepfake technology and the slow development of protective detection methods and regulations (Chen & Esmaeilzadeh, 2024). This gap poses challenges in preventing the potential harm of deepfakes. To mitigate these problems, experts in fields such as computer science, medicine, ethics, and law must work together.

This review addresses three key questions: What are the ethical and societal risks of deepfakes in medical imaging? How can we improve cybersecurity to protect medical images from deepfakes? What are the best deepfake detection methods and how can we standardize their testing?

The paper begins by explaining the main technologies involved in deepfake creation, including GANs, auto-encoders, and other deep learning techniques (Radford et al., 2015). It also discusses how these technologies are used in medical imaging. For example, deepfakes can be used for data augmentation when training AI models for diagnosis, maintaining patient privacy, and creating realistic simulations for medical education and training. These applications have the potential to improve diagnostic accuracy. They can also contribute to rare disease research and support personalized medicine (Nie et al., 2017). Despite these benefits, deepfakes have raised significant ethical and social concerns. It can lead to privacy violations and data

breaches. There is also the potential for misuse, including identity theft and fraud. Manipulation of medical images can also compromise diagnosis and patient safety (Finlayson et al., 2019). This can lead to misdiagnosis or incorrect treatment. In addition, the use of patient data to create deepfakes requires careful consideration of informed consent (Vayena et al., 2018). Patients need to understand how their data will be used and be aware of the potential consequences of deepfake technology. Deepfakes can also introduce bias into medical imaging models, leading to unfair or inequitable outcomes. As awareness of deepfakes increases, the public may begin to distrust medical images and healthcare institutions. This underscores the need for systems that can authenticate medical images and ensure their transparency.

This review also considers the evolving legal and regulatory landscape surrounding deepfakes in medical imaging. This includes an examination of existing laws related to privacy and medical malpractice, and the identification of gaps that need to be addressed by new regulations (Cochran & Napshin, 2021). A major focus will be on methods for detecting deep forgeries. These methods include traditional image forensics that look for inconsistencies in the image, deep learning-based detection methods that use convolutional neural networks (CNNs), recurrent neural networks (RNNs), and metadata analysis with blockchain for provenance tracking (Hsu et al., 2020). Figure 1 shows the conceptual framework of deepfake technology and its impact on medical imaging.

This paper explores possible future directions in the field. These include advances in detection techniques, such as the use of descriptive AI to improve transparency and multimodal analysis that combines image data with patient information (Xie et al., 2017). Improving cybersecurity measures is essential to create secure image storage systems and tamper-proof medical images. Establishing clear ethical guidelines and legal frameworks for the use of deepfakes, including patient consent procedures, is essential. The review also considers how deepfakes could be used positively in medical imaging, for example in personalized medicine or medical education. By addressing these challenges, the review aims to provide a clear understanding of the safe and responsible use of deepfake technology in medical imaging. This will guide future research and innovation, while ensuring that deepfake technology helps protect patient safety, maintain accurate diagnoses, and build public trust in healthcare.


**Fig. 1. Conceptual framework of deepfake technology and its impact on medical imaging**

The rest of this paper is organized as follows. Section 2 provides a background on deepfake technology. It also discusses its history and core principles. Section 3 reviews the methods for creating and detecting deepfakes. It focuses on advanced algorithms and techniques. Section 4 presents the experimental results. This section outlines the objectives, methods, results, and discussion of the results. Section 5 examines the challenges of detecting and preventing deepfakes in medical imaging. Section 6 provides policy recommendations to address these challenges. Section 7 identifies research gaps and suggests future directions

for improving detection and prevention. Finally, Section 8 concludes the paper with key takeaways and implications for the safe use of deepfake technology in healthcare.

## 2. BACKGROUND

Deep learning has transformed medical image analysis. This has enabled automated diagnosis, treatment planning, and drug discovery (Shen et al., 2017). These models can extract complex features from medical images, which has led to significant advances in clinical applications (Litjens et al., 2017). However, deep learning has also brought new challenges, particularly with the rise of deepfakes. In medical imaging, deepfakes use deep learning technologies, specifically Generative Adversarial Networks (GANs), to manipulate or create synthetic medical images. These images look real but are artificial and may represent medical conditions or anatomical structures. This raises unique challenges and ethical concerns that are not present in traditional image manipulation. Figure 2 presents a timeline illustrating the evolution of deepfake technology and its integration into medical imaging. This section examines how deep learning and deepfakes intersect in medical imaging.

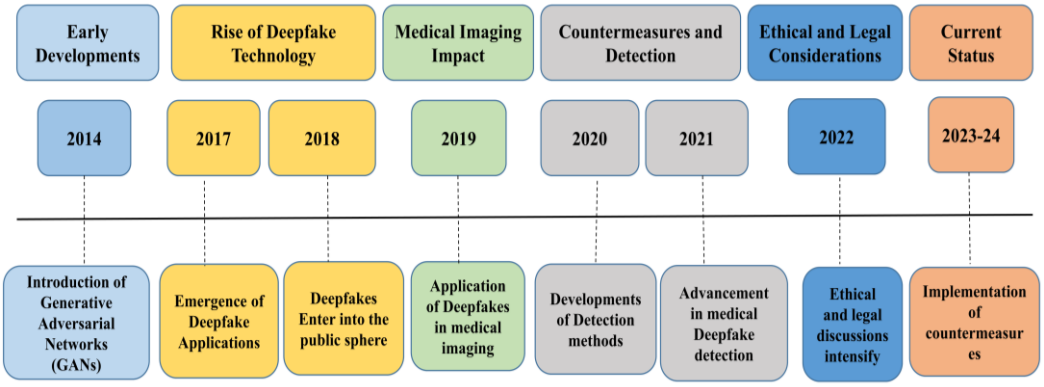



**Fig. 2. Timeline showing the evolution of deepfake technology and its integration into medical imaging**

Deep learning models, especially convolutional neural networks (CNNs), are useful for processing and understanding visual data. This makes them ideal for medical image analysis. CNNs have a layered structure that allows them to learn detailed image features. These range from simple edges and textures to complex anatomical structures. This has led to impressive advances in tasks such as image classification, segmentation, and object detection. The development of GANs has marked a significant step forward in deep learning research. GANs consist of two parts: a generator that creates fake data and a discriminator that tries to distinguish between real and fake data. Through adversarial training, GANs can generate realistic synthetic images, including medical images. GAN variants such as CycleGANs can learn to map between different image types. For example, they can convert MRI images into CT scans (Zhu et al., 2017). Another variant, StyleGANs, allows more control over the features of the generated images, further improving the quality of deepfakes in medical imaging (Karras et al., 2019). In addition to GANs, autoencoders, especially variational autoencoders (VAEs), are another powerful tool. Auto-encoders compress and reconstruct the data. They can also be used for tasks such as image denoising and anomaly detection. These advances in deep learning have made it easier to create deep fakes in medical imaging. Tab. 1 summarizes the state-of-the-art detection algorithms used in medical imaging. It highlights the methods, datasets, and contributions. This provides a clear overview of recent advances in medical image forgery detection. The ability to create and manipulate medical images using deep learning has opened up new opportunities in both research and clinical practice.

