Milestone 2 Documentation

Project Title: AI-Powered-Enhanced EHR Imaging & Documentation System
Team Name & Members: Team A

- Samriddhi Tiwary
- Hima Anjuri
- Pravalika
- Rajeswari
- Dhanasree
- Sushmasri
- Shazfa
- Shobith Reddy
- Gokula Prasath

Course / Semester / Instructor Name: Aryan Khurana

---

1. Introduction

The goal of our project is to integrate **Artificial Intelligence (AI)** and **Generative AI** into healthcare to automate data processing, image enhancement, and clinical documentation.
Milestone 1 marks the foundation — where we collected, cleaned, standardized, and organized medical data so it is ready for AI-driven analysis and automation in later stages.

This stage ensures that our project begins with **high-quality, structured, and ethically prepared data**, setting the base for model accuracy and reliability in Milestone 2 and beyond.

---

2. Dataset Resources

- **Dataset Used:** Heart CT and MRI Dataset

- **Source:** Kaggle (Open-access, anonymized dataset for research)
- **Type:** Unstructured Medical Imaging (CT & MRI)

Why We Chose This Dataset

- Focuses on **cardiac imaging**, one of the most vital areas of medical AI research.
- Includes both **CT and MRI modalities**, offering diversity in data for multimodal AI analysis.
- Open-source and privacy-compliant, suitable for educational and research-based use.

---

3. Data Description

The dataset includes:

- **Heart CT Scans:** Cross-sectional images capturing heart and vessel anatomy.
- **Heart MRI Scans:** Soft-tissue imaging for detailed cardiac structure.
- **Mixed Formats:** Images in .jpg, .png, and some DICOM files.
- **Size:** 300+ images from simulated patients.

Each image represents a unique patient scan that can later be paired with structured EHR-like synthetic data and ICD-10 codes.

---

4. Preprocessing Steps

a) Cleaning

- Removed **corrupted and duplicate** images using hashing and validation scripts.

- Converted all images to **.png** format to maintain consistency and prevent compression loss.
- Filtered out low-quality scans and organized them into CT and MRI folders.

b) Standardization

- **Resized** all images to **256×256 pixels** for uniformity across the dataset.
- Applied **consistent naming convention** (e.g., heartct_001.png, heartmri_002.png).
- Created separate folders for modalities to simplify further processing.

c) Normalization

- Normalized all pixel values between **0 and 1** using NumPy.
- This ensures stability during AI model training and eliminates scale bias.

---

5. Tools and Libraries Used

Tool / Library	Purpose
Python 3.10+	Base language for data processing
Pandas	For handling structured metadata (CSV files)
NumPy	For pixel normalization and array manipulation
PIL (Pillow)	For image conversion and resizing
OpenCV (cv2)	For image validation and visualization
Matplotlib	For visual inspection of images

Tool / Library	Purpose
Hashlib	For duplicate detection
os / shutil	For directory organization and file handling

---

6. Data Structure (Final Organization)

After preprocessing, the final data was organized in a structured and modular way to ensure traceability and reproducibility.

```
/data
  /images
    /CT
      heartct_001.png
      heartct_002.png
    /MRI
      heartmri_001.png
      heartmri_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
```

Folder Explanation

- **/data/images:** Contains cleaned and standardized CT & MRI images.
- **/data/ehr_notes:** Includes synthetic structured EHR records and generated clinical notes.
- **/data/mapping:** Contains ICD-10 code mappings for later automation stages.
- **/docs:** Documentation and notes on data sources, challenges, and cleaning workflows.
- **README.md:** Overview of dataset, methodology, and reproduction steps.

7. Challenges Faced

Challenge	Description	Mitigation
Large File Sizes	Some MRI scans were high-resolution and required optimization.	Processed in smaller batches and resized early.
Corrupted Files	Some files were unreadable or incomplete.	Used try/except scripts to skip invalid files.
Format Inconsistency	Mixed image formats (DICOM, JPG, PNG).	Converted all to PNG using Pillow.
Limited Local Resources	GPU/CPU limitations for large-scale processing.	Used lightweight libraries and smaller sample subsets.
Metadata Gaps	Metadata Gaps	Generated synthetic EHR metadata aligned by filename.

8. Ethical and Privacy Considerations

- Used anonymized, open-source datasets — no real patient data was used.
- All generated EHR notes are synthetic and non-identifiable.
- Maintained strict separation of raw and processed data to ensure data integrity.
- Followed FAIR (Findable, Accessible, Interoperable, Reusable) data principles for proper dataset handling.
- Ensured that enhancement and cleaning steps did not alter diagnostic details or introduce bias.

These steps make our data ethically compliant and aligned with research data governance standards.

---

9. References

- Heart CT and MRI Dataset (Kaggle)
- Dong, C., Loy, C. C., He, K., & Tang, X. (2016). *Image Super-Resolution Using Deep Convolutional Networks (SRCNN)*.
- Python Documentation: https://docs.python.org
- Pillow (PIL) Library: https://pillow.readthedocs.io
- OpenCV Documentation: https://docs.opencv.org

---

10. Summary

Milestone 1 established the data foundation for our entire AI-based healthcare project.
We collected, cleaned, standardized, and ethically organized the

Heart CT & MRI dataset into a format ready for deep learning applications.

The structured organization and documentation ensure that our data pipeline is:

- **Reproducible – Anyone can follow our process and recreate the dataset.**
- **Reliable – Verified for quality, uniformity, and ethical safety.**
- **Ready for AI – Suitable for enhancement (Milestone 2) and clinical integration (Milestones 3 & 4).**

This milestone reflects our commitment to data integrity, ethical AI usage, and real-world clinical readiness — setting a strong foundation for the rest of the project.