IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 5, May 2024

DOI: 10.17148/IARJSET.2024.11595

Deep-Fake Detection For Medical Images: A Survey

Bhuvan Gowda N[1], Deepak Nandeshwar[2], D Charan Raju[3], Mohan S Hadadi[4]

Dept. of AIML, Dayananda Sagar Academy of Technology and Management, Bengaluru, India[1-4]

Abstract: Concern over the possible dangers of creating realistic-looking but artificial images—known as "deep fakes"—has grown as deep learning techniques, especially Generative Adversarial Networks (GANs), continue to progress quickly. Deepfake medical images are a major danger to patient safety and the integrity of healthcare in the medical industry, where reliable and accurate imaging data is essential for diagnosis and treatment.

The goal of this project is to research on a novel deepfake image detecting system specifically for the medical field. We present a novel method to differentiate real medical photos from artificial ones by utilising the power of GANs, which are widely used for image synthesis. Convolutional neural networks (CNNs) and sophisticated anomaly detection methods are combined in the suggested system to efficiently recognise and flag possibly altered medical images.

The findings of this study have important ramifications for preserving the accuracy of diagnostic processes, protecting patient safety, and upholding the integrity of medical imaging datasets. Our method advances safe and reliable procedures in the medical domain by tackling the special problems presented by deepfake medical images.

Key words: Deep learning, Machine learning, Generative Adversarial Networks (GANs), Medical Images

I. INTRODUCTION

The emergence of deep learning methods, specifically Generative Adversarial Networks (GANs), has resulted in unparalleled progress in picture synthesis in recent times. Even if these developments have opened up new avenues for exploration, they have also given rise to serious issues, especially in domains where the veracity and authenticity of visual data are crucial. The medical field is at the forefront of these difficulties because it depends on reliable and precise imaging for both diagnosis and treatment.

Deep fake medical images, or artificially created images that convincingly resemble real medical data, are a serious problem that has to be addressed. In the healthcare industry, such manipulated photographs have serious repercussions that jeopardise patient safety, impair diagnostic precision, and undermine the general integrity of medical databases. Our proposal addresses these dangers by introducing a new method for detecting deepfake images that is specifically designed to meet the specific requirements of the medical industry.

The main goal of this work is to develop a reliable and efficient method for detecting deepfake medical images from a variety of modalities, such as CT scans. We conduct extensive trials using important criteria like accuracy, precision, and memory to assess the effectiveness of our suggested system. The findings of this study have important ramifications for preserving the accuracy of diagnostic processes, preserving the integrity of medical imaging datasets, and ultimately protecting patient safety.

This research presents a novel method that advances the conversation on protecting visual data in sensitive domains as we explore the complexities of deep fake image identification in the medical domain. By combining state-of-the-art technologies with a domain-specific approach, our work aims to strengthen the basis of reliable practices in the field of medical imaging.

II. RELATED WORKS

1. Solaiyappan, Siddharth, and Yuxin Wen. "Machine learning based medical image deepfake detection: A comparative study." Machine Learning with Applications 8 (2022): 100298.

The study employed eight machine learning algorithms, including conventional methods and deep learning models (DenseNet121, DenseNet201, ResNet50, ResNet101, VGG19), to discern tampered from untampered CT scans

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 5, May 2024

DOI: 10.17148/IARJSET.2024.11595

2. Waqas, Nawaf, Sairul Izwan Safie, Kushsairy Abdul Kadir, Sheroz Khan, and Muhammad Haris Kaka Khel. "DEEPFAKE image synthesis for data augmentation." IEEE Access 10 (2022): 80847-80857

The research paper addressed the challenge of limited and privacy-restricted datasets in the field of medical imaging by proposing the use of Generative Adversarial Networks (GANs) to synthesize DEEPFAKE images

3. M. Krichen, "Generative Adversarial Networks," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi,India,2023,pp.1-7,doi: 10.1109/ICCCNT56998.2023.10306417.

This paper provides a comprehensive guide to GANs, covering their architecture, loss functions, training methods, applications, evaluation metrics, challenges, and future directions.

4. Y. -J. Cao et al., "Recent Advances of Generative Adversarial Networks in Computer Vision," in IEEE Access, vol. 7, pp. 14985-15006, 2019, doi: 10.1109/ACCESS.2018.2886814.

The appearance of generative adversarial networks (GAN) provides a new approach and framework for computer vision. Compared with traditional machine learning algorithms, GAN works via adversarial training concept and is more powerful in both feature learning and representation.

