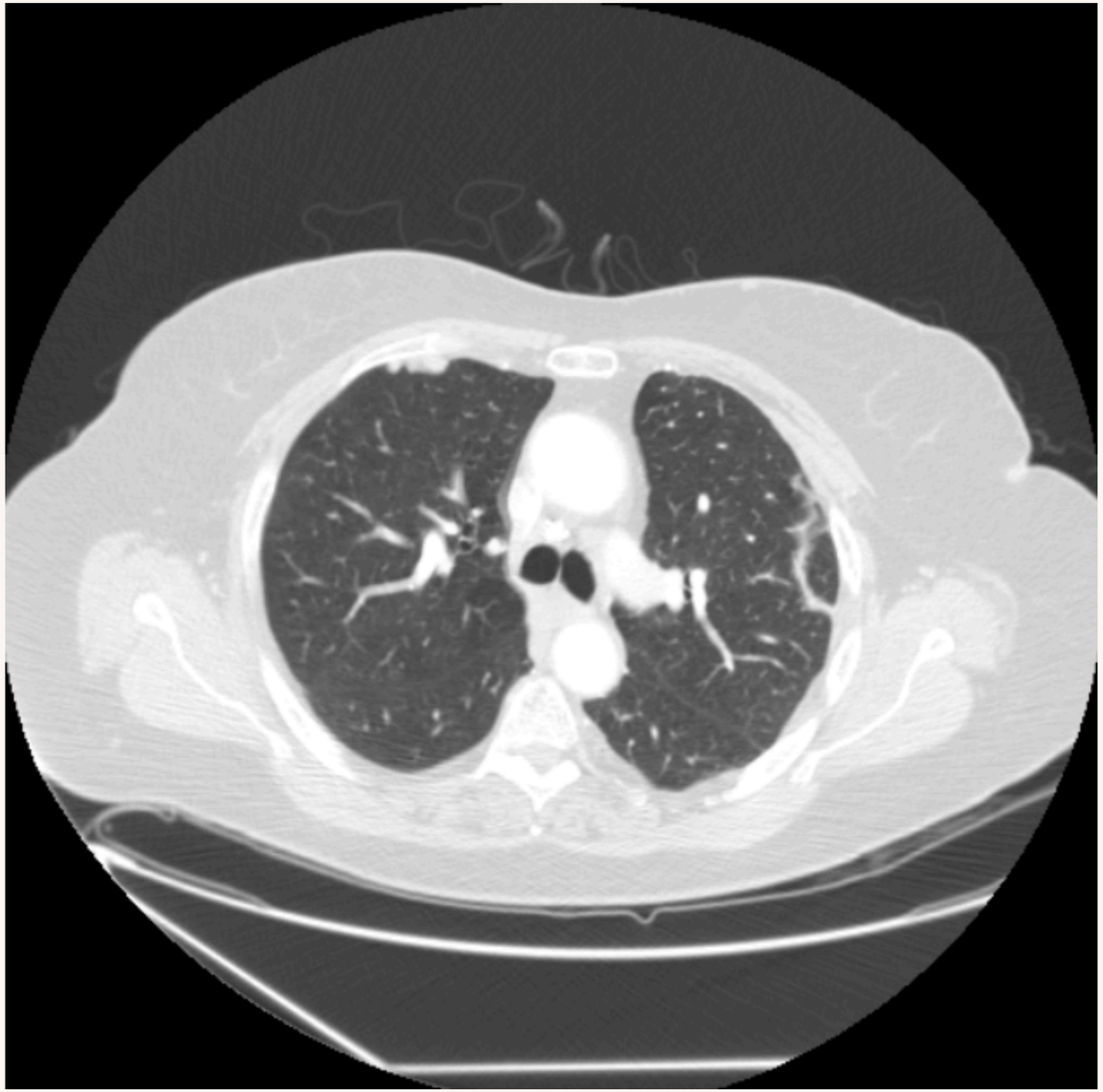
Detecting Image Forgery in Lung CT Scans

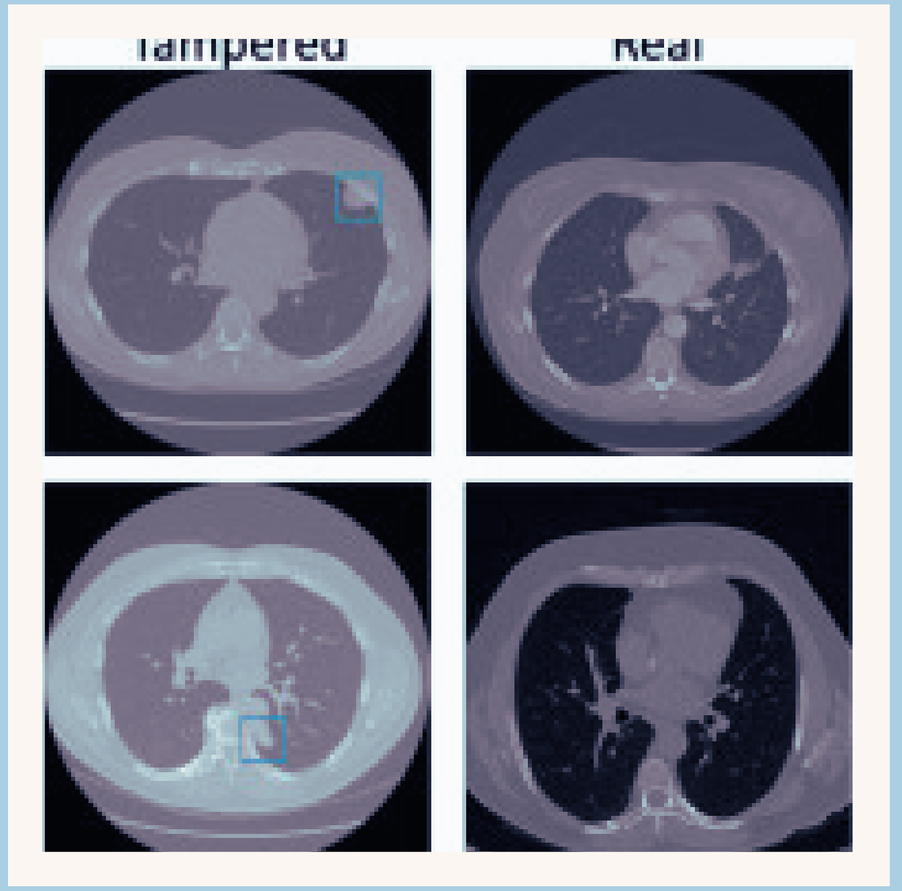
Saher Mohammed, Sneha Mahato & Uditi Bansal



