Integration of Imaging Features and Clinical Features for Early Detection of Lung Cancer

Leonard Garcia[1], Luoting Zhuang[2], Yannan Lin[2], William Hsu[2]

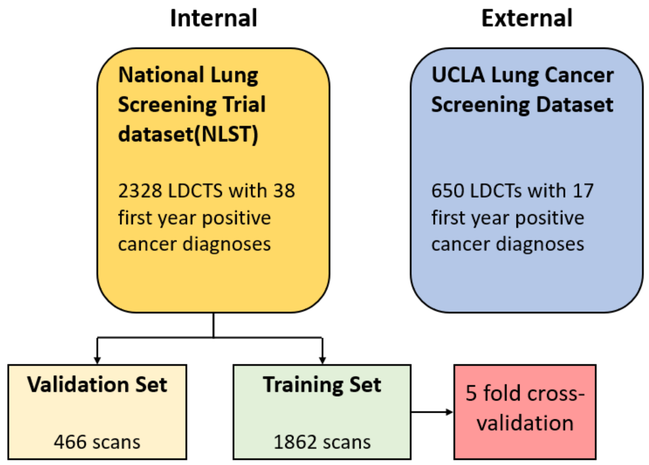
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, University of California Los Angeles, CA 90095, USA. [2] Medical & Imaging Informatics, Department of Radiological Sciences, David Geffen School of Medicine, University of California Los Angeles, CA 90095, USA.

Introduction

- **Lung Cancer Mortality:** Lung cancer remains the leading cause of cancer-related deaths globally.
- **Machine Learning in Detection:** Significant advancements have been made in using machine learning models to predict lung cancer from low-dose chest computed tomography (LDCT) scans.
- **Limitation of Current Models:** Existing imaging-based models often fail to consider crucial clinical information, potentially limiting their effectiveness.
- **Project Objective:** This project aims to develop multimodal machine learning algorithms that integrate both imaging features extracted from the lung cancer risk prediction model, Sybil [1], with clinical features from the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial data to enhance lung cancer risk prediction.

Methods

Datasets

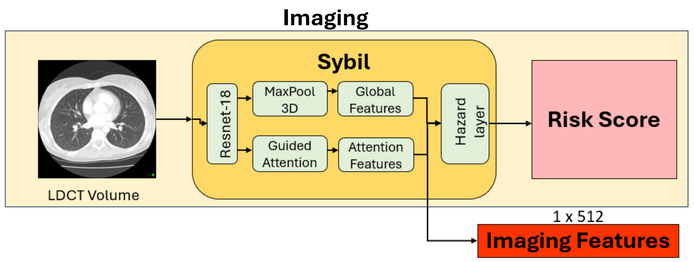


Scan Selection

NLST:
- Orientation: axial
- Convolution Kernel:
  - medium, sharp
- Slice Thickness: 1-3.2mm

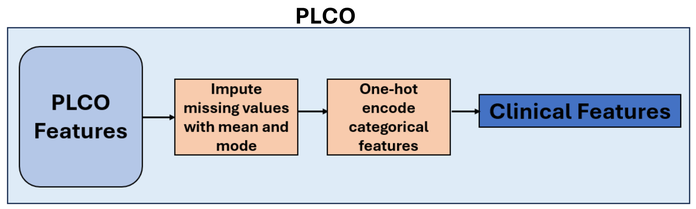
UCLA:
- Orientation: axial
- Convolution Kernel:
  - medium, sharp
- Slice Thickness: 1mm

Feature Extraction

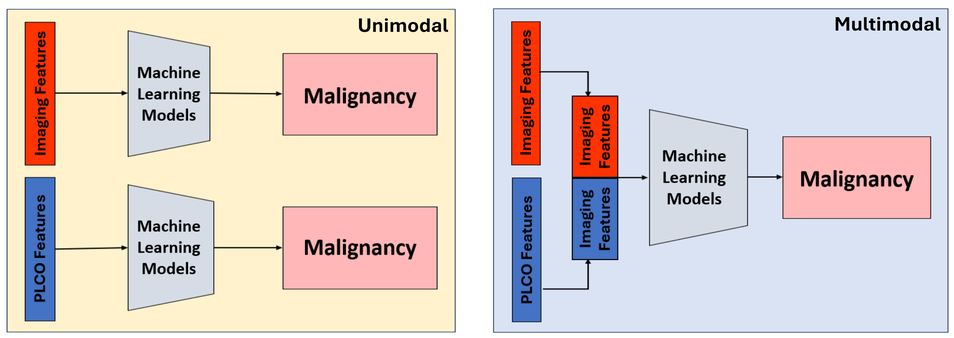


PLCO Features

- Age at time of screening
- Education Level
- Race/ethnicity
- Body Mass Index (BMI)
- Personal history of lung cancer
- Family history of lung cancer
- Personal history of chronic obstructive pulmonary disease (COPD)
- Smoking status
- Smoking intensity
- Smoking duration
- Smoking quit year



Model Architecture



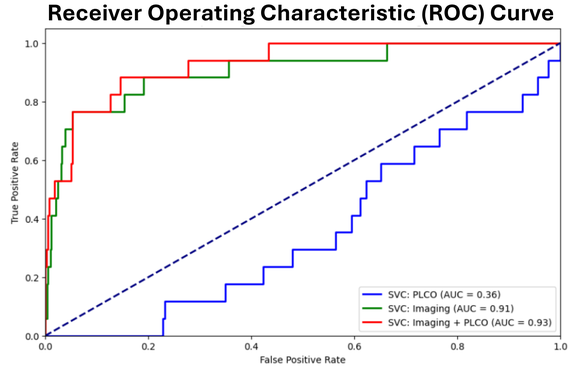
Results

Various machine learning models were trained on the NLST dataset and evaluated on the held-out NLST test set and UCLA dataset to predict lung cancer risk within one year using AUROC and AUPRC as performance metrics.

