1. Introduction

This project uses medical imaging data (MRI, CT, X-ray) along with Electronic Health Records (EHR) text data to develop AI models for healthcare tasks such as tumor detection, classification, pneumonia detection, and structured/unstructured data analytics.

The datasets were chosen because they are:

- **Open-source and freely available.**
- **De-identified or synthetic, ensuring privacy.**
- **Relevant** for real-world AI applications in radiology and healthcare NLP.

---

2. Dataset Sources

Brain MRI Images for Brain Tumor Detection

- **URL:** Kaggle Dataset
- **License:** Open-source, for research and educational purposes.
- **Type:** MRI brain images with tumor/normal labels.

IQ-OTH/NCCD Lung Cancer Dataset (CT Images, Augmented)

- **URL:** Kaggle Dataset
- **License:** Open-source, academic usage permitted.
- **Type:** CT lung images (normal vs cancerous).

Chest X-ray Images (Pneumonia)

- **URL:** Kaggle Dataset
- **License:** Open-source, academic use.
- **Type:** Chest X-ray images (normal vs pneumonia).
- **Content:** Separate folders for “PNEUMONIA” and “NORMAL” classes.

Cynthia Synthetic EHR Dataset (PDF version)

- **URL:** Kaggle Dataset
- **License:** Open-source, synthetic and de-identified.
- **Format:** 5 PDF files simulating Electronic Health Records.
- **Content:** Structured (demographics, diagnoses, medications) + unstructured (physician notes, discharge summaries).

Electronic Health Record (Anu Chhetry)

- **URL:** Kaggle Dataset
- **License:** Open-source, de-identified.
- **Format:** CSV file with structured EHR data.
- **Content:** Demographics, admissions, diagnoses, and hospital stay information.

---

3. Data Description

Brain MRI Dataset

- ~3,000 MRI images.
- Format: .jpg
- Labels: Tumor vs Normal.

Lung CT Dataset

- ~1,000 CT images (augmented).
- Format: .png
- Labels: Benign, Malignant, Normal.

Chest X-ray Pneumonia Dataset

- ~5,800 images divided into train/test/val sets.
- Format: .jpeg
- Labels: Normal vs Pneumonia (bacterial/viral).

Cynthia Synthetic EHR Dataset (PDFs)

- 5 synthetic EHR PDFs.
- Formats: .pdf
- Content: Structured (diagnoses, procedures, demographics) + unstructured (notes, summaries).

Electronic Health Record (Anu Chhetry)

- 1 structured dataset.
- Format: .csv
- ~1,400 records with 29 attributes.

---

4. Preprocessing Steps

Cleaning

- Removed duplicate and corrupted image files.
- Dropped incomplete/malformed rows in structured EHR datasets.

Standardization

- Resized all images to 256×256 pixels.
- Converted image formats to .png for consistency.
- Normalized EHR text (lowercasing, punctuation cleanup).

Labeling

- MRI & CT: Labels mapped to “Normal”, “Tumor”, “Cancer”.
- X-ray: Labels mapped to “Normal” vs “Pneumonia”.
- EHR: Diagnoses mapped to ICD-10 codes.

Tools Used

- Python (Pandas, NumPy).
- OpenCV & PIL (image resizing/conversion).
- NLTK/Regex (text preprocessing).

---

5. Data Structure (Final Organization)

```
/data
  /images
     /ct
        ct_001.png
        ct_002.png
    /mri
       mri_001.png
       mri_002.png
    /xray
      chest_xray_001.png
      chest_xray_002.png
  /ehr_notes
      ehr_cleansed.csv
```

note_1.txt

note_2.txt

/mapping

ICD-10_mapping.csv

/docs

datasources.md

challenges.md

cleaning_steps.md

ICD-10 mappingnotes.md

README.md

6. Challenges & Decisions

- Class imbalance: More pneumonia X-rays than normal → balanced with augmentation.
- Missing labels: Some EHR rows lacked ICD-10 mapping → excluded.
- PDF extraction issues: Cynthia EHR PDFs required extra parsing/cleanup.
- Corrupted files: ~20 images removed during preprocessing.

---

7. Ethical & Privacy Considerations

- All datasets are open-source and de-identified.
- No personally identifiable patient data (PII) is included.
- Cynthia EHR dataset is synthetic, ensuring no real patient privacy risk.
- Anu Chhetry's dataset is de-identified.

---

8. References

- Brain MRI Dataset – Kaggle: Link
- Lung CT Dataset – Kaggle: Link
- Chest X-ray Pneumonia Dataset – Kaggle: Link
- Cynthia Synthetic EHR Dataset – Kaggle: Link
- Electronic Health Record (Anu Chhetry) – Kaggle: Link

Milestone 1: Data Collection and Preprocessing

The goal of this milestone is to **prepare both imaging and clinical datasets for AI model training**. It involves:

- **Collecting medical imaging datasets** (X-ray, MRI, CT, ultrasound, DXA) to provide unstructured visual data.
- **Gathering structured and unstructured EHR content** such as patient notes and ICD coding data.
- **Cleaning, labeling, and standardizing the data** so it becomes consistent and ready for Generative AI models.