LUNG CT SCAN

PROBLEM STATEMENT

How can we reliably detect tampering in 3D lung CT scans to ensure medical image authenticity?



Background

- **India’s health insurance sector loses an estimated ₹8,000-10,000 crore annually to fraud, with a significant share involving tampered medical documents submitted to support false claims.**
- **Over 120 million medical images in India, including CT scans and MRIs, have been exposed online due to poor security, increasing the risk of unauthorized access, manipulation, and misuse of diagnostic data.**

https://www.firstpost.com/tech/news-analysis/over-120-million-x-rays-ct-scans-exposed-on-the-internet-due-to-carelessness-of-hospitals-report-7898691.html

- **Real-World Case Example (2022): On August 31, 2022, Mahesh Kathrani allegedly submitted forged MRI reports and fabricated documents from an imaging center to falsely claim benefits for a cerebrovascular accident (stroke) under a critical illness insurance policy linked to his home loan.**
- **Significance: This case demonstrates how tampered medical records can be used to support fraudulent insurance claims, highlighting the need for automated systems to detect manipulated medical imaging data.**

https://timesofindia.indiatimes.com/city/rajkot/jailed-physiotherapist-now-accused-of-falsely-claiming-rs-22-5-lakh-mediclaim/articleshow/122096169.cms

Potential Applications

- **Medical Image Authentication:** Verifying the integrity and originality of CT scans before diagnosis or treatment planning.
- **Cybersecurity in Healthcare:** Strengthening protection against malicious modification of medical imaging records.
- **Regulatory and Legal Investigations:** Supporting investigations in cases of medical fraud, insurance disputes, or legal evidence involving radiology images.

Potential Impact

- **Enhanced Patient Safety:** Prevents incorrect treatments that may result from manipulated medical images.
- **Reduced Healthcare Fraud:** Minimizes risks of intentional image manipulation for insurance, legal, or malicious purposes.
- **Better Data Integrity in Research:** Ensures reliability of medical datasets used in clinical studies and AI model training.

Literature Review

GAN-Based Medical Image Small Region Forgery Detection via a Two-stage Cascade Framework

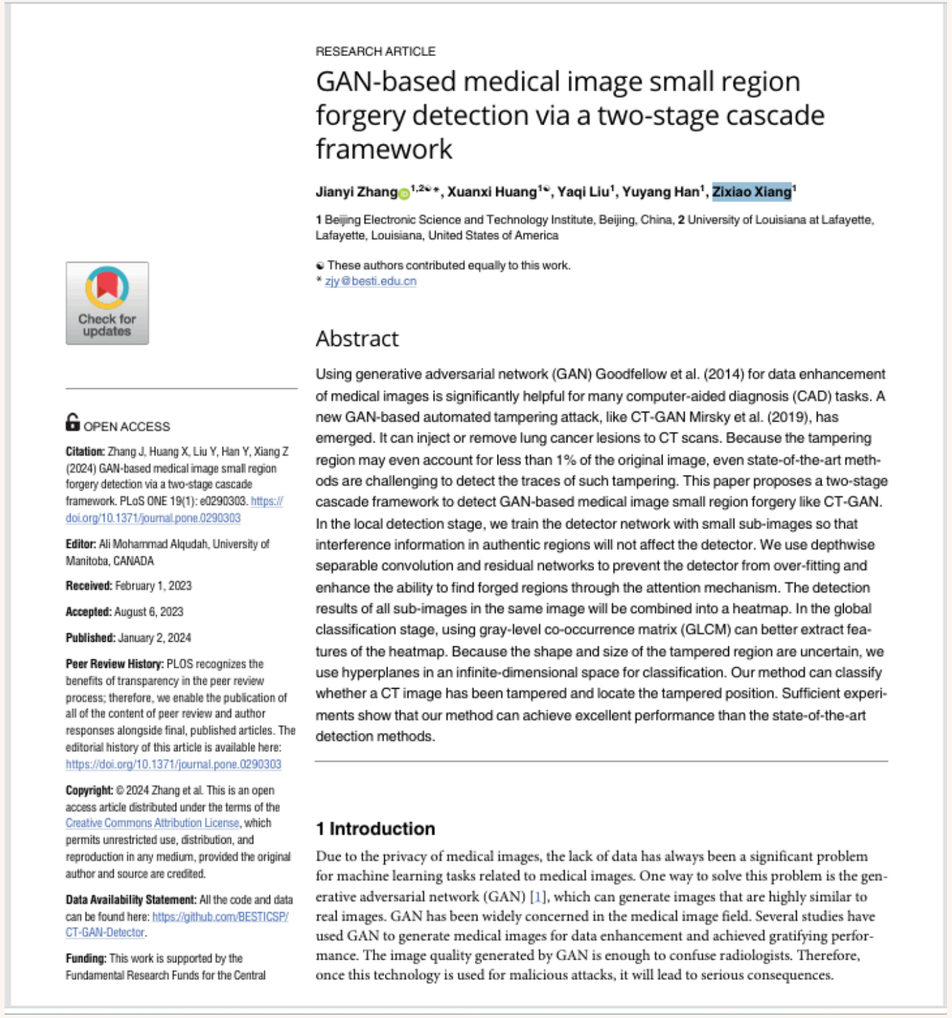
Jianyi Zhang, Xuanxi Huang, Yaqi Liu, Yuyang Han, Zixiao Xiang

Approach

- CT scans divided into multiple sub-images using sliding window.
- Deep learning model used to detect tampered regions.
- Focus on identifying GAN-generated manipulations in CT images.