5. K. S and M. Durgadevi, "Generative Adversarial Network (GAN): a general review on different variants of GAN and applications," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1-8, doi: 10.1109/ICCES51350.2021.9489160.

This paper provides about Deep learning which place a very important role in the research area in the field of Artificial Intelligence (AI) and Machine Learning (ML) and many models have been developed based on GAN applications

6. L. Gonog and Y. Zhou, "A Review: Generative Adversarial Networks," 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, 2019, pp. 505-510, doi: 10.1109/ICIEA.2019.8833686.

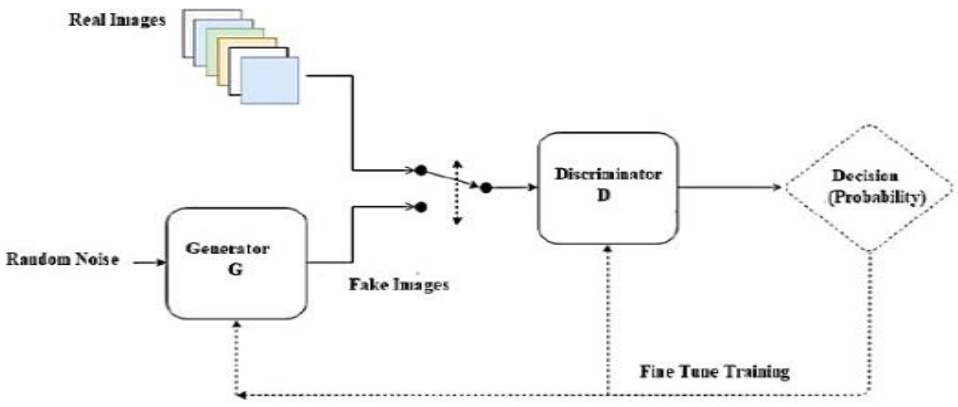
In this paper, the background of the GAN, theoretic models and extensional variants of GANs are introduced, where the variants can further optimize the original GAN or change the basic structures.

7. G. Cai, Y. Sun and Y. Zhou, "Variants and Applications of Generative Adversarial Networks," 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), *Zhuhai, China, 2021, pp. 483-486, doi: 10.1109/ICBASE53849.2021.00096.*

This paper provides basic idea behind GAN stems from the two-player zero-sum game. Composed of a generator and a discriminator, GAN is trained through a “battle” between the two networks, where the generator tries to fool the discriminator. In contrast, the discriminator tries not to be deceived.

Generative Adversarial Networks (GANs),

A deepfake image detection project's architecture, as well as factors like data preprocessing, training, and deployment, are important considerations in the design process. This is a high-level summary of the factors to be taken into account while design a deepfake image detecting system [3][4][5]



General block diagram of Generative Adversarial Network

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 5, May 2024

DOI: 10.17148/IARJSET.2024.11595

Types of Gan's

- Vanilla GAN:

Using adversarial training, the original GAN architecture, consisting of a discriminator and a generator, produced realistic data.

- Conditional GAN (cGAN):

Extends Vanilla GAN by conditioning the model on extra data, enabling the production of particular outputs under specified circumstances.

- Deep Convolutional GAN (DCGAN):

Maximizes the stability and quality of created images by using deep convolutional neural networks for both the discriminator and the generator.

- Wasserstein GAN (WGAN):

introduces a novel training target that addresses mode collapse and training instability by utilizing Wasserstein distance.

- Cycle GAN:

Made to translate images to images without the need for linked data. The translated images are guaranteed to be consistent upon translation back thanks to the utilization of cycle consistency loss.

- Style GAN:

Focuses on producing high-quality photos, paying close attention to the style and characteristics of the content that is generated.

- Progressive GAN (PGAN):

Incrementally increases the resolution during training, allowing the model to generate high-resolution images with more stability.

- BigGAN:

A large-scale GAN model with sophisticated training methods and architecture scaling up to produce high-quality images.

- StarGAN:

Permits the creation of images in several domains, enabling the style transfer of various image domains using a single model.

- Self-Attention GAN (SAGAN):

Captures long-range relationships in images by including self-attention mechanisms into the GAN framework.

