Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, No.2, February : 2025

MEDICAL DEEPFAKE IMAGE DETECTION USING DEEP LEARNING

Mrs. Pratibha Patil, Professor, Dept.Of Information Technology, International Institute of Information Technology,pune, Savitribai Phule Pune University.

Mr. Digvijay Patil, Professor, Savitribai Phule Pune University.
Mr. Jayesh Parate, Student, Dept.Of Information Technology, International Institute of Information Technology,pune, Savitribai Phule Pune University.
Mr. Kunal Kudande, Student, Dept.Of Information Technology, International Institute of Information Technology,pune, Savitribai Phule Pune University.
Mr. Atharva Pisal, Student, Dept.Of Information Technology, International Institute of Information Technology,pune, Savitribai Phule Pune University.
Mr.Shivam Tavaskar, Student, Dept.Of Information Technology , International Institute of Information Technology,pune, Savitribai Phule Pune University

ABSTRACT:
The rapid advancements in deep learning (DL) and artificial intelligence (AI) have raised serious security and privacy concerns, particularly with deepfakes—manipulated data that can tamper medical images. The inability to detect these medical deepfakes poses risks, including compromised hospital assets, political sabotage, insurance fraud and potential loss of life. To address this, we introduce a deep learning method, designed to detect deepfake lung CT scans and Knee Osteoarthritis X-ray. The model is based on a CNN model which is a modified EfficientNet framework. We have combined two datasets CT-GAN dataset and Knee Osteoarthritis X-ray dataset in order to train the model. Data pre-processing and augmentation methods are applied for data standardization and variation. The Model aims to achieve high detection accuracy and reduce the risk of deepfakes
Keywords: CT GAN(Generative Adversarial Network), Deep Learning, Efficient Net , VGG, X-Ray , Lung CT Scan, Convolutional Neural Network(CNN).

INTRODUCTION

The detection and diagnosis of spine disorders, such as disc degeneration and herniation, is an critical due to their impact on the quality of life, often leading to chronic lower back pain [1]. With Artificially generated images known as "medical deepfakes" are created by modifying or creating realistic medical data, such as X-rays, MRIs, and CT scans, using advanced AI techniques, especially deep learning. These techniques produce highly realistic artificial images. Deepfake photos started as a result of the 2014 development of Generative Adversarial Networks (GANs). With the advancement of GANs, deepfakes spread throughout a variety of areas, including healthcare, entertainment, and politics. Security risks could arise from the outdated software used by many CT machines in the medical profession.

The CT-GAN model, which can add or remove tumours from lung CT scans, is a prominent illustration of this. Radiologists were unable to identify modified photos in blind examinations; false-positive rates for extracted tumours were 94% and for added tumours were up to 99%. Detection rates for excised tumours were 87% and for additional tumours were only 60%, even in open examinations. These concerning findings highlight how susceptible modern medical practices are to deepfake attacks, and automated detection tools are desperately needed to protect patient safety.

Deepfake creation with GANs has been easier thanks to developments in AI and machine learning. To protect patients and preserve faith in healthcare systems, it is essential to identify deepfake photos in medical data.

Deepfake detection is one of the many applications of Convolutional Neural Networks (CNNs) in medical picture analysis. CNNs are frequently chosen for detecting fraudulent information because of

their capacity to automatically learn filters and features with little preprocessing. CNNs, for instance, can distinguish between benign and malignant growths in lung CT scans according to nodule size in order to diagnose lung cancer.
The suggested deep learning architecture is made to identify modifications in knee osteoarthritis X-rays and lung CT scans in order to overcome these difficulties. The model adds dense layers to the EfficientNet model in order to extract important information from the photos and categorise them with a high degree.