**Tab. 1. Summary of the state-of-the-art detection algorithms used in medical imaging**

| Authors | Methodology | Datasets | Contributions |
|---|---|---|---|
| Mirsky et al., 2019 | CT-GAN: Malicious tampering using GANs | Private 3D CT scans | Demonstrated the feasibility of injecting and removing tumors in 3D CT scans using GANs. Highlight risks associated with clinical and insurance fraud. |
| Reichman et al., 2021 | Deep learning-based framework, ConnectionNet | LuNoTim-CT Dataset | Proposed the ConnectionNet model for detecting manipulated regions in medical images. |
| Kim et al., 2022 | Sparse CNN with U-Net | Ocular Disease dataset | Developed a deepfake detection algorithm for medical data manipulation in eye disease images. |
| Budhiraja et al., 2022 | Medical deepfake detection using convolutional reservoir networks | Private dataset | Introduced convolutional reservoir networks to detect deepfakes in medical images and applied them to CT and X-ray modalities. |
| Alheeti et al., 2022 | Intelligent detection method for malicious tampering of cancer imagery | 3D Public CT-GAN Dataset | Focused on detecting manipulations in cancer images using deep learning techniques. |
| Sharafudeen & Vinod Chandra, 2023 | 3D Convolutional Neural Networks | 3D Public CT-GAN Dataset | Highlighting the effectiveness of 3D deep learning architecture in detecting medical deepfakes. |
| Karaköse et al., 2024 | YOLOv5 and YOLOv8 Comparative Analysis | Private dataset of lung CT, X-rays | YOLOv5 outperformed YOLOv8 with faster training and higher recall. Demonstrated robustness in detecting manipulated lung CT and osteoarthritis x-ray images. |
| S & Narayan, 2024 | LBP preprocessing, U-Net, SVM Classifier | LIDC-IDRI dataset, CT-GAN dataset | Developed a robust framework for detecting manipulated medical images by integrating multiple detection techniques. |
| Zhang et al., 2024 | Two-stage cascade framework (local detection + global classification) | CT scans with injected/removed lung cancer lesions | Proposed a method that achieves excellent performance in detecting small region forgeries in CT images generated by CT-GAN. |
| Latif et al., 2024 | Modified CNN | 3D Public CT dataset | Implemented AlexNet using transfer learning to detect manipulated 3D medical images. |

An important application of this method is data augmentation. Deepfakes can be used to create synthetic medical images that can increase the size of training datasets. This is especially helpful for rare diseases or conditions where it is difficult to collect large amounts of real patient data. This can improve the performance and robustness of AI models used for diagnosis and segmentation (Shorten & Khoshgoftaar, 2019). Another important application is anonymization. Deepfakes can replace identifiable features in medical images with synthetic features, protecting patient privacy. This makes it easier to share and analyze medical data while complying with privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) (Kaissis et al., 2020).

Deep fakes can also be used in medical simulations. This helps medical professionals practice complex procedures and diagnose rare cases in a safe environment. Deepfake technology can benefit several medical fields, including surgery and radiology, where visual interpretation and decision-making are critical (Mccormic et al., 2022). However, the use of deepfakes in medical imaging raises several ethical and societal concerns that need to be carefully considered. A major concern is privacy and data security. For example, synthetic medical images can be misused. Malicious actors could create fake images to make false medical claims, commit identity theft, or blackmail patients (Mirsky et al., 2019). Another concern is the potential impact on diagnostic integrity and patient safety. If deepfakes are used to subtly alter medical images, they could lead to misdiagnosis or incorrect treatment. This can have serious consequences for patient health and safety. The issue of obtaining informed consent must also be addressed. If patient data is used to create deepfakes, patients must be informed of how their data will be used. Clear guidelines should be established to ensure that consent is obtained ethically and transparently. This ensures patient autonomy and helps build trust in the medical

system. There is also the issue of bias and fairness in deep learning models. These models reflect the biases in the training data. If the data used to train a model is biased, the model may not work equally well for all demographic groups. In medical imaging, this can lead to inequitable access to quality care. Addressing these biases through careful data selection and ongoing evaluation is necessary to ensure fairness in deepfake applications (Mehrabi et al., 2022). The legal and regulatory landscape for deepfakes in medical imaging is still evolving. Existing privacy laws, such as HIPAA and GDPR, and medical malpractice laws may not fully address the challenges posed by deepfake technology. New legal frameworks are needed to address issues such as liability for misdiagnosis caused by fake images.

It is also important to establish standards for synthetic medical data and address intellectual property rights related to AI-generated images. Legal guidelines must also ensure the ethical use of patient data for deepfake research. Several studies have investigated deepfake techniques in medical imaging. Mirsky et al. (2019) used deep learning to show how CT scans could be vulnerable to tampering. They demonstrated how deep fakes could add or remove cancerous lesions in medical images. This raises concerns about the security of medical imaging systems. (Motamed et al., 2021) investigated how GANs can generate synthetic medical images for data augmentation. Their research showed that deepfakes could improve AI diagnostic tools, especially for rare diseases. However, they also emphasized the importance of validating synthetic datasets to avoid potential biases or errors.

The ethical and legal challenges associated with deepfakes have been widely discussed. Various frameworks have been proposed to address these issues. Some of these frameworks provide specific recommendations for medical imaging. Some researchers have explored the use of blockchain to securely track and authenticate medical images (Salah et al., 2019). Others have explored how explainable AI (XAI) can be used to improve transparency in deepfake detection systems (Tjoa & Guan, 2020).

## 3. TECHNIQUES FOR GENERATING AND DETECTING DEEPFAKES

Deepfakes have potential benefits in medical imaging but reliable detection of manipulated images is crucial. This section explores the techniques for creating and detecting deepfakes. Fig. 3 shows a taxonomy of deepfake generation and detection techniques and provides a clear overview of the methods. It also highlights the ongoing challenge of staying ahead of deepfake technology and the need for highly adaptable detection methods.

