Capstone Project

Version 1.1.0

Title: *Unveiling AI-Manipulated Medical Images and Localisation of Tampered Areas*

---

1. Introduction

Medical imaging is the backbone of modern healthcare, but recent advances in AI-based image manipulation (e.g., GANs) have introduced serious risks. Attackers can tamper with CT scans by:

- **Injecting fake tumors** into healthy scans (False Malignant → FM).
- **Removing real tumors** from cancerous scans (False Benign → FB).

Such manipulations can be used for **cyberattacks on medical facilities**, **sabotage of research centers**, or **insurance fraud**, thereby threatening patient safety and medical credibility.

This project aims to:

1. Detect manipulated CT images.
2. Classify them into **True Malignant (TM), True Benign (TB), False Malignant (FM), False Benign (FB)**.
3. Localize the tampered regions within the scan.

---

2. Proposed Methodology

We propose a **two-stage pipeline**:

Stage 1: Anomaly Detection & Classification (EfficientNet)

- **Backbone:** EfficientNet (lightweight, scalable, and effective for high-resolution CT scans).
- **Input Branches:**
    1. **Raw CT Images**
    2. **Fourier Domain (DFT Transformed Images)** → helps capture pixel-level inconsistencies left behind by GAN manipulations.
- **Fusion:** Parallel branches ensembled via a small **MLP or voting classifier**.
- **Output Classes:** TM, TB, FM, FB.

Training Strategy:

- Pretrained EfficientNet on medical imaging (transfer learning).
- **Freeze & unfreeze backbone layers** for fine-tuning.
- Semi-supervised approach: mix of image + metadata (CSV annotations).

Loss Function:

- Multi-class classification → **Categorical Cross-Entropy Loss**.
- Add **Focal Loss** variant to handle class imbalance.

Evaluation Metrics:

- **Accuracy, Precision, Recall, F1-Score**.
- **AUC-ROC per class**.
- **Confusion Matrix** (to measure misclassifications TB↔FB, TM↔FM).

Stage 2: Tampering Localisation (UNet)

- **Backbone:** UNet (supervised on bounding box/mask annotations).
- **Input:** CT scan slice + corresponding ground truth coordinates (from CSV).
- **Output:** Tampered region mask with confidence scores.
- **Headmap:** Generates a heatmap highlighting tampered hotspots.

Loss Function:

- **Dice Loss + Binary Cross-Entropy (BCE)** for pixel-level accuracy.
- **IoU Score** as evaluation metric.

---

3. Dataset

We leverage **multiple real and synthetic datasets**:

- **True Malignant (TM):** LIDC-IDRI dataset (real tumors, ~79,400 slices after augmentation).
- **True Benign (TB):** Miskey + Kaggle CT dataset (healthy scans, ~56,000 slices).
- **False Malignant (FM):** CT-GAN injected tumors (~27,000 slices).
- **False Benign (FB):** CT-GAN removed tumors (~89,800 slices).

Preprocessing:

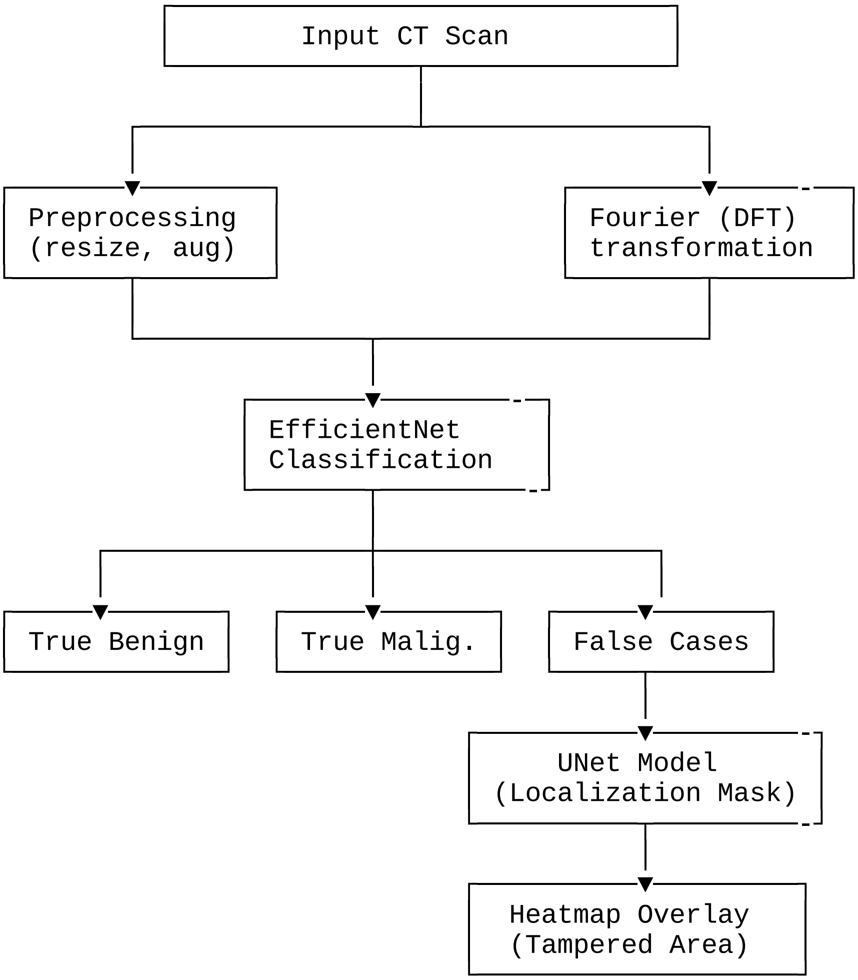
- Original format: DICOM (converted to `.npy`, size 512×512).
- Slice selection: ±5 slices around tumor regions.
- Augmentation:
  - Rotation (±15°).
  - Horizontal Flip.
  - Gaussian Noise.

Dataset Summary Table:

Category	Slice Count	Folder Count	Source/Dataset	Remarks
True Malignant	~79,400	1588	LIDC-IDRI + Augmentation	Real Tumors
True Benign	~56,000	1128	Miskey + Kaggle	Real Healthy
False Malignant	~27,000	540	CT-GAN + Self-generated	Injected Tumors
False Benign	~89,800	1796	CT-GAN + Self-generated	Removed Tumors

4. System Architecture

UML (High-Level Workflow)



5. Expected Outcomes

- **Stage 1 (EfficientNet):** Classifies CT scans into TM, TB, FM, FB.
- **Stage 2 (UNet):** Pinpoints manipulated regions and generates heatmaps.
- **Final Result:** A robust system capable of identifying and localizing AI-driven tampering in medical scans.

Some Anologies:

Using only Fourier vs only Images vs Both

- **Only Images (Raw CT):**
  Rrely on pixel intensity and texture. This works well for detecting real vs fake tumors visually, but subtle GAN artifacts may slip through.
- **Only Fourier:**
  Only capture artifacts and inconsistencies well, but lose anatomical context. The model may detect tampering, but not whether it corresponds to a clinically relevant region.
- **Both (Dual Branch):**
  Best of both worlds. The raw image gives **semantic/structural features** (tumor shape, organ boundaries), while the Fourier branch gives **artifact-level features** (pixel-level manipulation traces). Combining them usually improves generalization and robustnes