LITERATURE:
This literature survey explores various studies on the development and implementation of Medical Deepfake Detection, focusing on their impact, findings, research gaps.
A New Approach for Effective Medical Deepfake Detection in Medical Images (2024).Using various iterations of the YOLO object identification model, Karaköse et al. presented a method for identifying deepfake changes in CT scans and X-ray pictures. They tested a number of models, including YoloV3, YoloV5nu, YoloV5su, and YoloV8, and showed that YoloV5su performed better than the others, attaining a recall rate of 0.997 and outperforming other models, such as YoloV8x, by 60%. Their research highlights the necessity of real-time detection systems to safeguard patient safety and hospital infrastructure against the growing danger of deepfakes in medical imaging [1].
The CT-GAN framework, which was presented by Mirsky et al., is another important addition to the discipline. This work focusses on the injection and removal of medical illnesses like lung cancer in 3D CT scans, highlighting the susceptibility of medical systems to deepfake attacks. The scientists showed how CT-GAN was able to fool both human radiologists and the most advanced artificial intelligence technologies, with radiologists 60–99% of the time failing to recognise modified images. According to this study, there is a concerning chance that adversaries may take use of deepfake technology in the medical field, which could have detrimental effects including fraud and incorrect diagnosis [10].
The researchers examined deepfake detection methods for medical photos in a survey by Gowda et al., highlighting the urgent need for specialised models catered to medical data. The study covered the efficiency of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) in identifying irregularities in doctored medical images. The authors emphasised the critical need for precise and effective detection systems, pointing out the serious risks that deepfake medical images pose to patient safety and the integrity of diagnostic procedures. (IARJSET.2024.11595).[12]
A deep learning-based system called MedNet was presented by Albahli et al. with the purpose of identifying deepfakes in lung CT scans. The model, which is based on the EfficientNetV2-B4 architecture, improves the network's feature extraction capabilities by adding more dense layers and a spatial-channel attention mechanism. When MedNet was evaluated on the CT-GAN dataset, it detected phoney lung CT images with an accuracy of 85.49%, suggesting that attention processes can enhance detection performance. For accurate classification, our work emphasises the significance of concentrating on changed regions in medical images and the necessity of ongoing developments in deepfake detection methods. [3]
In the paper Machine learning based medical image deepfake detection: A comparative study, eight machine learning algorithms are evaluated, comprising five deep learning models (DenseNet121, DenseNet201, ResNet50, ResNet101, and VGG19) and three traditional approaches (Support Vector Machines, Random Forests, and Decision Trees). On raw photos, they discovered that traditional machine learning models performed better than deep learning ones, with Random Forests and SVM obtaining higher accuracy because of the complexities of deep learning models and their tiny dataset size.[2]
Jekyll: Using Deep Generative Models to Attack Medical Image Diagnostics.This paper presents "Jekyll," a GAN-based tool that fools diagnostic algorithms and human radiologists by creating fake medical images with attacker-selected disease conditions. Jekyll creates fakes by using unpaired data

from photos of both healthy and sick patients. This preserves the patient's identity while adding characteristics of the condition, making it difficult to detect using traditional techniques. The study analyses defensive systems and highlights the significance of these attacks, especially in financially driven fraud situations. It finds that generative approaches can be modified to bypass trained detection algorithms.[8]

Arshed et al.'s research work "A Deep Learning Model for Detecting Fake Medical Images to Mitigate Financial Insurance Fraud" makes a number of significant contributions. Stable Diffusion and Deep Learning for the Identification of False Images: In the context of insurance fraud, the study investigates CNNs and patch-based neural networks for identifying fake medical imaging. High-quality synthetic skin cancer images are produced using stable diffusion technology, and the deep learning models are trained and tested by comparing them to real photos. User research that evaluates human participants' capacity to discriminate between authentic and fraudulent photos is also included in the publication. This comparison showed that deep learning models outperformed human perception, which only managed 68% accuracy.[13]

In order to identify tampering in medical images, especially cancer imagery, the publication, "Intelligent Deep Detection Method for Malicious Tampering of Cancer Imagery," proposes a deep neural network (DNN)-based method. A collection of 3D CT lung images with both real and fictitious tumour insertions or removals was employed in the investigation. Three hidden layers and twenty-four input features made up the framework of the DNN model. The system's detection accuracy rate of 93.19% was far greater than that of earlier CNN-based models (81% accuracy), proving that DNN can successfully increase detection rates while lowering error and false alarm rates. The DNN's performance was compared to other methods, proving superior due to enhanced detection and minimized error. [7]

METHODOLOGY:

This suggested solution provides a thorough framework for efficiently identifying faked medical photos in response to the growing use of deepfake technology in the field of medical imaging. To ensure a strong dataset of authentic and deepfake medical images, including knee osteoarthritis X-rays and lung CT scans, the method starts with careful data collection and augmentation. The EfficientNet/VGG architecture is used to prepare the data for analysis by methodical preparation, which includes picture conversion and normalisation. The effectiveness and great accuracy of this model in processing high-resolution images led to its selection. The architecture is made to concentrate on areas of images that may have been manipulated by utilising advanced techniques like attention processes. For accurate classification, the classification layer uses binary cross-entropy loss. A variety of performance indicators, such as accuracy, precision, and F1-score, are used to train and verify the model. The technology is intended for implementation in clinical settings following thorough testing on unseen data, allowing for the real-time detection of deepfakes to protect patient safety and preserve the accuracy of medical diagnostics. In addition to addressing the technical difficulties of deepfake detection, this multipronged strategy seeks to smoothly incorporate into current medical imaging workflows, guaranteeing usefulness in real-world situations.