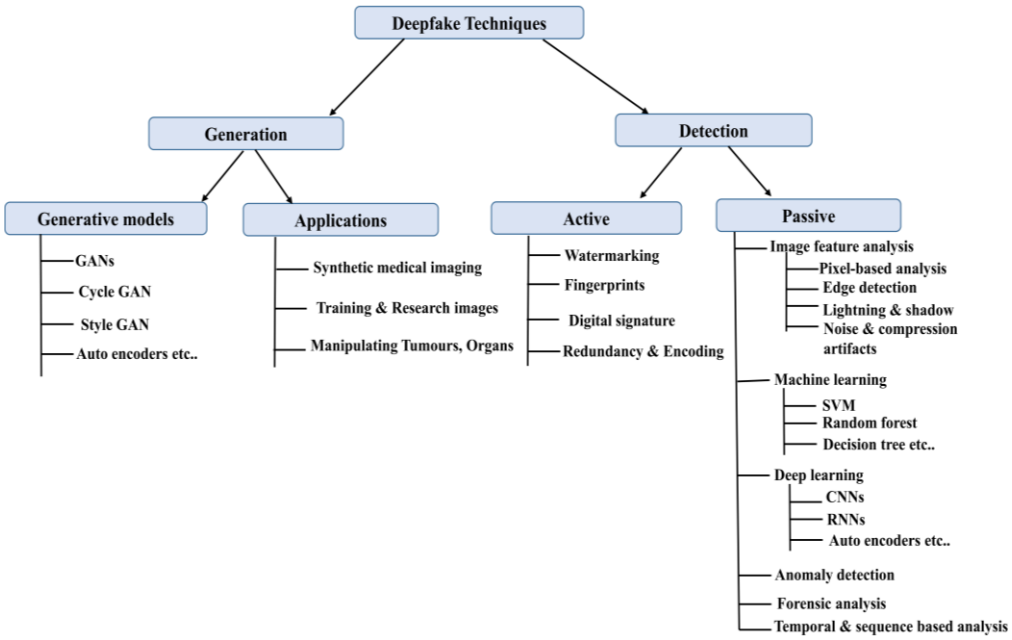
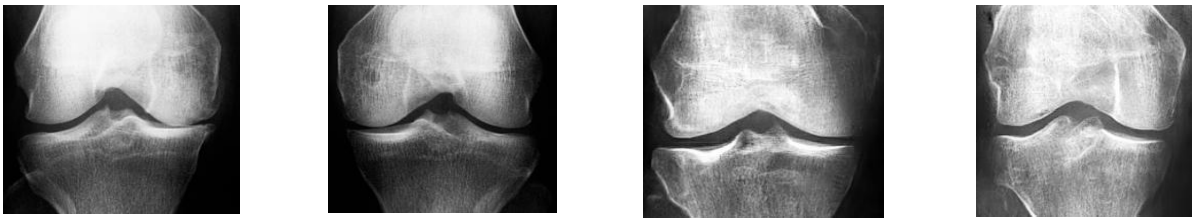



**Fig. 3. Taxonomy of deepfake generation and detection techniques**

### 3.1. Generation techniques

Deep learning is central to creating deepfake medical images. Generative Adversarial Networks (GANs) are especially important. Different GAN architectures have unique strengths and weaknesses. They are used in various medical imaging applications. Tab. 2 summarizes the advantages and limitations of several key GAN models. These include Pix2Pix GAN, CycleGAN, StyleGAN, StarGAN and Progressive GAN. Fig. 4 shows real and GAN-generated CT scans. Autoencoders are also used for deepfake generation. They can reconstruct and manipulate images. The following subsections explore these techniques in detail. Specific applications and ethical implications are also discussed.



**(a)** **(b)** **(c)** **(d)**

**Fig. 4. Real vs. GAN-Generated CT Scans. (a), (b) Real CT scans (Chen P, 2018). (c), (d) GAN-generated synthetic CT scans (Prezja et al., 2022). The high level of realism achieved by GANs poses significant challenges for current detection algorithms**

**Tab. 2. Summary of state-of-the-art generation techniques used in medical imaging**

| Techniques | Description | Applications in medical imaging | Advantages | Limitations |
|---|---|---|---|---|
| GANs | GANs have two nets, a generator that creates images and a discriminator that checks if they look real. | - Create fake MRI, CT, and X-ray images for training data.<br>- Simulate rare diseases for learning. | - Produces high-quality images.<br>- Works with many types of medical images. | - Training can be unstable.<br>- Requires large data sets. |
| Pix2Pix GAN | Pix2Pix GAN is a type of GAN that works with paired data to translate one image into another. | - Translate MRI scans to CT scans and vice versa.<br>- Segmenting body parts for medical analysis. | -Produces accurate translations with paired data. | - Requires paired data sets, which can be difficult to obtain.<br>- Less flexible when working with unpaired data. |
| CycleGAN | CycleGAN works like Pix2Pix but without the need for paired data. | - Convert MRI scans to CT scans.<br>- Improve image quality.<br>- Simulate treatment outcomes. | - Can use unpaired data, making it more flexible. | - May produce artifacts that degrade image quality.<br>- May not capture small details needed for diagnosis. |
| StyleGAN | StyleGAN creates images with control over details such as lighting and texture. | - Create varied synthetic medical images.<br>- Enhance features for training purposes. | - Creates detailed and realistic images. | - Requires a lot of computing power.<br>- Complex design that requires expertise. |
| StarGAN | StarGAN is a model that works with multiple types of images in one. | - Convert MRI, CT, and X-ray images with a single model.<br>- Enhance multiple features in an image. | - Efficient by using one model for multiple tasks.<br>- Scalable for various medical imaging tasks. | - Training is complex for multiple image types.<br>- Some transformations may degrade quality. |
| Progressive GAN | Progressive GANs systematically generate images, starting with low resolution and adding detail over time. | - Create high-resolution medical images.<br>- Simulate detailed body structures. | - Training is more stable.<br>- Produces high-resolution images with fine detail. | - Requires a lot of processing power.<br>- Takes more time to train. |
| Autoencoders | Autoencoders are used to compress and recover data. | - Create anonymous patient images.<br>- Identify problems by comparing real images to recreated images. | - Reduces data size.<br>- Produces various synthetic images. | - Images may not be as sharp as GANs.<br>- Limited control over image detail. |