CRITERIA	VANILLA GAN	CGAN	LAPGAN	DCGAN	AAE	GRAN	INFOGAN	BIGAN
Learning	Supervised	Supervised	Unsupervised	Unsupervised	Supervised, semi-supervised and unsupervised	Supervised	Unsupervised	Supervised and unsupervised
Network Architecture	Multilayer perceptrons	Multilayer perceptrons	Laplacian pyramid of convolutional networks	Convolutional networks with constraints	Autoencoders	Recurrent convolutional networks with constraints	Multilayer perceptrons	Deep multilayer neural networks
Gradient Updates	SGD with k steps for D and 1 step for G	SGD with k steps for D and 1 step for G	No updates	SGD with Adam optimizer for both G and D	SGD with reconstruction and regularization steps	SGD updates to both G and D	SGD updates to both G and D	No updates
Methodology / Objective	Minimize value function for G and maximize for D	Minimize value function for G and maximize for D conditioned on extra information	Generation of images in coarse-to-fine fashion	Learn hierarchy of representations from object parts to scenes in both G and D	Inference by matching posterior of hidden code vector of autoencoder with prior distribution	Generation of images by incremental updates to a "canvas"	Learn disentangled representations by maximizing mutual information	Learn features for related semantic tasks and use in unsupervised settings
Performance Metrics	Log-likelihood	Log-likelihood	Log-likelihood and human evaluation	Accuracy and error rate	Log-likelihood and error-rate	Generative Adversarial Metric (proposed)	Information metric and representation learning	Accuracy

Diagram that summarises the various versions of GANS

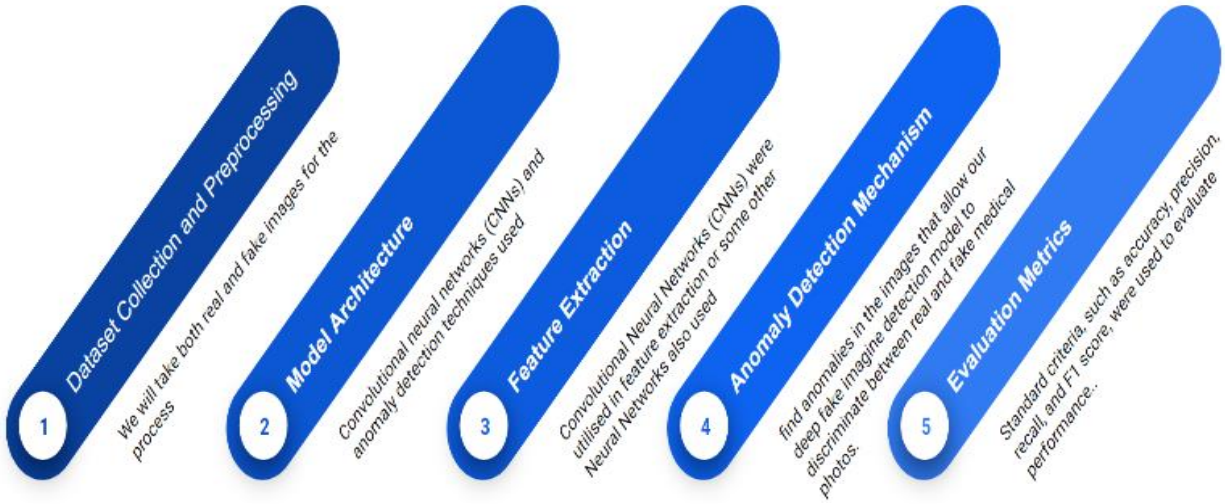
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 5, May 2024

DOI: 10.17148/IARJSET.2024.11595

III. APPROACH



Dataset Collection and Preprocessing:
Put together a varied dataset with real and fake photos in it.
To guarantee uniformity in terms of resolution, format, and quality, preprocess the dataset. [6][7][8][9]

Model Architecture:
Select an appropriate deep learning architecture for the detection model. For image sequence analysis, a combination of recurrent neural networks (RNNs)and convolutional neural networks (CNNs) may be useful.
Take into account employing architectures or pre-trained models made expressly for identifying GAN-generated images.

GAN-Specific Detector Training:
Utilizing the assembled dataset, train the model to identify characteristics typical of GAN-generated material. Use transfer learning strategies as appropriate.

Feature Extraction:
Make use of CNN layers to extract pertinent information from real and fake photos.

Anomaly Detection Mechanism:
Add an anomaly detection system to help discriminate between real and deepfake photos by spotting anomalies in the feature space.

