Detecting Medical Deepfakes: A Machine Learning Approach to CT Scan Authentication

Team members and emails:

Vishnu Gorur - vishnu_gorur@berkeley.edu Hyeong Geon Kim - hyeonggeonkim@berkeley.edu
Mannan Mishra - mannan_mishra@berkeley.edu Justin Nhan - nhanj@berkeley.edu

1 Motivation

The emergence of medical deepfakes represents a critical cybersecurity threat facing our current healthcare infrastructure. This dataset reveals that attackers can now manipulate medical imagery, with such precision, that even expert radiologists may struggle to distinguish between authentic and manipulated x-ray images.

The implications of these deep fakes drive our passion to develop countermeasures. Unethical groups could remove evidence of real cancers on medical images, potentially causing preventable deaths. Fraudulent scammers can also insert fake cancers, potentially resulting in financial fraud or unnecessary surgeries. Based on deep fake medical scans, insurance companies could be deceived into approving fraudulent claims or denying legitimate coverage. Most disturbingly, court cases involving medical or insurance malpractice can be compromised with tampered scans, resulting in a miscarriage of justice.

First this project will distinguish a real cancerous CT scan from a healthy CT scan irrespective of being real or fake. Building on top of that, we will explore methods to mitigate the risk of medical image tampering and the mistreatment of patients worldwide. By developing detection algorithms that can identify medical imaging deepfakes, we will gain an applied machine learning skill set while enabling patient safety, healthcare integrity, and judicial fairness in this rapidly evolving AI-equipped world.

2 Data

This project uses the Deepfakes: Medical Image Tampering Detection dataset from the UCI machine learning repository. The dataset includes 3D CT scans of human lungs, with some images manipulated to simulate realistic medical deepfakes. These manipulations involve inserting or removing cancerous lesions to create tampered examples that mimic real-world attacks on medical imaging systems.

The dataset contains over 20,000 DICOM image slices (512x512). Each image is labeled according to whether it was altered, and what type of tampering was performed: True-Benign, False-Benign, True-Malicious, False- Malicious.

3 Related Work

1. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning
2. Machine learning based medical image deepfake detection: A comparative study
3. Back-in-Time Diffusion: Unsupervised Detection of Medical Deepfakes

4 Github Repo

Github Repo Link: *https://github.com/jtn50/lungdeepfake*