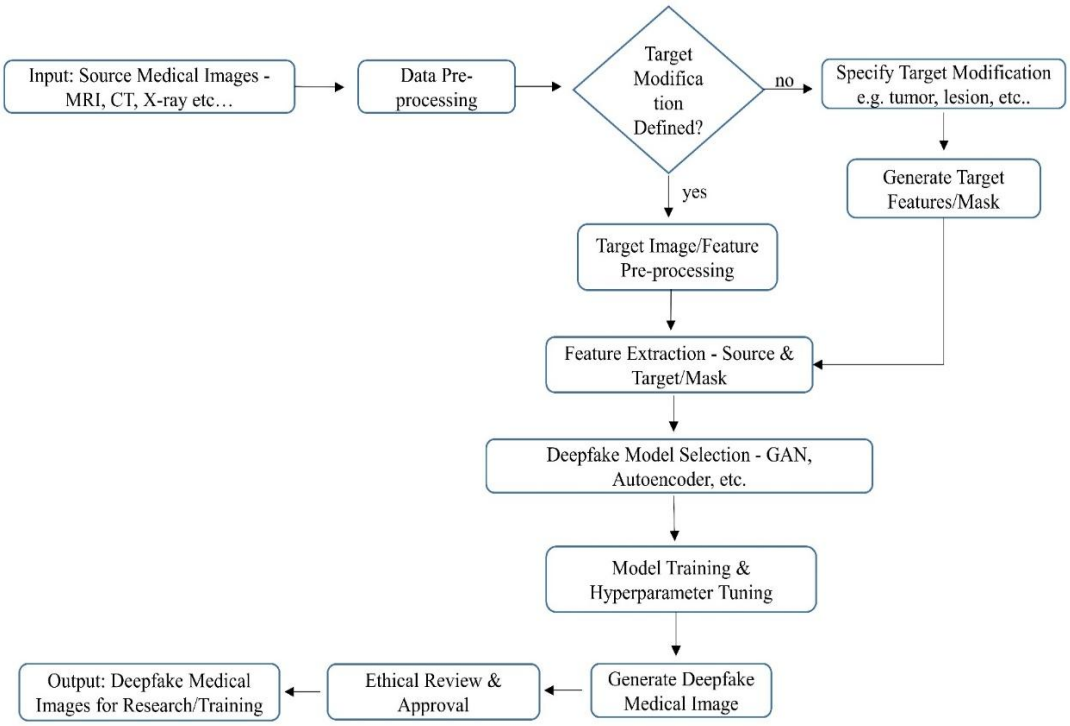



**Fig. 5. Flowchart for illustrating the steps involved in generating deepfake medical images**

## 3.2. Detection techniques

Deepfake detection is critical to patient safety and trust in healthcare. Medical images should be authentic and unaltered. This section examines several deepfake detection techniques. Some of these methods use markers for authentication. Others have analyzed images for signs of tampering. More advanced methods have used machine learning and deep learning. Emerging technologies such as blockchain and quantum computing are also being explored. Tab. 3 summarizes these techniques and highlights their strengths and weaknesses. Because each method has limitations, a multi-pronged approach is needed for robust deepfake detection in medical imaging.

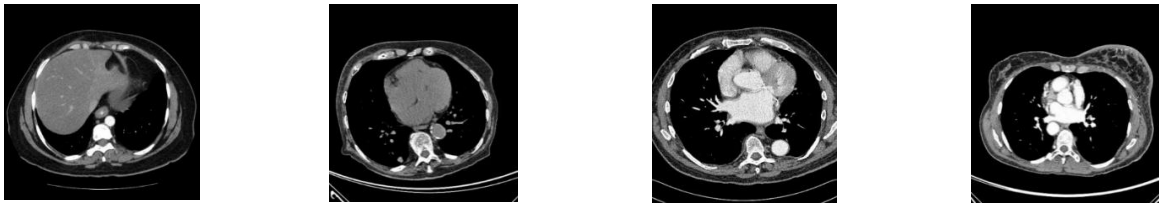
**Tab. 3. Summary of state-of-the-art detection techniques used in medical imaging**

| Techniques | Description | Advantages | Limitations |
|---|---|---|---|
| Active Detection Methods | Embed unique markers such as watermarks or digital signatures into images during creation for later authentication. | - Real-time verification<br>- Easy to implement. | - Can be modified by attackers.<br>- Adds additional data. |
| Passive Detection Methods | Analyzes features in images such as pixel patterns or noise levels to detect manipulation without adding markers. | - No changes to image creation required.<br>- Detects many manipulations. | - Requires complex analysis.<br>- High processing power.<br>- Effectiveness depends on image quality. |
| Machine Learning Models | Uses algorithms such as CNNs and RNNs to detect patterns of manipulation in medical images. | - High accuracy with large data sets.<br>- Can detect subtle changes.<br>- Can improve with more data. | - Requires large data sets<br>- Vulnerable to attacks<br>- Takes a lot of computing resources. |
| Deep Learning Models | Uses advanced neural networks to detect deep fakes by capturing complex features in images. | - High performance for complex tasks.<br>- Adaptable to different manipulations. | - Requires high resources<br>- Difficult to interpret decisions.<br>- Requires expert development. |
| Blockchain-Based Verification | Uses blockchain to create secure records of medical images and ensure they are not tampered with. | - Tamper proof.<br>- Increases transparency.<br>- Enables secure sharing. | - Difficult to integrate with systems.<br>- Scalability issues.<br>- Requires widespread adoption. |
| Quantum Computing Approaches | Uses quantum algorithms to accelerate and improve deepfake detection by analyzing data faster. | - High processing speed.<br>- Handles large amounts of data well.<br>- Can find complex patterns. | - Still in early stages, expensive.<br>- Needs more development for real-world use. |

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

The main goal of this experiment was to test how well different deep learning models can detect deepfakes in medical images. The study focused on models such as ResNet-50, EfficientNet-B0, DenseNet-121, VGG16, InceptionV3, and Vision Transformers (ViT) to see how they identify altered medical images. It also aimed to understand the strengths and weaknesses of each model and provide insights for improving deepfake detection in medical domains.

This study used the CT-GAN dataset (Mirsky et al., 2019), which is a well-known benchmark for deepfake detection in medical images, specifically lung CT scans. The dataset included four categories: True Benign (TB), True Malignant (TM), False Benign (FB), and False Malignant (FM). TB and TM images contain original CT scans that may or may not contain cancerous areas. These are considered true images for classification. FB and FM contain manipulated images where FB shows benign conditions with tumors removed and FM adds fake tumors to simulate malignant conditions. These are considered fake images for this experiment. Figure 6 shows examples of real and fake images from this dataset.