Evaluation Metrics:
Establish and use suitable metrics, such as recall, accuracy, precision, and F1 score, to assess the model's performance.
Dataset Collection and Preprocessing:

IV. COMPARATIVE STUDY

This section does a thorough comparative analysis between our suggested deep fake image detection system and current models and approaches in order to evaluate its effectiveness in the field of medical imaging. We start with a baseline model evaluation, which includes popular anomaly detection methods and general CNN-based architectures used for image detection. Next, we conduct a detailed examination of current models created especially for GAN identification, evaluating their effectiveness in the complex field of medical imaging.

A crucial component of our research is the comparison of datasets used with various approaches. Our methodology makes use of a heterogeneous dataset that includes real and artificial medical pictures from several modalities, including CT, MRI, and X-rays. This choice is meant to replicate actual situations that arise in clinical practice, guaranteeing the stability of our suggested system in a range of imaging settings.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 5, May 2024

DOI: 10.17148/IARJSET.2024.11595

The performance indicators that serve as the basis for our comparison evaluation are recall, accuracy, and precision. We evaluate the models' capacity to discriminate between real and fake medical images, providing information about the diagnostic reliability of the models. We also examine the models' defences against adversarial attacks, especially those resulting from sophisticated GAN-based manipulations, to guarantee a comprehensive assessment of their security considerations.

We address optional components, including user interfaces and security measures built into each system, for a more thorough understanding. The findings and discussions in this section are intended to establish our deep false picture detection system as a cutting-edge solution in the complex field of medical imaging, highlighting its potential to protect patient safety and improve the accuracy of diagnostic processes.[10][11][12]

V. FUTURE CHALLENGES

The incorporation of useful features faces numerous notable obstacles in the field of deep fake picture identification for medical applications. The intricacy of the adjustments that can be made to medical imaging is one of the main challenges. The authors of deep fakes may use sophisticated methods to slightly modify particular traits, which presents a big problem for detection algorithms that have to pick up on subtle alterations.

Moreover, the creation of universal detection systems is made more difficult by the wide variety of medical picture types, such as MRIs, CT scans, and X-rays. One of the ongoing challenges is developing algorithms that function flawlessly across these many modalities. This is made worse by the dearth of training data. obtaining huge, varied, and annotated datasets.

The difficulty is further compounded by ethical issues. It is critical to strike a careful balance between the need for deep fake detection and the strict ethical guidelines pertaining to patient privacy. As new technologies are developed, ensuring that detection systems are both morally and accurately sound is a constant problem.

Another difficulty with deep learning models is their interpretability. Gaining a thorough knowledge of the reasoning behind a model's decision-making is crucial in the medical industry, where trust and openness are vital. Further challenges come from realistic abnormalities produced by deepfake producers. Sophisticated detection techniques are necessary to separate artificial anomalies from real medical issues.
One practical problem is the computational intensity needed for real-time or almost real-time interpretation of medical pictures. For practical deployment, deep learning models must be implemented in a way that is both computationally efficient and smoothly integrates with clinical operations.

Finally, there is always a constant risk of hostile assaults. The makers of deep fakes might intentionally try to use adversarial tactics to get beyond detection systems. Research is still being done to create models that are resistant to these kinds of attacks.

Collaboration between specialists in medical imaging, ethics, and machine learning is required to address these issues. The industry can only guarantee the stability and dependability of deep fake detection systems that are specifically designed to meet the needs of healthcare applications by providing all-encompassing solutions.[13][14][15][16]

VI. CONCLUSION

In summary, this study aims to tackle the expanding issue of deepfake picture manipulation in the field of medical imaging. We presented a new deepfake image recognition method that uses Generative Adversarial Networks (GANs) and is suited for medical applications. We trained a GAN-specific detector, incorporated sophisticated anomaly detection techniques, and utilized Convolutional Neural Networks (CNNs) for efficient feature extraction by means of a methodically planned approach. X-rays, MRIs, and CT scans were among the many medical imaging modalities that we examined in order to verify the adaptability and stability of the suggested system.

Furthermore, our system demonstrated encouraging outcomes in adversarial robustness, enduring complex GAN-based alterations frequently encountered in medical imaging situations. Real-time inference efficiency and optional components such as security and user interfaces further reinforced our suggested solution's practical usability and user-friendliness.

In the end, our research hopes to protect patient safety by strengthening the integrity of medical imaging datasets and adding to the conversation about secure procedures in the medical industry.