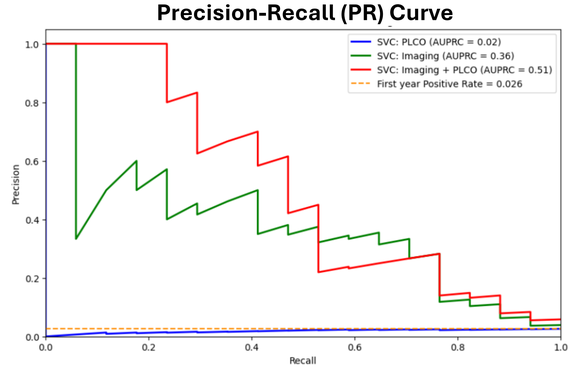
We compared three combinations of features:
1. Only PLCO clinical features
2. Only extracted imaging features
3. PLCO clinical and extracted imaging features together.

Model	Test AUROC	UCLA AUROC	Test AUPRC	UCLA AUPRC
PLCO Features Only				
Logistic Regression	0.579	0.584	0.046	0.052
Random Forest	0.583	0.624	0.040	0.052
XGBoost	0.547	0.586	0.026	0.044
SVC	0.594	0.358	0.021	0.018
Imaging Features Only				
Logistic Regression	0.900	0.890	0.286	0.372
Random Forest	0.880	0.848	0.227	0.343
XGBoost	0.852	0.909	0.141	0.454
SVC	0.908	0.907	0.301	0.383
PLCO + Imaging Features				
Logistic Regression	0.873	0.893	0.293	0.385
Random Forest	0.880	0.857	0.231	0.357
XGBoost	0.852	0.901	0.167	0.407
SVC	0.909	0.933	0.244	0.513

Receiver Operating Characteristic (ROC) Curve



Precision-Recall (PR) Curve

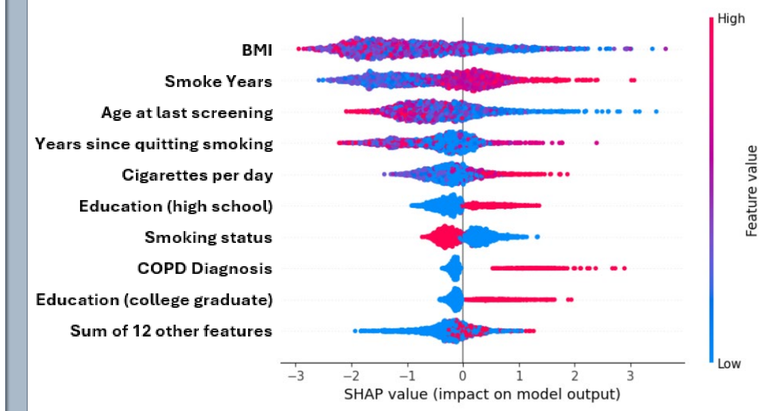


ROC and PR curves for the model that achieved the highest AUC and AUPRC on the UCLA dataset, SVC.

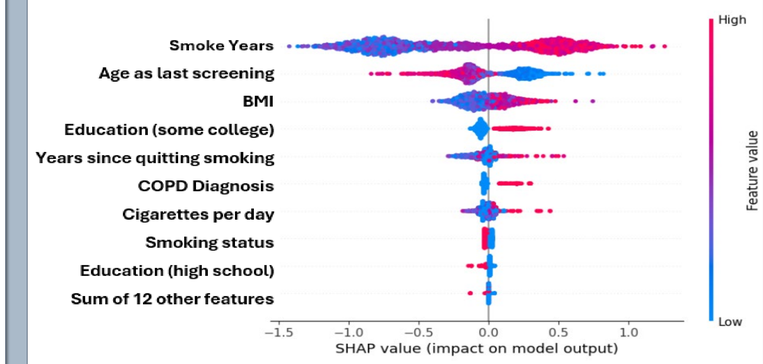
- SVC using both imaging and clinical features scored an AUC of 0.93 and AUPRC of 0.51, the best performance between all models and combination of features.

Results

SHAP Summary Plot (PLCO)



SHAP Summary Plot (PLCO + Imaging)



SHAP (**SH**apley **A**dditive ex**P**lanations) values represent the impact of each feature on the output of a machine learning model. Red dots represent high values, blue dots, low.

- Top and bottom figures display PLCO feature importance for SVC model trained on only PLCO features and PLCO + imaging features together, respectively.
- The three most influential features for both sets of training features include: the number of years smoked, age at last screening, and Body Mass Index (BMI).

Conclusions and Future Directions

- The results of this project suggest models utilizing both clinical PLCO and imaging features generally achieve higher AUROC and AUPRC values compared to models using only one type of feature.
- The improvement in model performance with the integration of imaging and clinical features supports the importance of leveraging multiple data types for enhancing lung cancer risk prediction.
- Future work includes developing more effective methods of combing clinical and imaging data, including deep learning and ensemble models.

References

[1] Mikhael PG, Wohlwend J, Yala A, Karstens L, Xiang J, Takigami AK, Bourgouin PP, Chan P, Mrah S, Amayri W, Juan YH. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. Journal of Clinical Oncology. 2023 Apr 20;41(12):2191-200.