Limitations

- Sub-image correlation ignored
- Adjacent slices in 3D CT volumes are not fully utilized.
- Reduced efficiency due to sliding window
- Large number of sub-images increases computational time.



RESEARCH ARTICLE

GAN-based medical image small region forgery detection via a two-stage cascade framework

Jianyi Zhang[1,2☯*], **Xuanxi Huang**[1☯], **Yaqi Liu**[1], **Yuyang Han**[1], **Zixiao Xiang**[1]

1 Beijing Electronic Science and Technology Institute, Beijing, China, **2** University of Louisiana at Lafayette, Lafayette, Louisiana, United States of America

☯ These authors contributed equally to this work.
* zjy@besti.edu.cn

Check for updates

OPEN ACCESS

Citation: 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

Editor: Ali Mohammad Alqudah, University of Manitoba, CANADA

Received: February 1, 2023

Accepted: August 6, 2023

Published: January 2, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0290303

Data Availability Statement: All the code and data can be found here: https://github.com/BESTICSP/CT-GAN-Detector.

Funding: This work is supported by the Fundamental Research Funds for the Central

Abstract

Using generative adversarial network (GAN) Goodfellow et al. (2014) for data enhancement of medical images is significantly helpful for many computer-aided diagnosis (CAD) tasks. A new GAN-based automated tampering attack, like CT-GAN Mirsky et al. (2019), has emerged. It can inject or remove lung cancer lesions to CT scans. Because the tampering region may even account for less than 1% of the original image, even state-of-the-art methods are challenging to detect the traces of such tampering. This paper proposes a two-stage cascade framework to detect GAN-based medical image small region forgery like CT-GAN. In the local detection stage, we train the detector network with small sub-images so that interference information in authentic regions will not affect the detector. We use depthwise separable convolution and residual networks to prevent the detector from over-fitting and enhance the ability to find forged regions through the attention mechanism. The detection results of all sub-images in the same image will be combined into a heatmap. In the global classification stage, using gray-level co-occurrence matrix (GLCM) can better extract features of the heatmap. Because the shape and size of the tampered region are uncertain, we use hyperplanes in an infinite-dimensional space for classification. Our method can classify whether a CT image has been tampered and locate the tampered position. Sufficient experiments show that our method can achieve excellent performance than the state-of-the-art detection methods.

1 Introduction

Due to the privacy of medical images, the lack of data has always been a significant problem for machine learning tasks related to medical images. One way to solve this problem is the generative adversarial network (GAN) [1], which can generate images that are highly similar to real images. GAN has been widely concerned in the medical image field. Several studies have used GAN to generate medical images for data enhancement and achieved gratifying performance. The image quality generated by GAN is enough to confuse radiologists. Therefore, once this technology is used for malicious attacks, it will lead to serious consequences.

Comparative Analysis of Deep Convolutional Neural Networks for Detecting Medical Image Deepfakes

Abdel Rahman Alsabbagh, Omar Al-Kadi

Method:

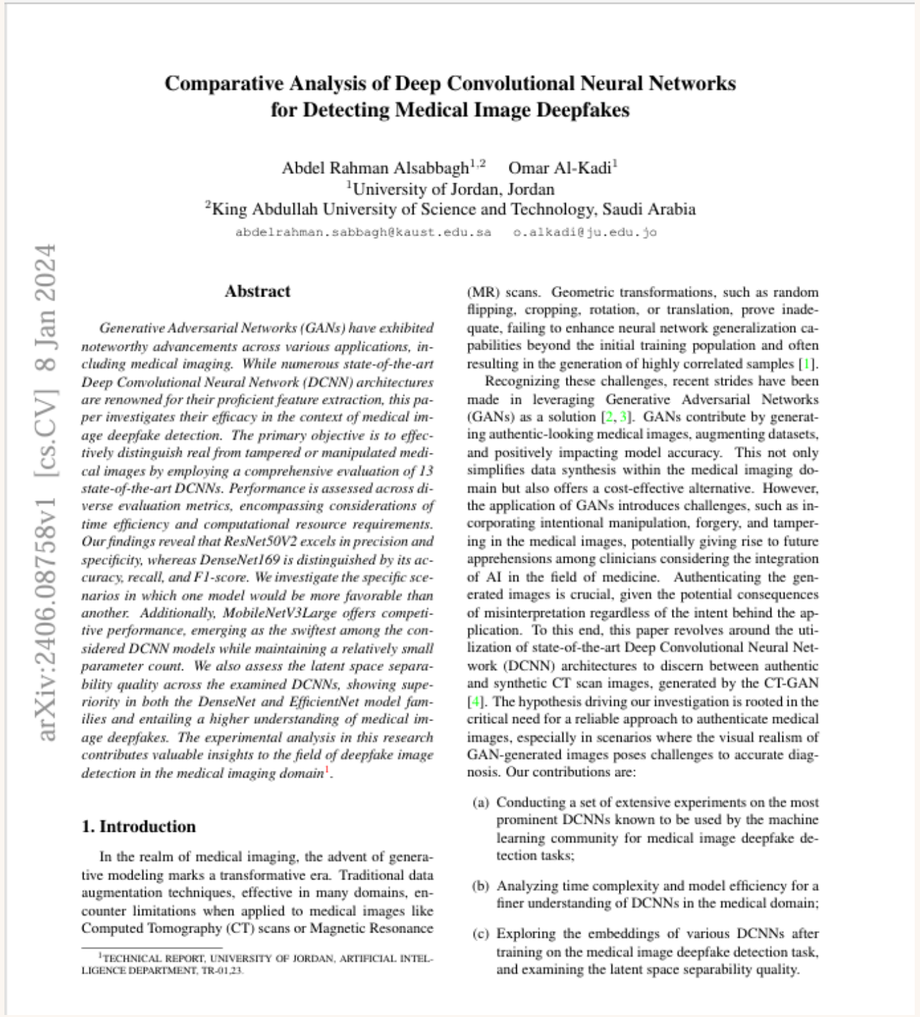
- Uses deep learning models for detecting medical image tampering
- Processes CT scan slices for classification