(e) (f) (g) (h)

**Fig. 6. Samples of CT images from CT GAN dataset True benign (e), True malignant (f), tumor removed using GAN (g), and tumor injected using GAN (h)**

All CT scans were in DICOM format, which contains important metadata used in medical imaging. To make the images compatible with deep learning models that work better with JPEG images, the DICOM files were converted to JPEG. This conversion kept the image resolution and quality high to avoid losing important details, although some metadata and image quality were slightly degraded.

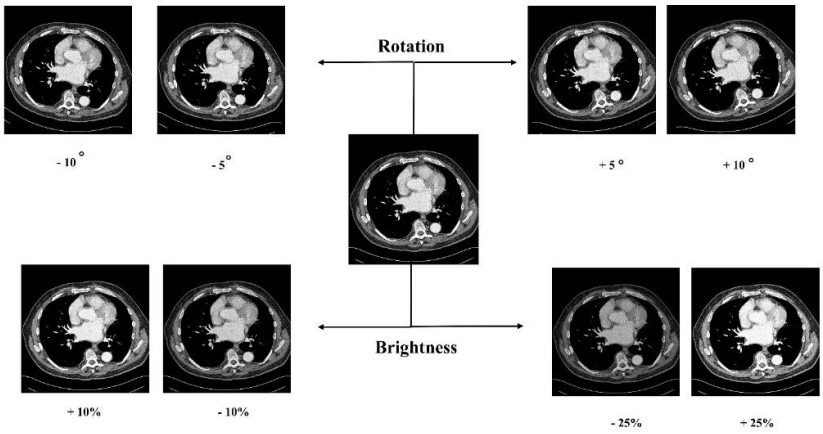



**Fig. 7. Illustration of rotation and brightness-based data augmentation on CT-GAN dataset**

To address the class imbalance in the dataset, data augmentation was applied. The dataset initially contained fewer fake images than real images. Data augmentation was applied only to the fake images. Eight data enhancements were applied to each fake image. These augmentations included rotations and brightness adjustments. The rotations included angles of ±5° and ±10°. This created four variations per image by rotation. Brightness adjustments included changes of ±10% and ±25%. This resulted in four brightness variations per image. Combining these enhancements creates eight different variations for each fake image. (Figure 7). This specific augmentation strategy directly addresses the class imbalance and increases the number of fake images in the dataset. These methods were chosen to improve the generalizability of the model without distorting the medical images.

The following preprocessing steps were applied to all images. First, the images were resized using bilinear interpolation. Images were resized to a uniform size of 224 × 224 pixels. This ensured consistency across the datasets. Center cropping was then performed. The images were center cropped to a size of 224x224 pixels. Pixel intensity normalization was also performed to standardize the data and improve model training. Tab. 4 shows the distribution of samples across the training, validation, and test sets.

**Tab. 4. Number of samples in the dataset splitting strategy**

| Split | Real images | Fake images | Total |
|---|---|---|---|
| Train | 1754 | 600 | 2354 |
| Val | 379 | 129 | 508 |
| Test | 379 | 126 | 505 |
| Total | 2512 | 855 | 3367 |

Google Colab Pro, which provides powerful computing resources, was used to train and evaluate the models. An NVIDIA L4 GPU with 24 GB of memory was used to speed up training and testing. The system also used CUDA 12.2 and cuDNN 8.9 for parallel processing. The models were developed in Python 3.8.8 using the PyTorch framework to ensure a stable platform for the experiments. Tab. 5 lists the hyperparameters and values used during training.

**Tab. 5. Hyperparameters and corresponding values**

| Hyperparameters | Value |
|---|---|
| Learning Rate | 0.001 |
| Batch Size | 64 |
| Optimizer | Adam |
| Epochs | 50 |
| Weight Decay | 0.0001 |
| Dropout Rate | 0.5 |
| Activation Function | ReLU |
| Loss Function | Weighted Cross-Entropy |

To evaluate the performance of the models, metrics such as accuracy, precision, recall, F1 score, confusion matrices, and confidence scores were used. The results of the experiments show the advantages and disadvantages of different deep learning models for deepfake detection in medical images. The performance metrics (Table 6) show that all models - ResNet-50 EfficientNet-B0 DenseNet-121 VGG16 InceptionV3 and Vision Transformers - performed well with high accuracy, precision, recall and F1 scores in distinguishing between real and fake lung CT scans.

InceptionV3 performed best with near perfect accuracy and F1 score. VGG16 ResNet-50 and DenseNet-121 also performed well, with accuracy and precision above 99% for the best models. These results suggest that classical convolutional neural networks (CNNs), originally designed for image classification, can be adapted to detect deepfake manipulations in medical images.

ViT and EfficientNet-B0 performed well, but showed lower results compared to the top CNN models. ViT, in particular, showed that while transformer-based models are successful in many computer vision tasks, they may need more adjustment or fine-tuning to work better for deep-fake detection in medical imaging. Factors such as small dataset size and subtle changes in the fake images may have affected the performance of these models.

**Tab. 6. Performance metrics of deep learning models on the CT-GAN dataset for medical image deepfake detection**

| Models | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Resnet 50 | 99.60 | 99.61 | 99.59 | 99.59 |
| Efficientnet b0 | 96.44 | 96.60 | 96.44 | 96.34 |
| VGG16 | 99.60 | 99.61 | 99.60 | 99.60 |
| InceptionV3 | 99.80 | 99.78 | 99.80 | 99.79 |
| DenseNet-121 | 99.01 | 99.05 | 99.01 | 99.02 |
| Vision Transformers (ViT) | 97.43 | 97.43 | 97.42 | 97.40 |

The confusion matrices (Figure 8) also show that certain models, such as EfficientNet-B0 and ViT, had a slightly higher rate of misclassifying fake images as real. This shows how difficult it is to detect subtle changes in medical images.

