# ED6001 – MEDICAL IMAGE ANALYSIS- MINI PROJECT

GROUP_MEMBERS

1.Duen Michael Chombo id24m803

2.Harepriya A G ED20BO23

3.Tegsemen Tizazu Eskezia MM24S800

Github link: https://github.com/duenchombo/ED6001-MEDICAL-IMAGE-ANALYSIS--MINI-PROJECT

# 1. Problem Definition & Dataset

## Problem Statement

Lung cancer remains the leading cause of cancer-related deaths worldwide, accounting for millions of fatalities each year. Alarmingly, only 16% of cases are diagnosed at an early, localized stage, where the five-year survival rate exceeds 50%. However, once detected at advanced stages, the survival rate drops drastically to nearly5%. Early diagnosis is therefore critical for improving patient outcomes.

With the advent of non-invasive imaging modalities such as Computed Tomography (CT), clinicians can obtain rich information on lung tissue characteristics. This enables the development of Computer-Aided Diagnosis (CAD) systems capable of assisting radiologists by classifying lung nodules as benign or malignant, based on extracted imaging biomarkers.

## Literature Context

Several studies have highlighted the potential of radiomics the extraction of high-dimensional quantitative features from medical images to provide reproducible and objective measures of tumor heterogeneity By combining radiomics features with machine learning models, CAD systems have demonstrated improved diagnostic accuracy and consistency compared to manual radiologist assessments, which often suffer from inter-observer variability..

By combining radiomics features with machine learning models, CAD systems have demonstrated improved diagnostic accuracy and consistency compared to manual radiologist assessments, which often suffer from inter-observer variability.

## Dataset Description

This project uses the **LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative)** dataset, a publicly available repository containing **expertly annotated CT scans** of the human thorax.

Each scan contains one or more nodules annotated by up to four radiologists, with malignancy ratings on a **five-point scale** (1 = highly benign, 5 = highly malignant).

After data curation, the dataset was reduced to **binary labels**:

- **0 – Benign** (malignancy ≤ 3)
- **1 – Malignant** (malignancy ≥ 4)

A total of **2,617 samples** were used, each represented by **1,619 radiomic features** extracted via the **PyRadiomics** library.

### Preprocessing Steps

1. **CT Scan Parsing** using the pylidc library to load scans and nodules.
2. **Radiomics Feature Extraction** using PyRadiomics with a YAML-based parameter configuration.
3. **Merging of Binary Labels and Feature Tables** resulting in a final dataset of shape **(2617, 1619)**.
4. **Normalization** of continuous features and **handling missing values** (mean imputation).
5. **Dataset Split** using **Stratified Group 10-Fold Cross-Validation** to preserve patient grouping and class balance.

---

# 2. Methodology & Implementation

### Model Selection

A **Support Vector Machine (SVM)** was selected as the primary model for classification due to its:

- Robust performance on **high-dimensional datasets**,
- Effectiveness in **binary classification**, and
- Ability to **maximize the margin** between malignant and benign classes.

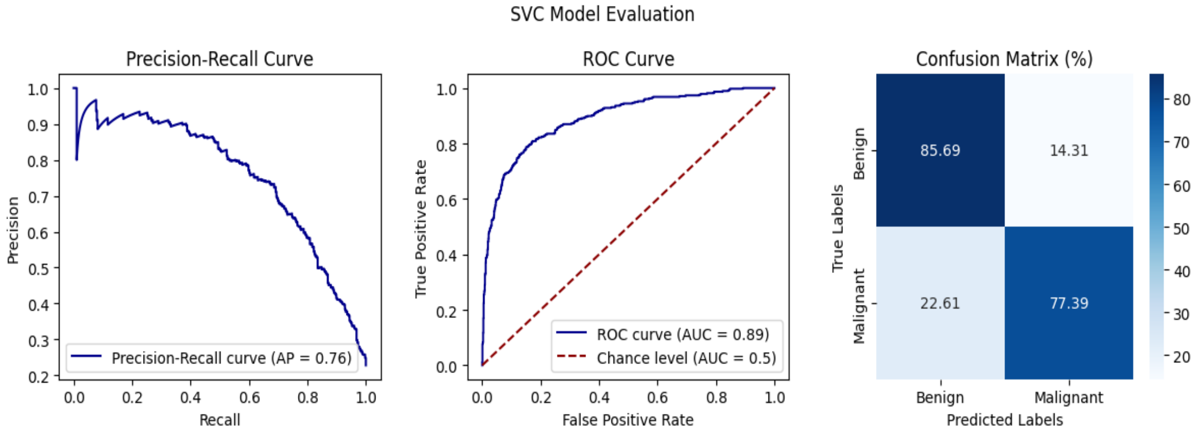
### Implementation Pipeline

1. **Feature Extraction:**
   Radiomics features (first-order, shape-based, GLCM, GLRLM, GLSZM, and GLDM) were extracted using PyRadiomics.
2. **Parameter Optimization:**
   - A **Grid Search** was conducted to identify the best hyperparameters for the SVM model (kernel, C, gamma).
   - The optimal configuration was selected based on **cross-validation accuracy and recall**.
3. **Cross-Validation Strategy:**
   - Applied **Stratified Group 10-Fold Cross-Validation**, ensuring that:
     - Each fold maintains class distribution (benign vs malignant),
     - Data from the same patient does not appear in both training and test sets.

4. **Evaluation Metrics:**
   The following metrics were computed for each fold and averaged across all runs:
   - **Accuracy**
   - **Balanced Accuracy**
   - **Precision**
   - **Recall**
   - **F1-Score**
   - **AUC (Area Under ROC Curve)**
   - **Average Precision (AP)** from Precision-Recall Curve
   - **Log Loss**
   - **Hamming Loss**
5. **Visualization:**
   - **Confusion Matrix** to assess classification distribution
   - **ROC Curve** to evaluate sensitivity vs specificity trade-off
   - **Precision-Recall Curve** to analyze the balance between precision and recall in imbalanced settings

---

# 3. Results & Evaluation Metrics



### Quantitative Results

| Metric | Balanced Accuracy (Scoring) | Recall (Scoring) |
|---|---|---|
| **AUC** | **0.89** | **0.87** |
| **Average Precision (AP)** | **0.75** | **0.73** |
| **Accuracy** | 0.83 | 0.80 |
| **F1-Score** | 0.77 | 0.74 |
| **Recall** | 0.78 | 0.81 |
| **Precision** | 0.76 | 0.71 |

### Qualitative Analysis

- With **balanced accuracy scoring**, the SVM classifier achieves **high discrimination capability** between benign and malignant nodules, as reflected by an **AUC of 0.89**.
  The confusion matrix shows that **most benign nodules are correctly classified**, while only **~23% of malignant cases** are misclassified.
- Under **recall-based scoring**, the model becomes **more sensitive to benign cases**, slightly sacrificing overall precision.
  The AUC and AP decrease marginally, suggesting a trade-off between minimizing **false negatives** and maintaining **balanced performance**.

### Visual Results

- **ROC Curve:** Demonstrates strong separation between classes (AUC ≈ 0.89).
- **Precision-Recall Curve:** Indicates good class balance and high model confidence.
- **Confusion Matrix:** Highlights effective classification of benign nodules, with controlled false positive rate.

---

## 4. Conclusion

This mini project successfully implemented 3D Lung Lesion Detection using Texture and Morphological Features The **SVM model**, when optimized and evaluated using rigorous cross-validation, achieved **strong discriminatory performance**, making it a reliable candidate for clinical CAD systems.