Key Features:

- Compares performance of different CNN-based models
- Evaluates effectiveness of models in detecting GAN-generated forgeries
- Focuses on improving accuracy and robustness across models

Limitations:

- Limited use of inter-slice (3D) relationships



Comparative Analysis of Deep Convolutional Neural Networks for Detecting Medical Image Deepfakes

Abdel Rahman Alsabbagh[1,2] Omar Al-Kadi[1]
[1]University of Jordan, Jordan
[2]King Abdullah University of Science and Technology, Saudi Arabia
abdelrahman.sabbagh@kaust.edu.sa o.alkadi@ju.edu.jo

Abstract

Generative Adversarial Networks (GANs) have exhibited noteworthy advancements across various applications, including medical imaging. While numerous state-of-the-art Deep Convolutional Neural Network (DCNN) architectures are renowned for their proficient feature extraction, this paper investigates their efficacy in the context of medical image deepfake detection. The primary objective is to effectively distinguish real from tampered or manipulated medical images by employing a comprehensive evaluation of 13 state-of-the-art DCNNs. Performance is assessed across diverse evaluation metrics, encompassing considerations of time efficiency and computational resource requirements. Our findings reveal that ResNet50V2 excels in precision and specificity, whereas DenseNet169 is distinguished by its accuracy, recall, and F1-score. We investigate the specific scenarios in which one model would be more favorable than another. Additionally, MobileNetV3Large offers competitive performance, emerging as the swiftest among the considered DCNN models while maintaining a relatively small parameter count. We also assess the latent space separability quality across the examined DCNNs, showing superiority in both the DenseNet and EfficientNet model families and entailing a higher understanding of medical image deepfakes. The experimental analysis in this research contributes valuable insights to the field of deepfake image detection in the medical imaging domain[1].

1. Introduction

In the realm of medical imaging, the advent of generative modeling marks a transformative era. Traditional data augmentation techniques, effective in many domains, encounter limitations when applied to medical images like Computed Tomography (CT) scans or Magnetic Resonance (MR) scans. Geometric transformations, such as random flipping, cropping, rotation, or translation, prove inadequate, failing to enhance neural network generalization capabilities beyond the initial training population and often resulting in the generation of highly correlated samples [1].

Recognizing these challenges, recent strides have been made in leveraging Generative Adversarial Networks (GANs) as a solution [2, 3]. GANs contribute by generating authentic-looking medical images, augmenting datasets, and positively impacting model accuracy. This not only simplifies data synthesis within the medical imaging domain but also offers a cost-effective alternative. However, the application of GANs introduces challenges, such as incorporating intentional manipulation, forgery, and tampering in the medical images, potentially giving rise to future apprehensions among clinicians considering the integration of AI in the field of medicine. Authenticating the generated images is crucial, given the potential consequences of misinterpretation regardless of the intent behind the application. To this end, this paper revolves around the utilization of state-of-the-art Deep Convolutional Neural Network (DCNN) architectures to discern between authentic and synthetic CT scan images, generated by the CT-GAN [4]. The hypothesis driving our investigation is rooted in the critical need for a reliable approach to authenticate medical images, especially in scenarios where the visual realism of GAN-generated images poses challenges to accurate diagnosis. Our contributions are:

(a) Conducting a set of extensive experiments on the most prominent DCNNs known to be used by the machine learning community for medical image deepfake detection tasks;

(b) Analyzing time complexity and model efficiency for a finer understanding of DCNNs in the medical domain;

(c) Exploring the embeddings of various DCNNs after training on the medical image deepfake detection task, and examining the latent space separability quality.

[1]TECHNICAL REPORT, UNIVERSITY OF JORDAN, ARTIFICIAL INTELLIGENCE DEPARTMENT, TR-01,23.

Leveraging 3D CNNs for Robust Detection of GAN-Generated Medical Image Forgeries

Sandhya L S and Ajeesh Ramanujan

Method:

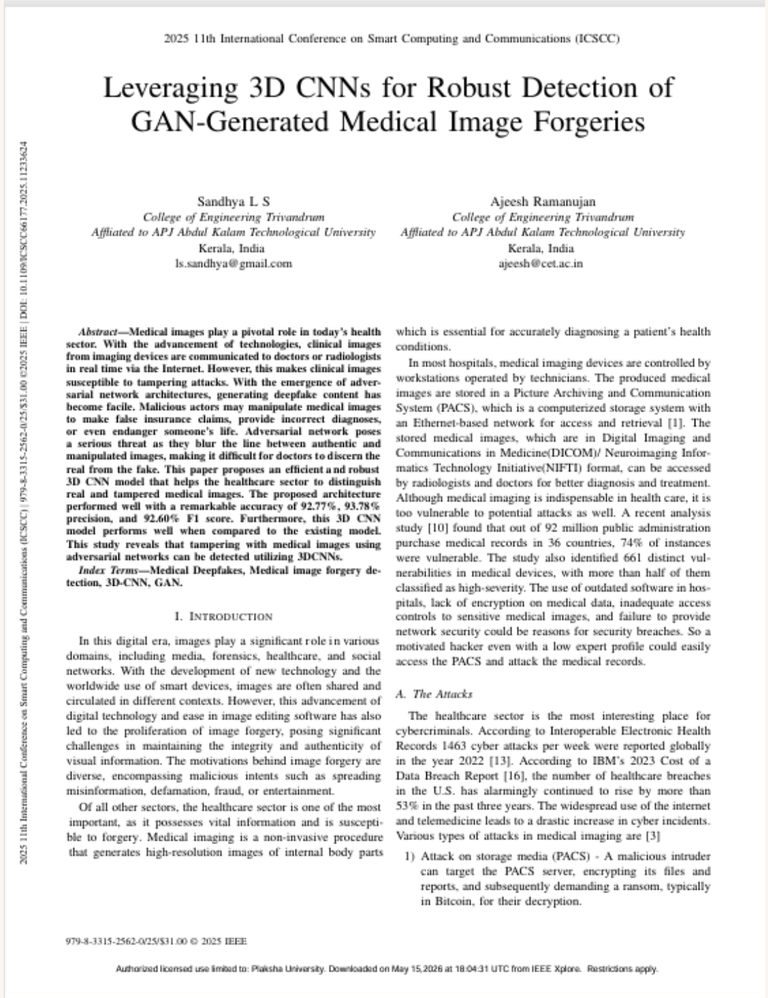
- Utilizes 3D Convolutional Neural Network (3D CNN) for detecting GAN-based medical image forgeries
- Processes multiple CT slices together (volumetric analysis)

Key Features:

- Captures inter-slice spatial relationships
- Uses 3D cubes (multi-slice inputs) instead of single images

Limitations:

- Requires high computational power (GPU-intensive)
- Trained on limited dataset



2025 11th International Conference on Smart Computing and Communications (ICSCC)

Leveraging 3D CNNs for Robust Detection of GAN-Generated Medical Image Forgeries

Sandhya L S
College of Engineering Trivandrum
Affliated to APJ Abdul Kalam Technological University
Kerala, India
ls.sandhya@gmail.com

Ajeesh Ramanujan
College of Engineering Trivandrum
Affliated to APJ Abdul Kalam Technological University
Kerala, India
ajeesh@cet.ac.in

2025 11th International Conference on Smart Computing and Communications (ICSCC) | 979-8-3315-2562-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSCC64177.2025.11233624

***Abstract*—Medical images play a pivotal role in today's health sector. With the advancement of technologies, clinical images from imaging devices are communicated to doctors or radiologists in real time via the Internet. However, this makes clinical images susceptible to tampering attacks. With the emergence of adversarial network architectures, generating deepfake content has become facile. Malicious actors may manipulate medical images to make false insurance claims, provide incorrect diagnoses, or even endanger someone's life. Adversarial network poses a serious threat as they blur the line between authentic and manipulated images, making it difficult for doctors to discern the real from the fake. This paper proposes an efficient a nd robust 3D CNN model that helps the healthcare sector to distinguish real and tampered medical images. The proposed architecture performed well with a remarkable accuracy of 92.77%, 93.78% precision, and 92.60% F1 score. Furthermore, this 3D CNN model performs well when compared to the existing model. This study reveals that tampering with medical images using adversarial networks can be detected utilizing 3DCNNs.**

***Index Terms*—Medical Deepfakes, Medical image forgery detection, 3D-CNN, GAN.**

I. Introduction

In this digital era, images play a significant role in various domains, including media, forensics, healthcare, and social networks. With the development of new technology and the worldwide use of smart devices, images are often shared and circulated in different contexts. However, this advancement of digital technology and ease in image editing software has also led to the proliferation of image forgery, posing significant challenges in maintaining the integrity and authenticity of visual information. The motivations behind image forgery are diverse, encompassing malicious intents such as spreading misinformation, defamation, fraud, or entertainment.

Of all other sectors, the healthcare sector is one of the most important, as it possesses vital information and is susceptible to forgery. Medical imaging is a non-invasive procedure that generates high-resolution images of internal body parts which is essential for accurately diagnosing a patient's health conditions.

In most hospitals, medical imaging devices are controlled by workstations operated by technicians. The produced medical images are stored in a Picture Archiving and Communication System (PACS), which is a computerized storage system with an Ethernet-based network for access and retrieval [1]. The stored medical images, which are in Digital Imaging and Communications in Medicine(DICOM)/ Neuroimaging Informatics Technology Initiative(NIFTI) format, can be accessed by radiologists and doctors for better diagnosis and treatment. Although medical imaging is indispensable in health care, it is too vulnerable to potential attacks as well. A recent analysis study [10] found that out of 92 million public administration purchase medical records in 36 countries, 74% of instances were vulnerable. The study also identified 661 distinct vulnerabilities in medical devices, with more than half of them classified as high-severity. The use of outdated software in hospitals, lack of encryption on medical data, inadequate access controls to sensitive medical images, and failure to provide network security could be reasons for security breaches. So a motivated hacker even with a low expert profile could easily access the PACS and attack the medical records.