Data collection And Augmentation:

Datasets: Gather datasets containing both authentic and deepfake medical images, like X-rays or CT scans.

Image Standardization: Resize all images to match Efficient Net’s input dimensions (e.g., 380x380 for EfficientNet).

Augmentation: Apply techniques like rotation, contrast adjustment, brightness modulation, and flipping to enhance dataset variety and improve generalization.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, No.2, February : 2025

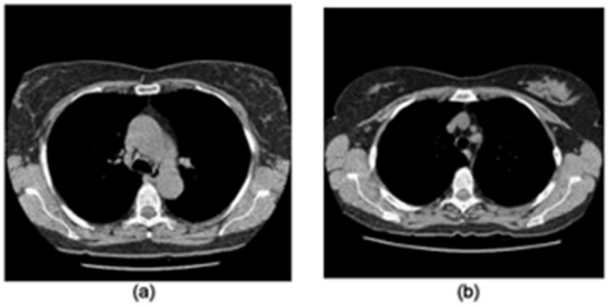

Figure 1 : a)Real Image b) Fake Image

used to enhance local contrast and accentuate anatomical structures, improving the ability of models to detect lesions[2].

Data augmentation was another crucial preprocessing step, Some of the augmentation methods employed in this study were: horizontal flip, random rotation of images, brightness scaling, and a Gaussian noise factor. These transformations helped the models generalize better.

The studies also indicated that preprocessing techniques such as cropping, resizing, and augmentation greatly enhanced the general performance of CNN models.

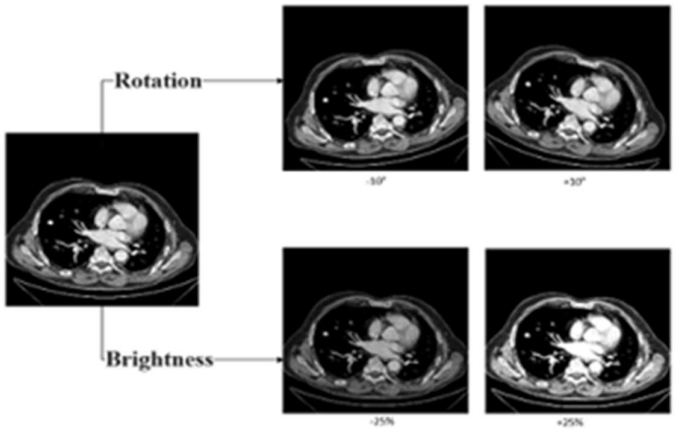

Figure 2 : Data Augmentation

PREPROCESSING:

Image Conversion: Convert DICOM or other specialized medical image formats to JPEG or PNG, using a standardized resolution. We have to Combine Lung CT Scan and Knee Osteoarthritis Dataset for better results. Normalization: Normalize pixel values (e.g., to a 0-1 range) to match the input requirements of EfficientNet.

EfficientNet Model:

Use EfficientNet because of its great accuracy for high-resolution photos and effective compound scaling. Adjust depth, width, and resolution based on dataset size and available computational resources.

Model Architecture and Layers:

1.Feature Extraction Layers: The EfficientNet architecture can be used to extract intricate features while maintaining computing efficiency in feature extraction layers. Attention Mechanisms: To highlight important feature areas and concentrate on areas that are probably altered, incorporate a spatial-channel attention mechanism.

2.Fully Connected Layers: Use fully connected layers to consolidate extracted features for classification.

3.Classification Layer

Add a SoftMax or sigmoid layer to classify each image as either Real or Deepfake.

4. Training and Validation

Training Process: Train the model on the training set with validation using appropriate metrics.

Performance Metrics: Track accuracy, precision, recall, and F1-score to evaluate model performance.

5. Evaluation and Testing

Evaluation Dataset: Test the model on unseen medical images to assess its ability to generalize.

Confusion Matrix: Improve model reliability and comprehend misclassifications by using confusion matrix analysis.