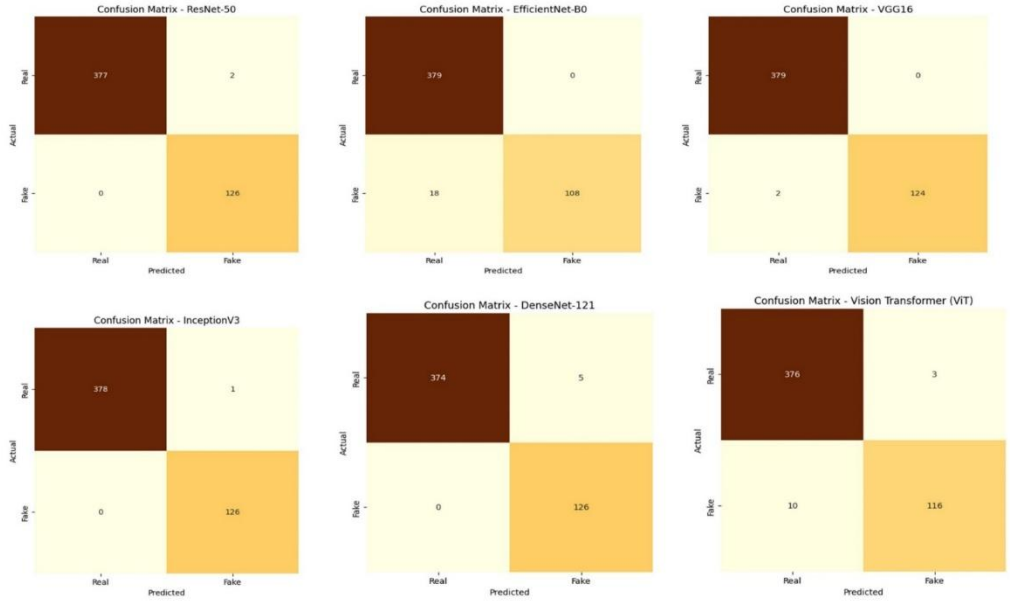



**Fig. 8. Confusion matrix-based comparative analysis of deep learning models on the CT-GAN dataset for medical deepfake detection**

The distributions of confidence scores (Figure 9) illustrate the challenges of deepfake detection. Most models produced high confidence scores for real images, indicating a high level of confidence in their authenticity.

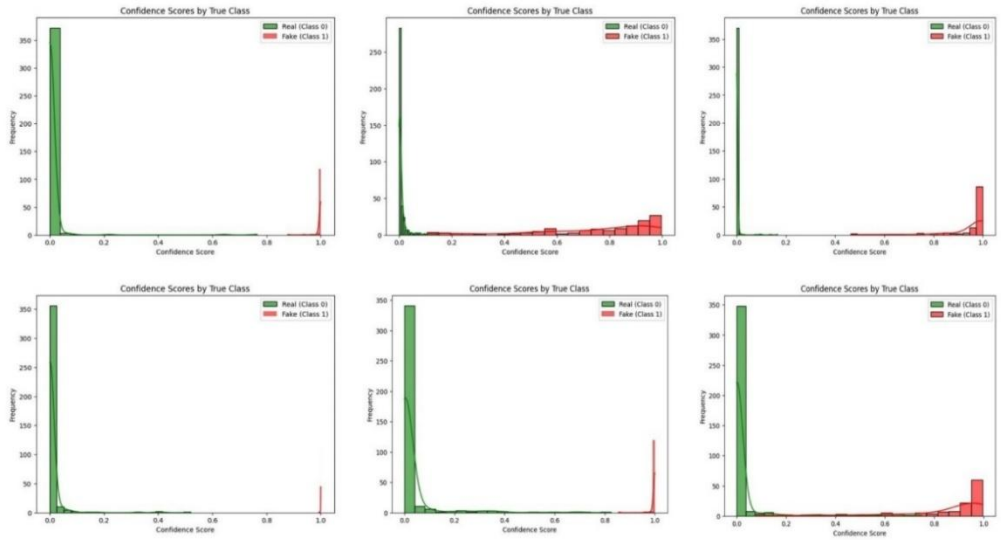



**Fig. 9. Confidence score distribution for real and fake classes across different models. The plots (from top left to bottom right) represent the models in the following order: ResNet-50, EfficientNet-B0, VGG16, InceptionV3, DenseNet-121, and Vision Transformer (ViT)**

For the fake images, however, the confidence scores varied more significantly. The best models achieved high confidence scores for the fake images. This indicated a clear ability to discriminate between real and fake images under optimal conditions. However, models with lower performance showed a wider distribution of confidence scores for fake images. This score distribution suggests greater uncertainty in deep fake detection

for these models and highlights a key limitation of the dataset. The CT-GAN dataset had fewer fake images than real images. This imbalance and the limited diversity of the dataset can adversely affect model performance. Models may not generalize well to new deep fakes. Although deep learning models are good at detecting real images, it is difficult to detect subtle deep fakes. In this study, only the CT-GAN dataset was used due to the lack of publicly available deepfake datasets. The lack of diverse datasets is a major challenge in medical deepfake detection research.

It is important to recognize the limitations of this experimental study. This study provides a definitive assessment of whether an image has been manipulated; however, it does not determine the specific location of the manipulation. Determining the exact location of the manipulation requires further research and model development. These limitations suggest that while deep learning models can effectively classify manipulated images, further development is needed to improve their ability to detect and localize changes.

These results have important implications for clinical practice. The high accuracy of the best models suggests that deep learning-based tools can be used in medical imaging systems to flag suspicious images and help physicians maintain diagnostic accuracy. However, not all models perform equally well, so careful model selection and validation are important.

## 5. CHALLENGES IN DETECTING AND PREVENTING DEEPFAKES IN MEDICAL IMAGING

Detecting and preventing deepfakes in medical imaging is challenging. A key challenge is the rapid evolution of deepfake technology (Cheng, 2024). This makes detection very difficult. The lack of standardized data sets makes it difficult to develop reliable methods. There are also many ethical and legal concerns about privacy, data security, bias, and accountability. Tab. 7 provides more details about these challenges. It also lists implications and potential solutions.

**Tab. 7. Challenges in detecting and preventing deepfakes in medical imaging**

| Challenge | Description | Implications | Solutions |
|---|---|---|---|
| Rapid Advancement of Deepfake Technology | Generative models such as GANs and StyleGAN are improving and making it harder to identify fake medical images. | - Fake images are harder to detect.<br>- Increased risk of misdiagnosis and harm. | - Build systems that adapt to new deepfake models.<br>- Use advanced tools to catch small errors. |
| Lack of Standardized Datasets | There are not enough datasets of real and fake medical images for training. | - Difficult to train and improve detection models.<br>- Difficult to reproduce research results. | - Collaborate with hospitals and medical institutions to create shared datasets.<br>- Ensure that datasets include all types of medical images (MRI, CT, and X-ray). |
| Ethical and Legal Issues | Concerns about patient privacy, data security, algorithmic bias, and legal liability when using deepfake detection systems. | - Risk of data breaches and misuse<br>- Bias in AI models can lead to unfair healthcare outcomes. | - Establish clear ethical rules and legal standards for AI in medical imaging.<br>- Use synthetic or anonymized data to protect privacy. |
| Algorithmic Bias and Fairness | AI systems can be biased by unbalanced training data. | - Unfair diagnostic results.<br>- Loss of trust in AI tools. | - Train models on diverse data sets representing all groups.<br>- Be transparent about model training and evaluation. |
| Legal Responsibility and Accountability | Uncertainty about who is responsible when AI systems fail. | - Fear of legal repercussions makes hospitals reluctant to use AI tools. | - Establish clear legal rules of responsibility and accountability.<br>- Promote transparency in the use of AI systems. |
| Integration into Clinical Workflows | Detection systems may be too slow or complex to use in daily clinical practice. | - Disrupts clinical operations.<br>- Delays patient care. | - Enhance detection systems to be faster and more efficient.<br>- Leverage a robust infrastructure to support real-time processing. |
| Privacy Concerns and Data Security | Using large amounts of sensitive medical data to train deepfake detection models can create privacy risks. | - Risk of data breaches and privacy violations.<br>- Loss of confidence in healthcare systems. | - Use strong encryption and data backup.<br>- Ensure compliance with privacy laws through regular audits. |