A. The Attacks

The healthcare sector is the most interesting place for cybercriminals. According to Interoperable Electronic Health Records 1463 cyber attacks per week were reported globally in the year 2022 [13]. According to IBM's 2023 Cost of a Data Breach Report [16], the number of healthcare breaches in the U.S. has alarmingly continued to rise by more than 53% in the past three years. The widespread use of the internet and telemedicine leads to a drastic increase in cyber incidents. Various types of attacks in medical imaging are [3]

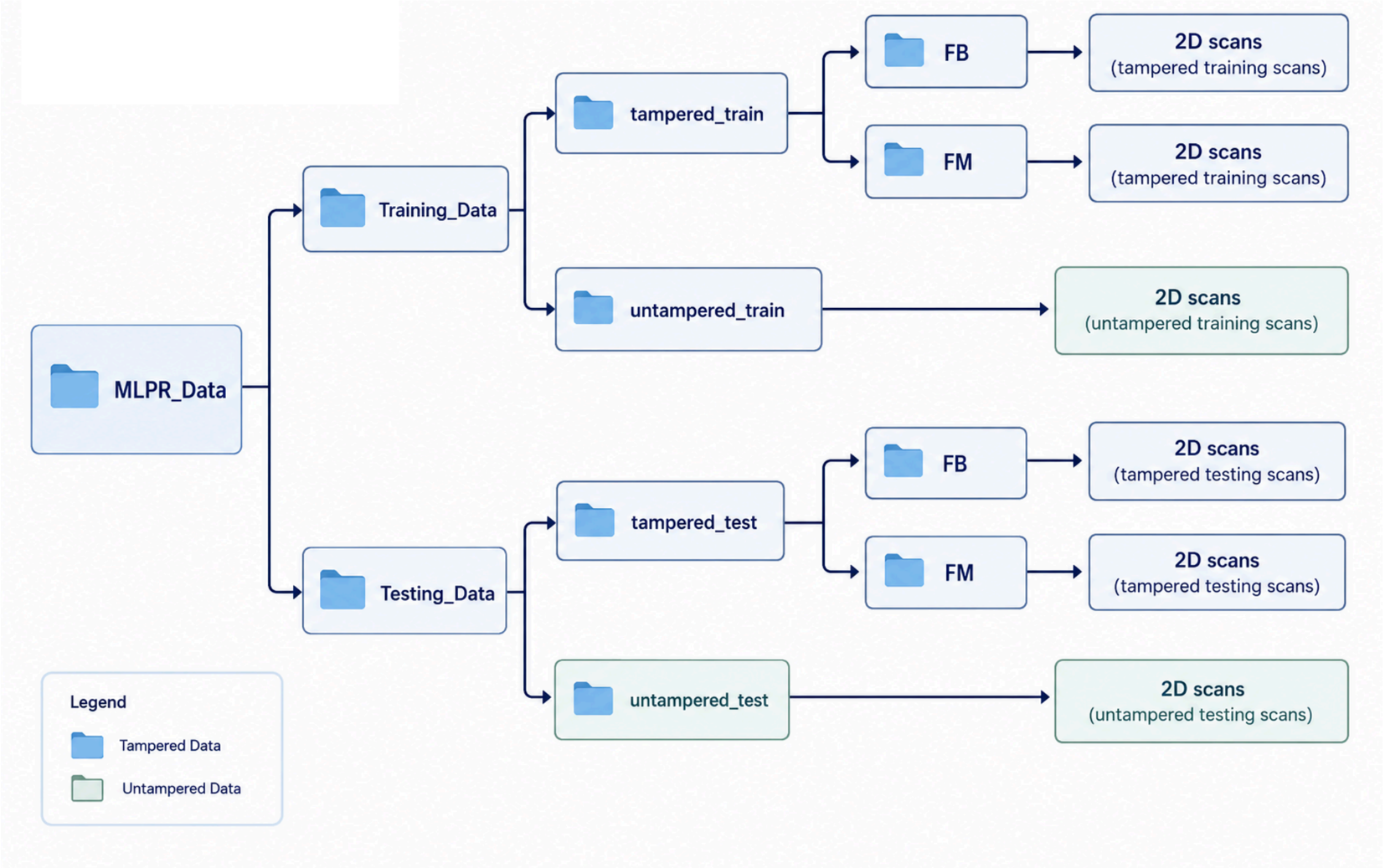
1) Attack on storage media (PACS) - A malicious intruder can target the PACS server, encrypting its files and reports, and subsequently demanding a ransom, typically in Bitcoin, for their decryption.

Dataset Organization & Preprocessing

Dataset Summary

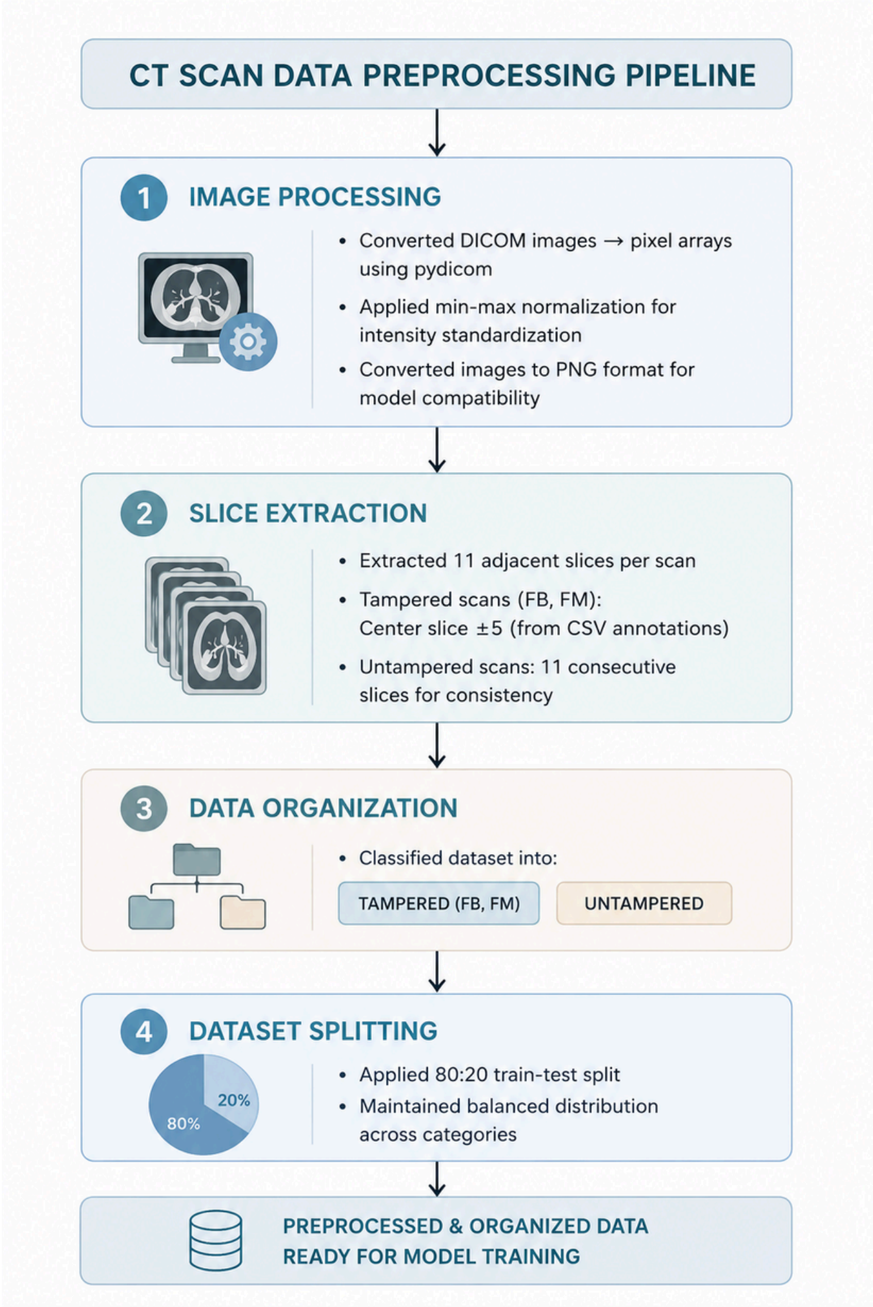
- tampered_train : 917
- untampered_train: 917

