

A Deep Learning-Based Approach for Medical Deepfake Detection Using YOLO Architectures

Tambe Vaishnavi Ramhari¹, Dr. Hemantkumar B. Jadhav, Prof. Jagruti R. Mahajan⁴, Prof. Pragati B. Chandane, Dr. Pradeep M. Patil⁵

¹PG Scholar, Adsul's Technical Campus, Ahilyanagar

²Professor, Adsul's Technical Campus, Ahilyanagar

³Professor, Adsul's Technical Campus, Ahilyanagar

⁴Assistant Professor, Adsul's Technical Campus, Ahilyanagar

⁵Principal, Adsul's Technical Campus, Ahilyanagar

Abstract: The rapid advancement of Artificial Intelligence and Deep Learning technologies has significantly increased the creation of deepfake content, including manipulated medical images. Altered medical scans and X-ray images can mislead healthcare professionals, resulting in incorrect diagnoses, delayed treatments, financial losses, and serious risks to patient safety. Therefore, the identification of fake medical images has become an important challenge in the healthcare sector. This research proposes an efficient deep learning-based approach for detecting manipulated medical images using multiple YOLO object detection models. Two separate datasets containing Knee Osteoarthritis X-ray images and Lung CT scan images were prepared for experimentation. Data preprocessing and augmentation techniques were applied to improve image quality, maintain dataset consistency, and increase model robustness. Each image in the datasets was categorized as either real or fake for supervised learning.

• **Keywords:** Medical deepfake detection, Deep Learning, YOLOv5, Lung CT Scan, Knee Osteoarthritis X-ray

I. INTRODUCTION

Deepfakes, created through advanced machine learning techniques, have emerged as a powerful and increasingly popular way to generate highly realistic fake videos and images, often by swapping one person's face with another. While the technology demonstrates impressive capabilities, it has also raised serious concerns about misuse, as malicious actors may exploit deepfakes to spread misinformation, manipulate opinions, or violate privacy. This growing threat has prompted researchers and technologists to develop deep learning-based detection systems aimed at safeguarding digital media integrity.

II. LITERATURE REVIEW

Deepfake detection has become one of the most important research areas in Artificial Intelligence and medical image security. Researchers have proposed several deep learning-based methods for identifying manipulated medical images generated using GAN architectures.

One of the earliest studies focused on manipulated retinal images using the Ocular Disease Intelligent Recognition dataset. Researchers evaluated the capability of ophthalmologists and general practitioners in identifying fake medical images. The results demonstrated that human experts faced significant difficulty distinguishing fake images from real images. To improve performance, a U-Net-based deep learning model was implemented, achieving significantly better detection accuracy.

Another important study used Sparse Convolutional Neural Networks (Sparse CNNs) to identify manipulated images generated using CycleGAN architectures. The researchers created datasets containing real, manipulated, and hybrid medical images. The CNN-based system successfully detected hidden visual inconsistencies and synthetic patterns introduced during manipulation.

Lung CT scan deepfake detection has also attracted major attention due to the risk of fake cancer insertion and removal. Researchers proposed EfficientNet-V2-based architectures integrated with attention mechanisms for detecting manipulated CT scans. The proposed models focused on suspicious image regions and achieved high classification performance.

III. DATASET DESCRIPTION

A. Dataset Overview

In this study, two different medical datasets were used for detecting manipulated medical images:

1. Knee Osteoarthritis X-ray Dataset

2. Lung CT Scan Dataset

The datasets contain both real and fake medical images generated using deepfake techniques.

B. Knee Osteoarthritis X-ray Dataset

The Osteoarthritis dataset contains real and GAN-generated fake X-ray images. Real images were collected from publicly available Osteoarthritis severity grading datasets, while manipulated images were generated using GAN-based architectures.

Dataset Characteristics

Category	Number of Images
Real Images	1600
Fake Images	1600
Total Images	3200

Sample Dataset Representation

The dataset includes different Osteoarthritis severity levels. All images were labeled into two classes:

- Real
- Fake

C. Lung CT Scan Dataset

The Lung CT scan dataset contains manipulated and genuine CT images collected from CT-GAN-based datasets.

Dataset Characteristics

Category	Number of Images
Real CT Images	700
Fake CT Images	700
Total Images	1400

The manipulated images contain inserted or removed pulmonary nodules created using deepfake methods.

D. Data Augmentation

To improve model performance and increase dataset variation, multiple augmentation techniques were

applied:

- Rotation
- Brightness Adjustment
- Contrast Enhancement
- Image Flipping
- Orientation Correction

E. Data Preprocessing

Medical images collected from different sources varied in size and format. Therefore, preprocessing was performed for standardization.