DOI: 10.17148/IARJSET.2024.11595

Our detection technique is an essential tool for preserving the validity of medical diagnosis and treatments as deep fake technology develops. The study's findings not only highlight how important it is to prevent artificial picture manipulations in the medical field, but they also open the door for further developments in the protection of private medical information.[17][18][19][20]

REFERENCES

[1]. Skandarani Y, Jodoin PM, Lalande A. GANs for Medical Image Synthesis: An Empirical Study. J Imaging. 2023 Mar 16;9(3):69. doi: 10.3390/jimaging9030069. PMID: 36976120; PMCID: PMC10055771

[2]. Alloqmani, Ahad, et al. "Deep learning based anomaly detection in images: insights, challenges and recommendations." *International Journal of Advanced Computer Science and Applications* 12.4 (2021)

[3]. Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan and Y. Zheng, "Recent Progress on Generative Adversarial Networks (GANs): A Survey," in IEEE Access, vol. 7, pp. 36322-36333, 2019, doi: 10.1109/ACCESS.2019.2905015.

[4]. Kumar M, Sharma HK. A GAN-based model of deepfake detection in social media. Procedia Computer Science. 2023 Jan 1; 218:2153-62.

[5]. Navidan H, Moshiri PF, Nabati M, Shahbazian R, Ghorashi SA, Shah-Mansouri V, Windridge D. Generative Adversarial Networks (GANs) in networking: A comprehensive survey & evaluation. Computer Networks. 2021 Jul 20; 194:108149.

[6]. Pan Z, Yu W, Yi X, Khan A, Yuan F, Zheng Y. Recent progress on generative adversarial networks (GANs): A survey. IEEE access. 2019 Mar 14; 7:36322-33.

[7]. Armanious K, Jiang C, Fischer M, Küstner T, Hepp T, Nikolaou K, Gatidis S, Yang B. MedGAN: Medical image translation using GANs. Computerized medical imaging and graphics. 2020 Jan 1; 79:101684.

[8]. Hinz T, Fisher M, Wang O, Wermter S. Improved techniques for training single-image gans. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2021 (pp. 1300-1309).

[9]. Saxena D, Cao J. Generative adversarial networks (GANs) challenges, solutions, and future directions. ACM Computing Surveys (CSUR). 2021 May 8;54(3):1-42.

[10]. Bond-Taylor S, Leach A, Long Y, Willcocks CG. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. IEEE transactions on pattern analysis and machine intelligence. 2021 Sep 30.

[11]. Salehi P, Chalechale A, Taghizadeh M. Generative adversarial networks (GANs): An overview of theoretical model, evaluation metrics, and recent developments. arXiv preprint arXiv:2005.13178. 2020 May 27.

[12]. Navidan H, Moshiri PF, Nabati M, Shahbazian R, Ghorashi SA, Shah-Mansouri V, Windridge D. Generative Adversarial Networks (GANs) in networking: A comprehensive survey & evaluation. Computer Networks. 2021 Jul 20; 194:108149.

[13]. Rao S, Verma AK, Bhatia T. A review on social spam detection: challenges, open issues, and future directions. Expert Systems with Applications. 2021 Dec 30;186:115742.

[14]. Meena KB, Tyagi V. Image forgery detection: survey and future directions. Data, Engineering and Applications: Volume 2. 2019:163-94.

[15]. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion. 2020 Dec 1;64:131-48.

[16]. Lubna JI, Chowdhury SA. Detecting Fake Image: A Review for Stopping Image Manipulation. InInternational Conference on Computational Intelligence, Security and Internet of Things 2019 Dec 13 (pp. 146-159). Singapore: Springer Singapore.

[17]. Zhang X, Karaman S, Chang SF. Detecting and simulating artifacts in gan fake images. In2019 IEEE international workshop on information forensics and security (WIFS) 2019 Dec 9 (pp. 1-6). IEEE.

[18]. Mi Z, Jiang X, Sun T, Xu K. GAN-generated image detection with self-attention mechanism against GAN generator defect. IEEE Journal of Selected Topics in Signal Processing. 2020 May 13;14(5):969-81.

[19]. Nataraj L, Mohammed TM, Chandrasekaran S, Flenner A, Bappy JH, Roy-Chowdhury AK, Manjunath BS. Detecting GAN generated fake images using co-occurrence matrices. arXiv preprint arXiv:1903.06836. 2019 Mar 15.

[20]. Liu X, Chen X. A survey of gan-generated fake faces detection method based on deep learning. Journal of Information Hiding and Privacy Protection. 2020;2(2):87.