- tampered_test: 249
- untampered test: 249



Link to Dataset: https://archive.ics.uci.edu/dataset/520/deepfakes+medical+image+tamper+detection

https://www.cancerimagingarchive.net/collection/lidc-idri/



Rationale: Based on findings from prior literature, tampered/tumor regions often extend to adjacent slices within the [−5, +5] range; therefore, these neighboring slices were also included in the tampered dataset to preserve contextual information.

To examine our preprocessing approach, we visualized the labeled slices and the adjacent ones within different ranges. Figure 1 shows three separate *False Malignant* scans with their tumor region with a range of $[-5, +5]$ of the original label. We can see that indeed the tumor does not show up only on the labeled slice, but also on the neighbor slices, which means in return that there are more than 113 tampered images. At this stage, there were two choices to get a hold of as much tampered image data as possible: a) visualize all labeled slices and their adjacent slices and select manually all of the slices that include a tumor at the region of interest, or b) automatically select the labeled slices and their adjacent slices within a range of $[-5, +5]$ of the labeled slice. The initial option is deemed impractical; therefore, the latter is chosen. However, it is not a strict rule that all slices within the range of $[-5, +5]$ of the labeled slice are consistently tampered, as illustrated in Figure 1. We acknowledge the potential for errors associated with this choice. Nevertheless, in a significant majority of cases, nearly all adjacent slices within the range $[-5, +5]$ do exhibit tampering. After this preprocessing step, we ended up having $1,243$ images for the *fake* class.

Alsabbagh, A. R., & Al-Kadi, O. S. (2024). Comparative analysis of deep convolutional neural networks for detecting medical image deepfakes. arXiv.

Proposed Enhancements to Address Limitations

We address the limitation of sub-image correlation by incorporating contextual residual learning using adjacent slices and multi-scale feature extraction.

Additionally, we employ a computationally efficient transfer learning-based architecture to reduce model complexity and enable faster training and inference while maintaining high detection performance.

Methodology

1 DATASET ACQUISITION & LABELING

DATASET

Kaggle CT Scan Dataset

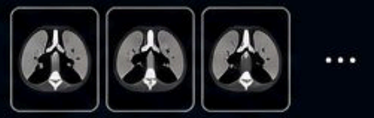
STRUCTURE

LABELING

Class	Label
Tampered	1
Untampered	0

2 CONTEXTUAL FORENSIC PREPROCESSING

ADJACENT SLICE SELECTION



Collect Previous, Current, Next Slices

RESIDUAL EXTRACTION

$$R = I - G(I)$$

Channel 1: Previous

Channel 2: Current

Channel 3: Next

3-CHANNEL FORENSIC INPUT

3 IMAGE PREPROCESSING & AUGMENTATION

PREPROCESSING

- Resize to 224 × 224
- Normalize [0, 1]
- ResNet50V2 Preprocess

AUGMENTATION (TRAINING DATA)

Horizontal Flipping

Small Rotations

Random Zoom

4 TRAIN-VALIDATION SPLIT

DATASET

80% Training Set

20% Validation Set

5 TRANSFER LEARNING & MODEL ARCHITECTURE

BACKBONE

ResNet50V2 (Pretrained)

CLASSIFICATION HEAD

(Custom Layers) → Sigmoid Output

6 TRAINING STRATEGY

INITIAL TRAINING

- Top layers trained
- Early stopping
- Class weights applied

FINE-TUNING

- Additional layers unfrozen
- Lower learning rate
- Adapt to CT tampering artifacts

7 THRESHOLD OPTIMIZATION

Predicted probabilities converted to binary labels using an optimized threshold.