6. Deployment and Real-Time Testing

Real-Time Application: Implement the model in a research or clinical setting, potentially combining it with a medical imaging system to detect deepfakes in real time.

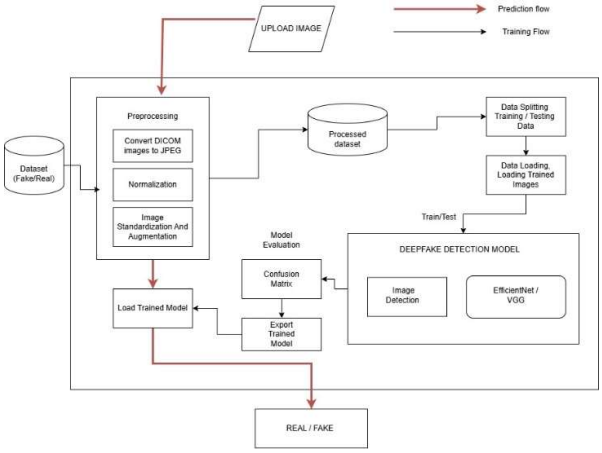

Figure 3: Architecture Diagram

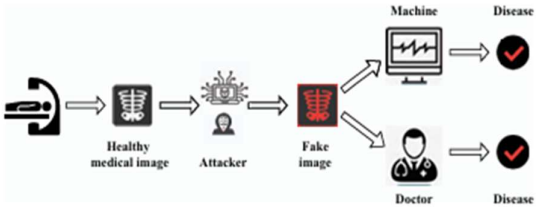

Figure 4: Medical DeepFake Injection

Infiltration into Medical Systems:

Distributing Deep Fakes: After being created, deepfake medical photos can be shared via email, file-sharing websites, or direct uploads to imaging systems.

Insider Threats: Sometimes, people who have access to medical systems (such as technicians or healthcare professionals) would purposefully insert deep fakes into the process, either maliciously or to skew the results.

Social Engineering:

Phishing Attacks: Cybercriminals may employ social engineering techniques, such as phishing attacks, to persuade medical personnel to examine or believe distorted images. This could result in the accidental incorporation of deepfakes into medical decision-making procedures.

Convolutional Neural Networks (CNNs):

In order to differentiate between authentic and fraudulent medical photos, a CNN model for deepfake image detection uses its capacity to recognise intricate details in images. Data collection is the first step in the process, which involves finding real medical images from trustworthy databases and

creating altered (deep fake) copies utilising sophisticated methods like GANs or publicly accessible datasets containing phoney medical imagery. Preprocessing the data is crucial; it entails standardising the input shape by downsizing photos to a uniform size, like 224x224 pixels. In order to stabilise and speed up model training, image normalisation is also used to scale pixel values, usually between 0 and 1. Rotations, flips, and brightness modifications are examples of data augmentation techniques that can be used to expose the model to a range of styles. Multiple convolutional and pooling layers are used in the CNN architecture to capture complex characteristics, and fully connected layers are then used for classification. Depending on whether the task involves binary or multi-class classification, the final output layer usually employs either a softmax or sigmoid activation function. A dataset that has been divided into training, validation, and testing sets is used to train the model. An appropriate optimiser, such as Adam, is then used to optimise a loss function, such as binary cross-entropy. The model can successfully identify and categorise medical deepfake images after it has been taught, helping to preserve the reliability and integrity of medical imaging procedures.

EFFICIENTNET:

EfficientNet is a deep learning model type for image classification that scales its architecture in a novel way to improve efficiency and performance. This is how it operates:

Compound Scaling: EfficientNet increases the network's size by balancing three parameters: resolution (image size), width (channels in each layer), and depth (layers). This improves the model's accuracy without requiring a significant amount of processing resources.

Convolutional Layers: These layers identify fundamental aspects of an image, such as textures and edges. As the model becomes more sophisticated, it can identify increasingly intricate patterns that are necessary to differentiate between authentic and fraudulent medical photos.

Squeeze-and-Excitation (SE) Blocks: By varying the amount of attention paid to certain elements, these unique blocks assist the model concentrate on the most crucial areas of the image.

Activation and Normalisation: The model employs batch normalisation to maintain training stability and the Swish activation function, which facilitates a smooth gradient flow during training.

Pooling and Fully Connected Layers: By reducing the quantity of the data, pooling layers improve the efficiency of the model. The last fully connected layers employ a sigmoid function for binary outputs and deal with the actual classification problem (such as a real or fake image).