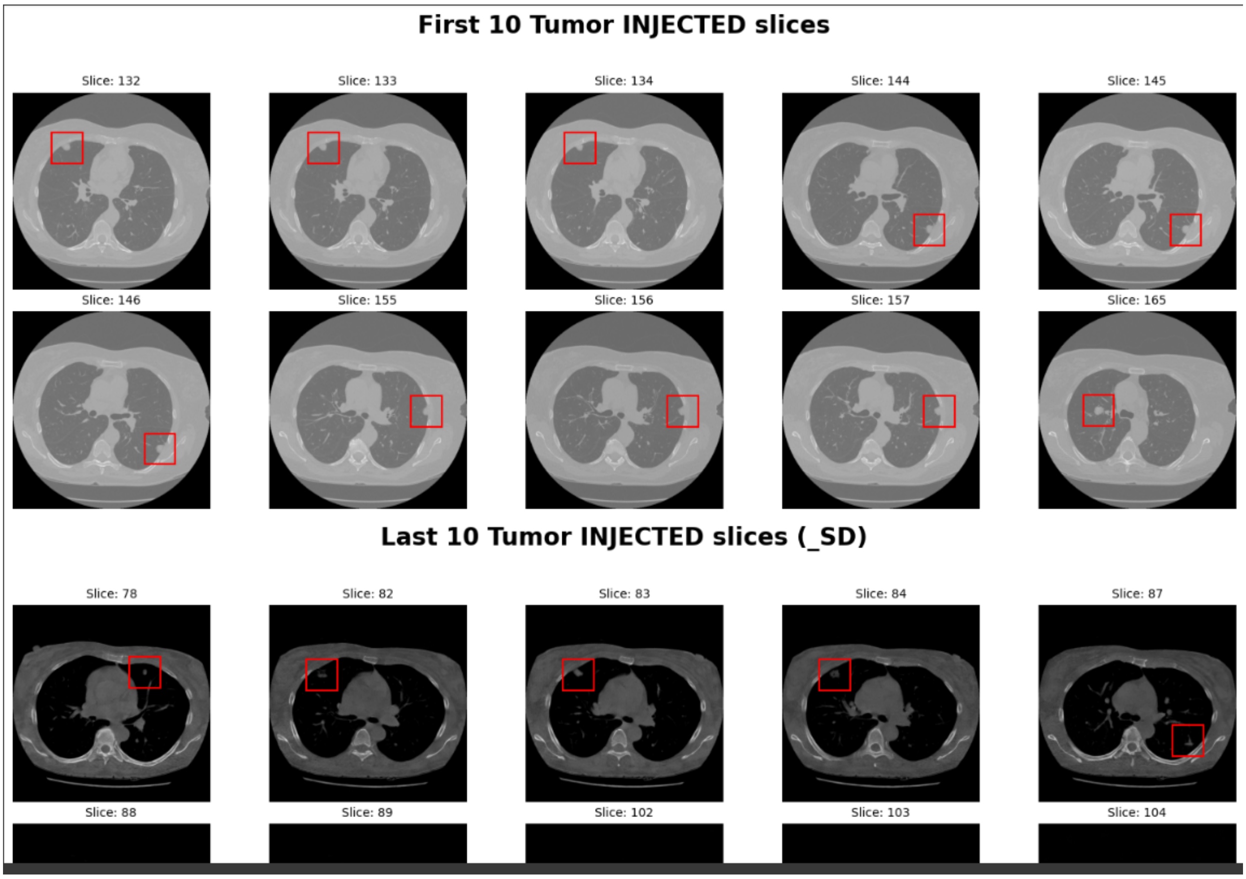
Trade-offs

- **Pros of both:**
  - Higher robustness against unseen GAN manipulations.
  - Captures both semantic (tumor/organ structure) and forensic (artifact) cues.
  - Useful if dataset is diverse and adversary may use different GANs.
- **Cons of both:**
  - More computationally expensive (two branches).
  - Fusion strategy (MLP/voting) needs tuning.
  - Risk of overfitting if dataset is small.