8 EVALUATION

- Accuracy
- Precision
- Recall
- F1-Score
- Classification Report
- Confusion Matrix

Actual \ Predicted	0	1
0	TN	FP
1	FN	TP

Baseline Performance of ResNet50V2

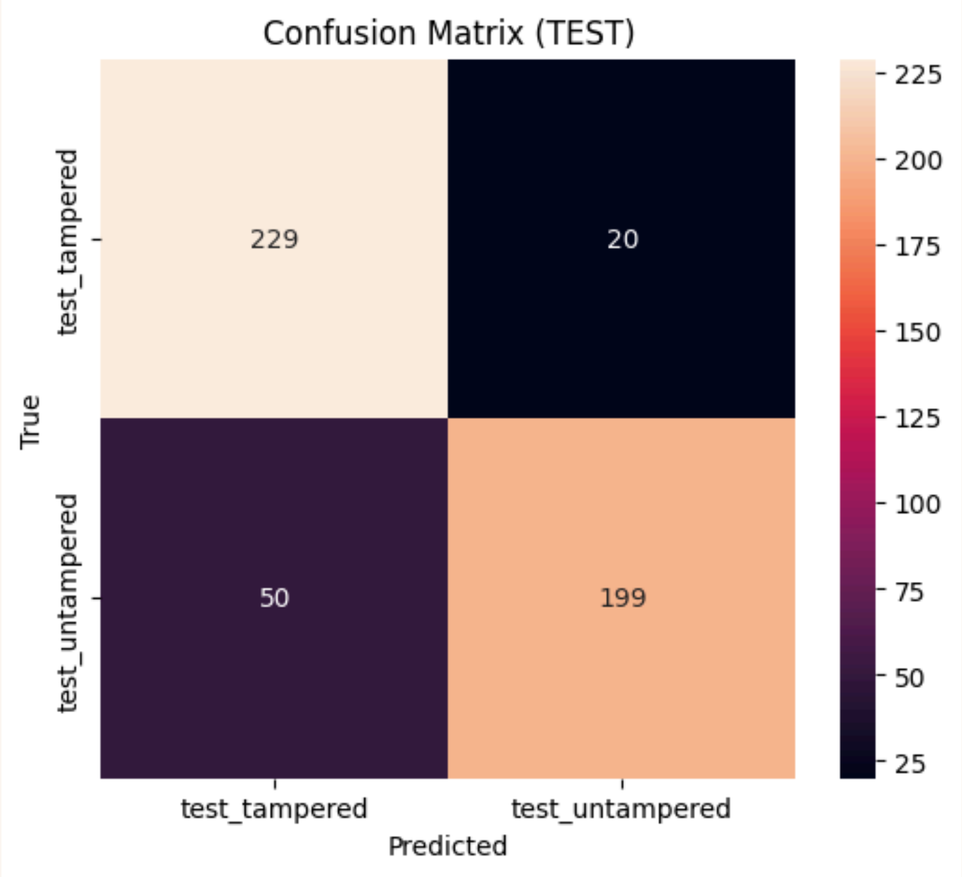
```
Classification Report:

                precision    recall  f1-score   support

  test_tampered       0.82      0.92      0.87       249
test_untampered       0.91      0.80      0.85       249

       accuracy                           0.86       498
      macro avg       0.86      0.86      0.86       498
   weighted avg       0.86      0.86      0.86       498

Test Accuracy: 0.8594377510040161
```



Performance benchmarking

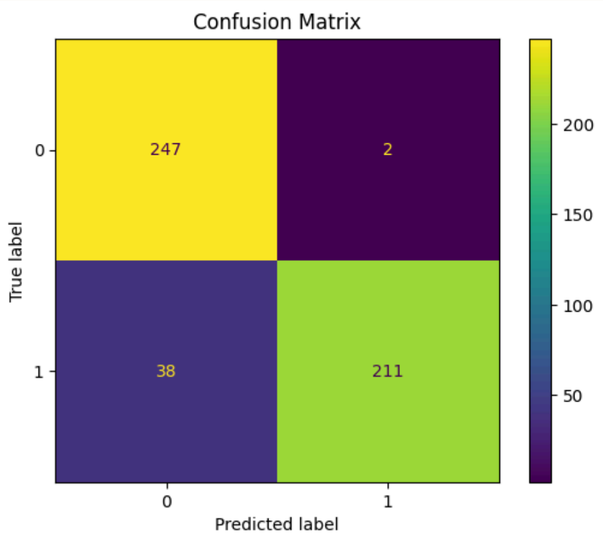
OUR RESULTS

```
Classification Report:

              precision    recall  f1-score   support

         0.0       0.87      0.99      0.93       249
         1.0       0.99      0.85      0.91       249

    accuracy                           0.92       498
   macro avg       0.93      0.92      0.92       498
weighted avg       0.93      0.92      0.92       498
```



Results of Base Model + Inter-slice Context Residual Learning

RESULTS OF THE PAPER

TABLE III
DETAILS PERFORMANCE METRICS

Metric	value
Accuracy	**92.77**
Precision	93.78
Recall	92.77
F1 score	92.60

Sandhya L. S., & Ramanujan, A. (n.d.). Leveraging 3D CNNs for robust detection of GAN-generated medical image forgeries.

Bibliography

Datasets:

1. https://archive.ics.uci.edu/dataset/520/deepfakes+medical+image+tamper+detection
2. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254

Research Papers:

1. https://pmc.ncbi.nlm.nih.gov/articles/PMC10760893/pdf/pone.0290303.pdf
2. https://www.scribd.com/document/902263598/Detection-of-GAN-manipulated-Medical-Images-Through-Deep-Learning-Techniques
3. https://www.usenix.org/system/files/sec19-mirsky_0.pdf
4. https://iweeexplore.ieee.org/abstract/document/11233624?casa_token=AwEt96tbcrQAAAAA:aAtKAqLvgrwiyfIMo8tjHQ4cL94HeFqPJ6CS9oPLUVkfJxAehr-ovZZsfI2Kc2jUjwyxAzew1T9NjQ
5. https://www.sciencedirect.com/science/article/pii/S0010482524013337
6. https://pubmed.ncbi.nlm.nih.gov/33125324/

Thank You