Training: To enhance generalisation, the model is trained on a dataset of pre-processed and occasionally enhanced images. It updates the model weights using optimisers like Adam and learns by minimising a loss function like binary cross-entropy.

Inference: Following training, the model is able to effectively classify fresh photos and determine if they are authentic or fraudulent.

CONCLUSION :

In conclusion, patient safety and the integrity of medical imaging are seriously threatened by the development of deepfake technology. Using cutting-edge methods like EfficientNet for feature extraction and classification, this paper has presented a thorough framework for identifying medical deep fakes. The suggested solution highlights the significance of ongoing validation and real-time application in healthcare settings as deep fakes continue to develop. Our methodology mphasizes the serious ethical implications of deepfakes in healthcare in addition to attempting to address the technological issues of detection. We can improve the safety and reliability of medical diagnostics by implementing this detection system in clinical settings, thereby safeguarding patient welfare and maintaining the reputation of medical practitioners. To stay up with developments in deepfake production and maintain the efficiency of detection techniques in an increasingly digital healthcare environment, future research should investigate adaptive learning processes and larger datasets.

REFERENCES

[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

[2] Solaiyappan, S., & Wen, Y. (2022). Machine learning based medical image deepfake detection: A comparative study.MachineLearning With Applications,8,100298.https://doi.org/10.1016/j.mlwa.2022.100298

[3] Albahli, S., & Nawaz, M. (2023). MedNet: Medical deepfakes detection using an improved deep learning approach. Multimedia Tools and Applications, 83(16), 48357 48375. https://doi.org/10.1007/s11042-023-17562-5Zheng, G., Pei, S., Xie, Y., & Xu, D. (2020). Automated detection of lumbar disc degeneration based on MRI images: a comprehensive review. IEEE Access, 8, 101-112.

[4] Y. Aslam and N. Santhi, ''A review of deep learning approaches for image analysis,'' in Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT), Nov. 2019, pp. 709–714, doi: 10.1109/ICSSIT46314.2019.8987922.

[5] P.Chen. (2018). Knee Osteoarthritis Severity Grading Dataset. Mendeley Data, V1. [Online]. Available: https://www.10.17632/56rmx5bjcr.1

[6] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, ''GAN-based synthetic brain MR image generation,'' in Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), Apr. 2018, pp. 734–738.

[7] . K. M. A. Alheeti, A. Alzahrani, N. Khoshnaw, and D. Al-Dosary, ''Intelligent deep detection method for malicious tampering of cancer imagery,'' in Proc. 7th Int. Conf. Data Sci. Mach. Learn. Appl. (CDMA), Mar. 2022, pp. 25–28.

[8] Mangaokar, N., Pu, J., Bhattacharya, P., Reddy, C. K., & Viswanath, B. (2020). Jekyll: Attacking Medical Image Diagnostics using Deep Generative Models. *Jekyll: Attacking Medical Image Diagnostics Using Deep Generative Models*. https://doi.org/10.1109/eurosp48549.2020.00017

[9] Jaiswal A, Gianchandani N, Singh D, Kumar V, Kaur M (2021) Classification of the COVID 19 infected patients using DenseNet201 based deep transfer learning. J Biomol Struct Dynamics 39(15):5682–5689

[10] Mirsky Y, Mahler T, Shelef I, Elovici Y (2019) {CT-GAN}: Malicious Tampering of 3D Medical Imagery using Deep Learning. In: 28th USENIX Security Symposium (USENIX Security 19), pp. 461–478 12. Nawaz M, Javed A, Irtaza A (2022) ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. Visual Comput 1–22

[11] Deep-Fake detection for medical Images: a survey. (2024). *International Advanced Research Journal in Science, Engineering and Technology*, *11*(5), 630. https://doi.org/10.17148/IARJSET.2024.11595

[12] Gowda, B., Nandeshwar, D., Raju, D. C., & Haddi, M. S. (2024). Deep-Fake Detection For Medical Images: A Survey. *International Advanced Research Journal in Science, Engineering and Technology*, *11*(5), 630. https://doi.org/10.17148/IARJSET.2024.11595

[13] Arshed, M. A., Mumtaz, S., Gherghina, Ș. C., Urooj, N., Ahmed, S., & Dewi, C. (2024). A deep learning model for detecting fake medical images to mitigate financial insurance fraud. *Computation*, *12*(9), 173. https://doi.org/10.3390/computation12090173