Preprocessing Steps

1. Image Resizing
2. Noise Removal
3. Format Conversion
4. Image Normalization
5. Bounding Box Labeling

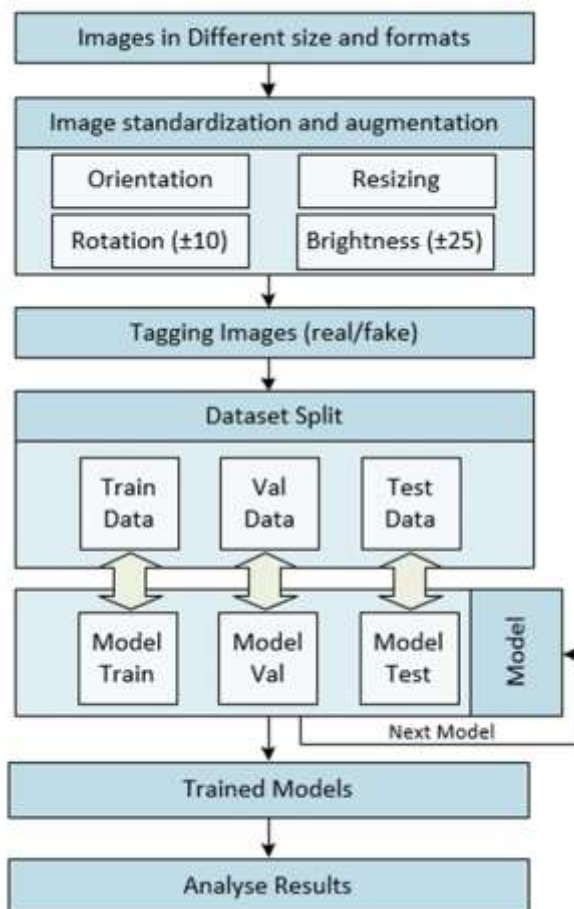
Image Dimensions

Dataset	Resized Dimension
X-ray Images	210 × 210
CT Scan Images	416 × 416

DICOM images were converted into JPEG format before training.

IV. PROPOSED METHODOLOGY

The proposed system uses YOLO-based deep learning architectures to identify manipulated medical images. The given diagram represents the complete workflow of the medical deepfake detection system, starting from image collection to final result analysis. The architecture illustrates how medical images are processed, augmented, classified, trained, and evaluated using deep learning models.



Step 1: Dataset Collection

Medical images were collected from publicly available datasets containing Knee Osteoarthritis X-rays and Lung CT scans.

Step 2: Data Preprocessing

Preprocessing operations included:

- Image resizing
- Image normalization
- Noise reduction
- DICOM conversion
- Bounding box annotation

Step 3: Data Augmentation

Augmentation improved dataset diversity and model robustness.

Applied techniques:

- Rotation
- Brightness adjustment
- Contrast enhancement
- Horizontal flipping

Step 4: YOLO Model Training

The following YOLO architectures were trained:

- YOLOv3
- YOLOv5nu
- YOLOv5su
- YOLOv8n
- YOLOv8s
- YOLOv8m
- YOLOv8l
- YOLOv8x

The models learned hidden manipulation patterns and image inconsistencies.

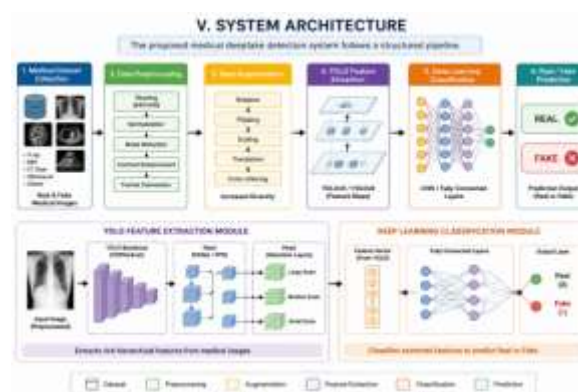
Step 5: Prediction and Classification

The trained models classified medical images into:

- Real Image
- Fake Image

V. SYSTEM ARCHITECTURE

The proposed medical deepfake detection system is theoretically designed as a multi-stage deep learning framework that identifies manipulated medical images using image preprocessing, feature extraction, and classification techniques. The architecture combines computer vision algorithms with deep neural networks to detect forged or artificially generated medical images with high accuracy and computational efficiency.



YOLO Architecture

YOLO is a CNN-based object detection algorithm that predicts bounding boxes and class probabilities simultaneously.

Advantages of YOLO

- Fast execution speed
- Real-time detection
- High accuracy
- Low computational complexity
- Efficient feature extraction

VI. RESULTS AND PERFORMANCE EVALUATION

6.1 Experimental Setup

The proposed deepfake detection system was implemented using the Python programming language with the support of GPU acceleration to achieve faster training and inference. The model training and evaluation were carried out in a high-performance computing environment to efficiently process large image datasets and optimize deep learning computations. The system utilized advanced object detection architectures from the YOLO family for identifying manipulated facial regions in images and video frames.