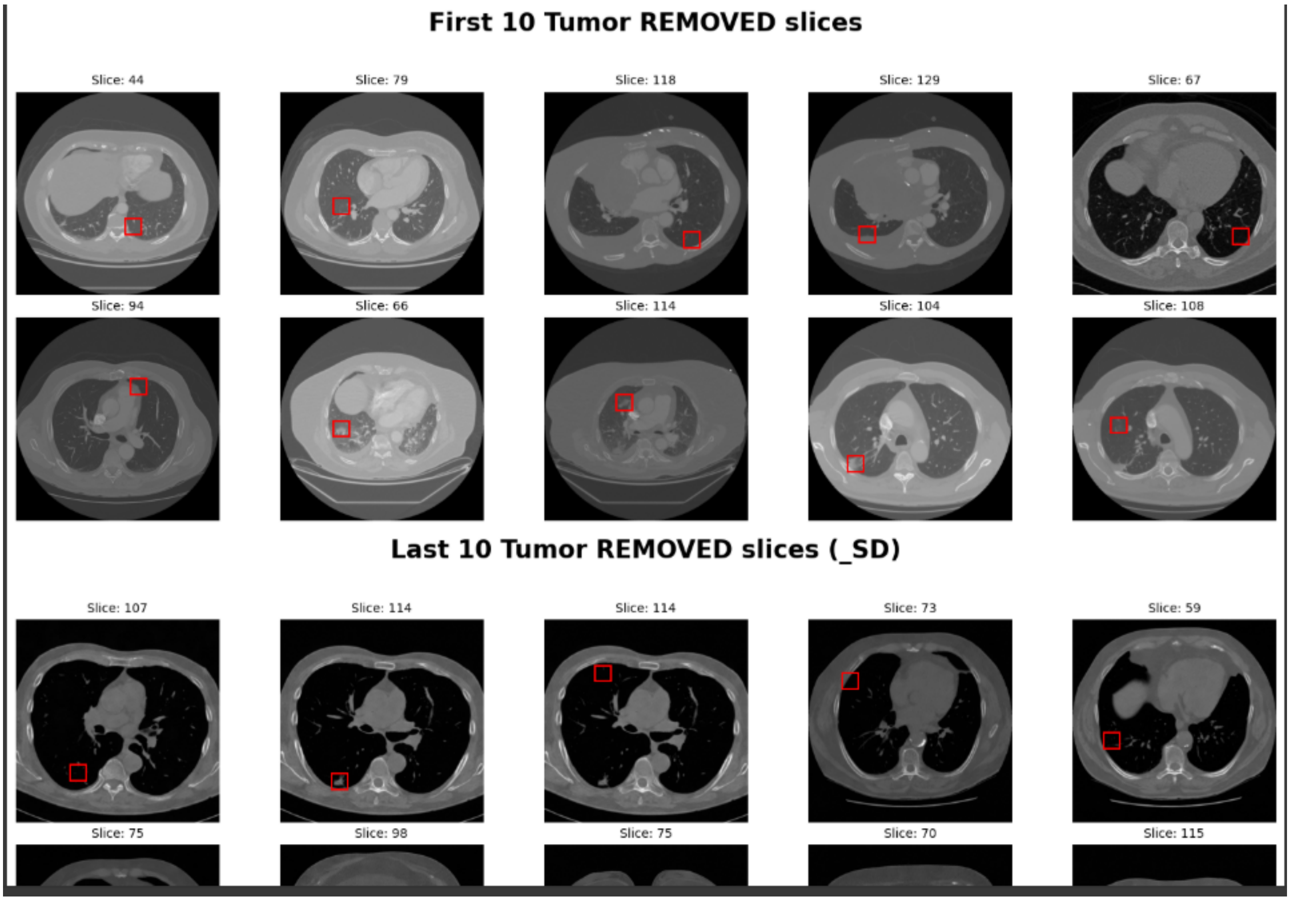
[Start with BOTH also later comare the metrics with individual]

DataSet Images:

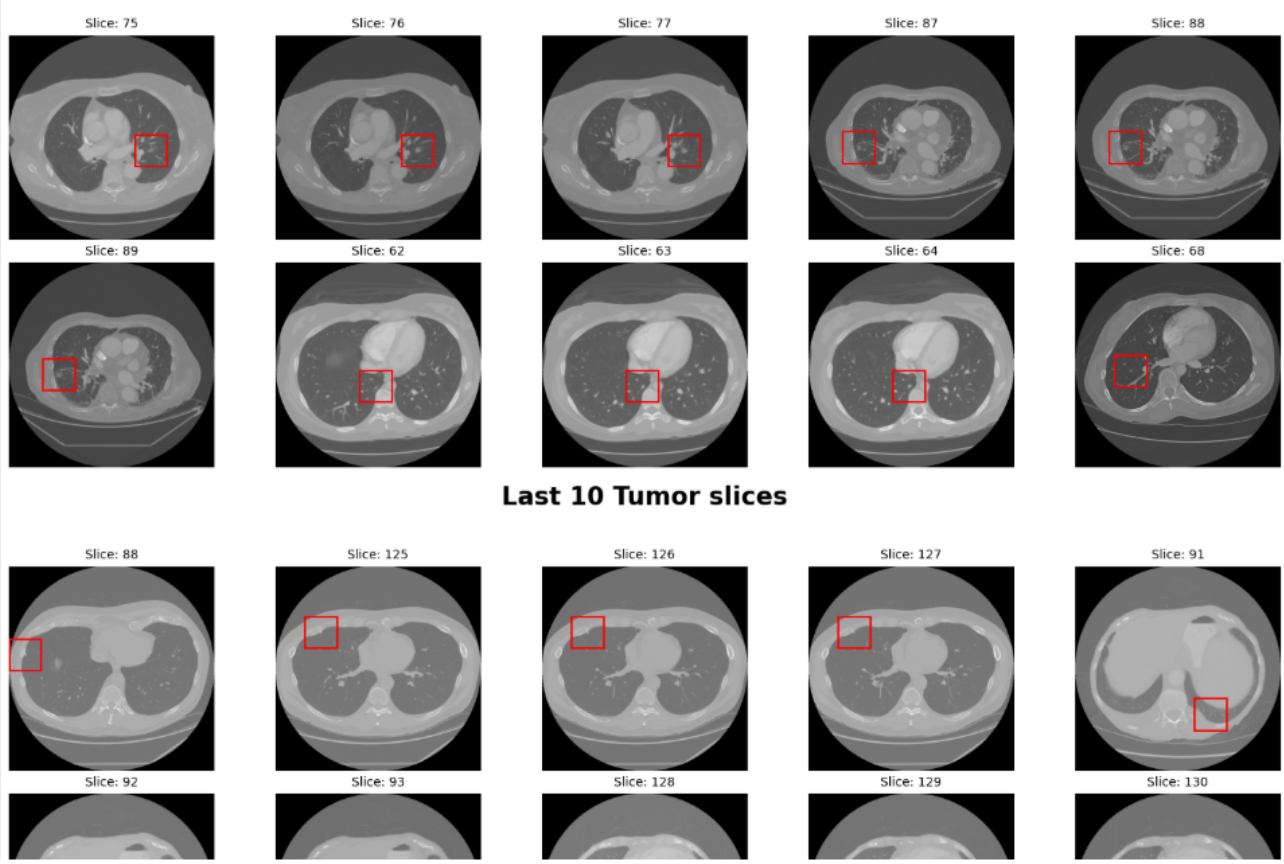
FM:



FB:


First 10 Tumor REMOVED slices
Slice: 44
Slice: 79
Slice: 118
Slice: 129
Slice: 67
Slice: 94
Slice: 66
Slice: 114
Slice: 104
Slice: 108
Last 10 Tumor REMOVED slices (_SD)
Slice: 107
Slice: 114
Slice: 114
Slice: 73
Slice: 59
Slice: 75
Slice: 98
Slice: 75
Slice: 70
Slice: 115

TM:



TB:

```
plot_slices(df.tail(10), "Last 10 True Benign slices")
```