## 6. POLICY RECOMMENDATIONS TO ADDRESS DEEPFAKES IN MEDICAL IMAGING

Deepfakes are a growing problem in medical imaging. Explicit policies are needed to manage their impact. To protect medical data, these policies should emphasize strict regulations, collaboration, ethical use of AI, and cybersecurity.

Clear regulations are critical to addressing the challenges of deepfakes in healthcare. International cooperation is essential to create a global exchange of digital medical information. Organizations such as the WHO(World Health Organization ) and the ITU(International Telecommunication Union ) should make efforts to develop universal standards and guidelines (Łabuz, 2023). These guidelines should define acceptable use and outline penalties for misuse. Collaboration between AI experts and medical professionals is important to ensure ethical and effective implementation. GDPR provides a useful framework for data protection and privacy. Its principles can be adapted to address the risks of deepfakes. National regulations need to be updated to address issues such as liability for misdiagnosis, standards of proof, and intellectual property concerns. Regulatory agencies, such as the FDA, need to update their guidelines to incorporate AI-powered tools. This will ensure the safe and responsible use of deepfakes in medical imaging.

Deepfakes pose significant challenges that require a collaborative effort to solve. Stakeholders must work together, including technology companies, medical institutions, researchers, policymakers, and patient advocates. Public-private partnerships can accelerate the development of detection tools and improve data security. Joint projects can explore innovative solutions, such as the use of blockchain for secure provenance tracking. Data sharing plays an important role in improving detection accuracy by providing access to larger and more diverse data sets. Therefore, strict privacy regulations must be enforced to protect sensitive medical information. Organizations such as HL7 can update data exchange standards and promote interoperability to ensure seamless and secure information exchange.

Ethical considerations are essential to the use of AI in medical imaging. Therefore, AI systems must be developed and used responsibly. This includes transparency and accountability. Deepfake detection models should be easy to understand. Informed consent was necessary when collecting patient data. It is also important to address potential biases in the algorithms. Independent audits can help ensure that ethical standards are met. Professional organizations such as RSNA and ACR should work together to create ethical guidelines and certification programs.Implement these programs to promote responsible use and accountability.

Medical imaging systems are sensitive and must be well protected. Robust cybersecurity is essential to combat tampering and misuse. This requires enhanced data security measures, secure image storage and transmission mechanisms. Tamper-proof authentication methods are critical to protecting data. Research and development should focus on securing the digital infrastructure. Technologies such as blockchain and watermarking can improve data integrity and traceability. Security audits should be conducted regularly to identify vulnerabilities and improve security.

## 7. RESEARCH GAPS AND FUTURE DIRECTIONS

Several key research gaps hinder the safe and effective use of deepfakes in medical imaging. However, these gaps must be addressed. Future research directions should be explored. These steps are outlined below.

Current detection models lack access to diverse medical image datasets that include both real and fake images. Future research should focus on creating datasets that cover many medical conditions and imaging types (Seow et al., 2022). These datasets should include diverse patient populations to improve model performance. Collaboration with hospitals is needed to collect these images while respecting privacy laws.

Solving the challenge of deepfakes requires expertise from AI, cybersecurity, and ethics. Combining knowledge from these fields will help create more secure and ethical detection systems (Malatji & Tolah, 2024). Future studies should combine AI experts, cybersecurity experts, and ethicists to develop holistic solutions.

Current detection methods are often too slow for clinical use. Research should focus on creating lightweight models that can analyze images in real time without sacrificing accuracy. This would help integrate detection systems into healthcare workflows and enable rapid verification of medical images (Javed et al., 2024).

Many current models focus on image artifacts without considering the clinical context. Adding patient history and diagnostic data can improve the accuracy of deepfake detection. Context-aware models are better at identifying inconsistencies that are consistent or inconsistent with expected clinical scenarios.

Adding explicable AI to deepfake detection models can clarify their decision-making processes. Future research should explore methods to make AI models more understandable to clinicians. This will aid in the acceptance and validation of these tools in medical practice (Tsigos et al., 2024).

Addressing these gaps will help advance the use of deepfake technology in medical imaging in a safer, more ethical, and more effective manner. This will lead to better patient care and increased confidence in the healthcare system.

## 8. CONCLUSIONS

This review examines the role of deepfakes in medical imaging, highlighting their potential and challenges. The rapid growth of deepfake technology requires continuous improvements in detection methods. Deepfakes raise important ethical concerns in healthcare, such as risks to patient privacy, diagnostic accuracy, and public trust. Strong ethical guidelines and regulations are essential for the responsible use of deepfakes in medical imaging. Future studies should generate large and diverse ethically sourced datasets for training and testing deepfake detection systems. These datasets are critical for improving the accuracy and reliability of detection tools. It is also important to develop explainable real-time detection tools that can be easily integrated into clinical workflows.

Beyond detection, deepfakes have the potential to improve personalized medicine and medical education. They can be used to generate synthetic patient data for AI training, simulate rare diseases for education, and create personalized models for surgical planning. However, these applications require careful consideration of ethical issues and risks. A collaborative approach involving clinicians, researchers, ethicists, policymakers, and technologists is critical to developing ethical and effective AI systems in healthcare. This collaboration will ensure that deepfakes are used responsibly in healthcare, while protecting ethical standards, patient trust and safety.