The training process involved preprocessing the dataset, resizing input images, augmentation, feature extraction, and model optimization. GPU-based parallel computation significantly reduced the

overall training time and improved model convergence.

Hardware Configuration

Component	Specification
Processor	AMD Ryzen 9 7900X
RAM	64 GB
GPU	NVIDIA RTX 4090
Operating System	Windows 11
Programming Language	Python 3.x
Deep Learning Framework	PyTorch
IDE/Environment	Jupyter Notebook / VS Code

The NVIDIA RTX 4090 GPU provided efficient CUDA-based acceleration, enabling faster training of deep learning models with high-resolution image datasets.

6.2 Hyperparameter Configuration

Hyperparameters play an important role in determining the learning capability and overall performance of the deep learning model. Several experiments were conducted to identify the optimal parameter values that provide better accuracy and stable convergence.

Hyperparameters Used

Parameter	Value
Epochs	50
Batch Size	16
Learning Rate	0.001
Optimizer	Adam
Image Size	640 640
Loss Function	Cross-Entropy Loss

Parameter	Value
Activation Function	ReLU
Confidence Threshold	0.25

The Adam optimizer was selected because of its adaptive learning capability and faster convergence properties. A learning rate of 0.001 ensured stable training without overshooting the minima.

6.3 YOLO Model Performance Comparison

Different YOLO architectures were evaluated to determine the most effective model for deepfake detection. The comparison was performed using recall and mAP metrics.

Performance Comparison Table

Model	Recall	mAP50	Performance
YOLOv3	0.96	0.97	Good
YOLOv5nu	0.91	0.973	Moderate
YOLOv5su	0.997	0.998	Best
YOLOv8n	0.98	0.99	Excellent
YOLOv8x	0.994	0.997	Excellent

Experimental Analysis

The analysis showed that almost all YOLO architectures achieved excellent performance in distinguishing manipulated Osteoarthritis X-ray images. However, Lung CT scan images were comparatively more difficult to classify due to subtle manipulation patterns.

Among all architectures, YOLOv5su achieved the highest recall value of 0.997 while maintaining faster execution speed compared to larger YOLOv8 models. YOLOv5nu produced the lowest recall value of 0.91.

The results demonstrate that lightweight YOLO architectures can provide highly accurate medical deepfake detection while supporting real-time

processing.

VI. CONCLUSION

This study proposed a deep learning-based medical deepfake detection framework using multiple YOLO architectures. The rapid growth of deepfake technology has created major challenges in healthcare systems because manipulated medical images may lead to incorrect diagnosis, insurance fraud, and severe threats to patient safety. Two different datasets consisting of Knee Osteoarthritis X-ray images and Lung CT scan images were used for experimentation. Data preprocessing and augmentation techniques were applied to improve dataset quality and model robustness.

Different YOLO architectures including YOLOv3, YOLOv5nu, YOLOv5su, YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x were evaluated. Experimental results demonstrated that YOLO-based models achieved high accuracy in detecting manipulated medical images.

Among all architectures, YOLOv5su achieved the best performance with a recall value of 0.997 while maintaining faster execution speed than larger YOLOv8 models. The results indicate that YOLO-based architectures provide an effective balance between computational efficiency and detection accuracy.

The proposed framework can be applied in hospitals, healthcare centers, and medical imaging systems to improve diagnostic reliability and prevent malicious manipulation of medical data.

Future work may include:

- Expanding the dataset size
- Integrating Transformer architectures
- Real-time deployment
- Multi-modal medical image analysis
- Cloud-based healthcare security systems

VII. REFERENCES

1. Goodfellow, I., et al. "Generative Adversarial Nets." Neural Information Processing Systems, 2014.
2. Solaiyappan, S., and Wen, Y. "Machine Learning Based Medical Image Deepfake Detection: A Comparative Study." Machine Learning with Applications, 2022.
3. Mirsky, Y., et al. "CT-GAN: Malicious Tampering of 3 D Medical Imagery Using Deep Learning." USENIX Security Symposium, 2019.
4. Kim, Y. S., et al. "Development of Deepfake-Based Deep Learning Algorithm for Medical Data Manipulation Detection." Webology, 2022.
5. Albahli, S., and Nawaz, M. "MedNet: Medical Deepfakes Detection Using Improved Deep Learning." Multimedia Tools and Applications, 2023.
6. Krizhevsky, A., et al. "ImageNet Classification with Deep Convolutional Neural Networks." NIPS, 2012.
7. Zhu, J., et al. "Unpaired Image-to-Image Translation Using CycleGAN." ICCV, 2017.
8. Budhiraja, R., et al. "Medical Deepfake Detection Using Convolutional Reservoir Networks." IEEE, 2022.
9. Du, J. "Understanding Object Detection Based on CNN Family and YOLO." Journal of Physics, 2018.
10. Karaköse, M., et al. "A New Approach for Effective Medical Deepfake Detection in Medical Images." IEEE Access, 2024.