Descriptive Report: Research Advancement Module

1. Introduction

This report describes the progress made in the "Research Advancement" module, focused on developing a multimodal Artificial Intelligence (AI) model to improve the accuracy of medical diagnoses in pulmonary diseases. The project aims to integrate image data (X-rays, CT scans) and text data (medical histories, reports) for a more holistic patient assessment.

2. Motivation and Justification

AI-based medical diagnostic systems have the potential to revolutionize medicine, providing faster and more accurate diagnoses. However, many current systems work with unimodal data, analyzing only medical images or clinical texts in isolation. This limitation prevents a holistic evaluation of the patient, leading to possible diagnostic failures. This study aims to fill this gap by developing a study that seeks to understand and enable multimodal AI models that integrate image data (magnetic resonances, X-rays) with textual data (medical records, reports), increasing the accuracy and personalization of diagnoses in diseases involving the lung.

Reasons for the study:

- Academic: Contribute to the literature of AI applied to health, especially in the field of multimodality.

- Social: Improve the quality of medical diagnoses, reducing clinical errors and improving treatments.

- Practical: Develop a system applicable to hospitals and clinics, promoting efficiency in diagnosis.

3. Research Problem

AI-based medical diagnostic systems still face challenges in interpreting complex data, especially when considering multiple sources of information. Unimodal models are limited because they analyze images or texts separately, without exploring the synergy between the two types of data. Given this context, the following research question arises: How can a multimodal artificial intelligence model improve the accuracy of medical diagnoses by integrating image and text data?

This investigation seeks to answer:

- What are the main limitations of unimodal models in the detection and analysis of diseases involving the lung?

- How can the fusion of textual and visual data increase accuracy in predicting medical diagnoses?

- Which multimodal learning techniques are most effective for correlating information from images and clinical texts?

- What impact can the explainability of multimodal models have on the acceptance of these systems by health professionals?

These questions guide the research and will allow for a detailed analysis of the applicability of multimodal AI in diagnostic medicine.

4. Objectives

General Objective

Develop an AI model that combines image and text data for more accurate and personalized medical diagnoses in lung diseases.

Specific Objectives

- Review existing approaches to multimodal AI in the health area.

- Collect and process medical image data and clinical texts.

- Develop independent models for image and text analysis.

- Integrate the models into a single multimodal architecture.

- Ensure the security of medical and patient data.

5. Scope

The project will cover from initial research to the implementation of a simple functional multimodal AI model to identify diseases that affect the lung. It includes literature review, data collection and processing, development of individual models, multimodal integration, testing and possibly clinical validation.

6. Research Methodology

- Research Type: Exploratory and applied.
- Approach: Quantitative and experimental.
- Data Collection Methods: Public medical databases (MIMIC-IV, CheXpert).
- Analysis Techniques: Training of deep learning models (CNN for images, NLP for texts), statistical validation with performance metrics (accuracy, F1-score).

7. Data Inclusion/Exclusion Criteria - Sprint 2

This document defines the inclusion and exclusion criteria for the selection of medical image data (X-rays and computed tomography scans of the chest) to be used in the lung disease detection project. These criteria will be refined throughout the project as we explore the data and gain more knowledge about the domain.

Inclusion Criteria

The following criteria MUST be met for a dataset or an individual image to be included in the study:

1. Image Modality:
    - ○ The image MUST be of the chest.
    - ○ Accepted modalities are radiography (X-ray) and computed tomography (CT).
2. Labels/Diagnoses:
    - ○ The image MUST have an associated label, indicating the presence or absence of pulmonary disease. Ideally, the label should specify the type of disease (e.g., pneumonia, COPD, pulmonary nodule, etc.). In cases of nodules, the location (coordinates) of the nodule is also desirable.
    - ○ Labels must be provided by radiologists or specialists, or derived from reliable radiological reports.
3. Image Quality:
    - ○ The image MUST be of sufficient quality to allow for visual analysis and processing by algorithms.
    - ○ A minimum resolution will be defined after the initial exploratory data analysis (Sprint 3). For now, we will avoid images with severe artifacts, excessive noise, or apparent low resolution.
4. Documentation:
    - ○ The dataset MUST have complete documentation that allows understanding of the origin and characteristics of the data.

8. Data Sources

The data sources considered for the project include:

- NIH Chest X-ray Dataset
- LIDC-IDRI
- MIMIC-CXR
- Open-i
- Kaggle (search for relevant datasets)

9. Sprint 3: Initial Dataset Analysis

Sprint Objective: To perform the initial exploration, quality assessment, and exploratory analysis of the identified chest X-ray datasets, aiming to select the most promising ones for the next project stages.

Activities Performed:

1. Initial Exploration of Identified Datasets:

We conducted an initial analysis of the structure, size, and format of each of the identified datasets: NIH Chest X-ray Dataset, LIDC-IDRI, MIMIC-CXR, and Open-i. We documented the main characteristics of each dataset, including the number of images, the presence of metadata (reports, demographic information, etc.), and the complexity of the data organization. We identified the different types of files present in each dataset (images, annotation files, text files with reports, etc.).

2. Data Quality Assessment:

A preliminary assessment of data quality was conducted in each dataset.

- Checking for the presence of corrupted or illegible images.
- Analyzing the consistency and integrity of the available metadata.
- Identifying possible class imbalance problems (if applicable and available in the metadata).
- For the LIDC-IDRI dataset, we began analyzing the complexity of multiple radiologist annotations.

3. Exploratory Data Analysis:

- Analysis of descriptive statistics to understand the data distribution (e.g., age distribution, gender, most frequent pathologies, when applicable).

- Explored the distribution of different pathologies mentioned in the reports (when available).

- For the MIMIC-CXR dataset, we started the exploration of the structure and content of the radiologic reports.

Data Source Evaluation

Objective: To evaluate the characteristics, quality, potential, and limitations of the identified chest X-ray datasets for the project.

Datasets Evaluated:

NIH Chest X-ray Dataset

- Description: A large dataset containing over 100,000 chest X-ray images from patients at the National Institutes of Health (NIH). The images are accompanied by textual reports that indicate the presence or absence of up to 14 different pathologies.

- Accessibility: Publicly available for download.

- Structure: The images are organized into folders by pathology. The reports are in separate text files.

- Quality: Image quality may vary. The reports are generated through an automated process and may contain inaccuracies.

- Potential: Useful for multiple pathology classification tasks due to the large volume of data and the presence of reports.

- Limitations: Reports may be less detailed compared to complete radiological reports. The absence of detailed annotations (bounding boxes, segmentations) limits the use for tasks that require precise location of pathologies.

LIDC-IDRI (Lung Image Database Consortium image collection)

- Description: A dataset composed of chest computed tomography (CT) scans with pulmonary nodule annotations performed by multiple radiologists. Although the main focus is CT, some institutions also included X-ray images.

- Accessibility: Requires registration and approval for access through the Cancer Imaging Archive (TCIA). The availability of X-ray images may be limited compared to CT images.

- Structure: Data is organized by patient and study.