### Authors Contributions

*Pradeepan P. and Gladston RAJ S. conducted the experiments and wrote the manuscript. Gladston RAJ S. and Juby GEORGE. reviewed and edited the manuscript. All authors have read and agreed to the final version of the manuscript.*

### Conflicts of Interest

*The authors declare no conflict of interest.*

### REFERENCES


Alheeti, K. M. A., Alzahrani, A., Khoshnaw, N., & Al-Dosary, D. (2022). 'Intelligent deep detection method for malicious tampering of cancer imagery. *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)* (pp. 25–28). IEEE. http://dx.doi.org/10.1109/CDMA54072.2022.00010

Budhiraja, R., Kumar, M., Das, M. K., Bafila, A. S., & Singh, S. (2022). 'MeDiFakeD: Medical deepfake detection using convolutional reservoir networks,. *2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT)* (pp. 1–6). IEEE. http://dx.doi.org/10.1109/GlobConPT57482.2022.9938172

Chen, P. (2018). *Knee osteoarthritis severity grading dataset*. Mendeley. Retrieved March 30, 2025 from https://data.mendeley.com/datasets/56rmx5bjcr/1

Chen, Y., & Esmaeilzadeh, P. (2024). Generative AI in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, *26*, e53008. https://doi.org/10.2196/53008

Cheng, X. (2024). Refining CycleGAN with attention mechanisms and age-Aware training for realistic Deepfakes. *Heliyon*, *10*(16), e36665. https://doi.org/10.1016/j.heliyon.2024.e36665

Cochran, J. D., & Napshin, S. A. (2021). Deepfakes: Awareness, concerns, and platform accountability. *Cyberpsychology, Behavior and Social Networking*, *24*(3), 164–172. https://doi.org/10.1089/cyber.2020.0100

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science* , *363*(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 289-293). IEEE. https://doi.org/10.1109/ISBI.2018.8363576

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). Deep fake image detection based on pairwise learning. *Applied Sciences*, *10*(1), 370. https://doi.org/10.3390/app10010370

Javed, M., Zhang, Z., Dahri, F. H., & Laghari, A. A. (2024). Real-time Deepfake video detection using eye movement analysis with a hybrid deep learning approach. *Electronics*, *13*(15), 2947. https://doi.org/10.3390/electronics13152947

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

Karaköse, M., Yetış, H., & Çeçen, M. (2024). A new approach for effective medical deepfake detection in medical images. *IEEE Access*, *12*, 52205-52214. https://doi.org/10.1109/ACCESS.2024.3386644

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4396-4405). IEEE. https://doi.org/10.1109/CVPR.2019.00453

Kim, Y. S., Song, H. J., & Han, J. H. (2022). A study on the development of deepfake-based deep learning algorithm for the detection of medical data manipulation. *Webology*, *19*(1), 4396–4409. https://doi.org/10.14704/web/v19i1/web19289

Łabuz, M. (2023). Regulating deep fakes in the Artificial Intelligence act. *Applied Cybersecurity & Internet Governance*, *2*(1), 1-42. https://doi.org/10.60097/acig/162856

Latif, G., Brahim, G. B., Mohammad, N., & Alghazo, J. (2024). Combating medical image tampering using deep transfer learning. *AIP Conference Proceedings*, *3034*, 040002. http://dx.doi.org/10.1063/5.0194668

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*, *5*, 883-910. https://doi.org/10.1007/s43681-024-00427-4

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mirsky, Y., Mahler, T., Shelef, I., & Elovici, Y. (2019). CT-GAN: Malicious tampering of 3D medical imagery using deep learning. *28th USENIX Security Symposium (USENIX Security 19)* (pp. 461-478). USENIX Association.

Motamed, S., Rogalla, P., & Khalvati, F. (2021). Data augmentation using generative adversarial networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. *Informatics in Medicine Unlocked*, *27*, 100779. https://doi.org/10.1016/j.imu.2021.100779

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., & Shen, D. (2017). Medical image synthesis with context-aware generative adversarial networks. In M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, & S. Duchesne (Eds.), *Medical Image Computing and Computer Assisted Intervention − MICCAI 2017* (Vol. 10435, pp. 417–425). Springer International Publishing. https://doi.org/10.1007/978-3-319-66179-7_48

Prezja, F., Paloneva, J., Pölönen, I., Niinimäki, E., & Äyrämö, S. (2022). DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. *Scientific Reports*, *12*, 18573. https://doi.org/10.1038/s41598-022-23081-4

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv, abs/1511.06434*. https://doi.org/10.48550/arXiv.1511.06434

Reichman, B., Jing, L., Akin, O., & Tian, Y. (2021). Medical image tampering detection: A new dataset and baseline. In A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, & R. Vezzani (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges* (Vol. 12661, pp. 266–277). Springer International Publishing. https://doi.org/10.1007/978-3-030-68763-2_20

S, A., & Narayan, S. (2024). Detection of GAN-manipulated medical images through deep learning techniques. *2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)* (pp. 1-6). IEEE. https://doi.org/10.1109/AMATHE61652.2024.10582065

Salah, K., Rehman, M. H. U., Nizamuddin, N., & Al-Fuqaha, A. (2019). Blockchain for AI: Review and open research challenges. *IEEE Access*, *7*, 10127–10149. https://doi.org/10.1109/access.2018.2890507

Seow, J. W., Lim, M. K., Phan, R. C. W., & Liu, J. K. (2022). A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, *513*, 351–371. https://doi.org/10.1016/j.neucom.2022.09.135

Sharafudeen, M., & Chandra, S. S. (2023). Medical deepfake detection using 3-dimensional neural learning. *Artificial Neural Networks in Pattern Recognition: 10th IAPR TC3 Workshop* (pp. 169-180). Springer International Publishing. https://doi.org/10.1007/978-3-031-20650-4_14

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*,60. https://doi.org/10.1186/s40537-019-0197-0

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE transactions on neural networks and learning systems*, *32*(11), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314

Tsigos, K., Apostolidis, E., Baxevanakis, S., Papadopoulos, S., & Mezaris, V. (2024). Towards quantitative evaluation of explainable AI methods for deepfake detection. *3rd ACM International Workshop on Multimedia AI against Disinformation*. *ArXiv, abs/2404.18649*. https://doi.org/10.48550/arXiv.2404.18649

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 39–52. https://doi.org/10.22215/timreview/1282

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., & Yuille, A. (2017). Adversarial examples for semantic segmentation and object detection. *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 1378-1387). IEEE. https://doi.org/10.1109/ICCV.2017.153

Zhang, J., Huang, X., Liu, Y., Han, Y., & Xiang, Z. (2024). GAN-based medical image small region forgery detection via a two-stage cascade framework. *PloS One*, *19*(1), e0290303. https://doi.org/10.1371/journal.pone.0290303

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2242-2251). IEEE. https://doi.org/10.1109/ICCV.2017.244