INTEGRATION OF IMAGING AND CLINICAL FEATURES FOR EARLY DETECTION OF LUNG CANCER

A Thesis

Presented to

The Faculty of the Department of Computer Science

California State University, Los Angeles

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Computer Science

By

Leonard Garcia

December 2024

© 2024

Leonard Garcia

ALL RIGHTS RESERVED

The thesis of Leonard Garcia is approved.

Navid Amini, Committee Chair

Mohammad Pourhomayoun

Dr. Eun-Young (Elaine) Kang, Department Chair

California State University, Los Angeles

December 2024

TABLE OF CONTENTS

ABSTRACT

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

By

Leonard Garcia

Lung cancer is the leading cause of cancer-related mortality. Much work has been done on developing both statistical models and machine learning models that can predict the likelihood of lung cancer development from clinical features and low-dose chest computed tomography scans, respectively. Although promising, single modality models can overlook valuable information. Therefore, this project developed machine learning algorithms to integrate imaging features extracted from a lung cancer risk prediction model, Sybil, and PLCO clinical features to provide a more accurate and holistic lung cancer risk assessment. We concatenated imaging and PLCO features and trained multimodal machine learning models on NLST data and evaluated on a UCLA dataset. The best-performing of these multimodal models achieved an AUROC score of 0.933 and AUPRC of 0.513, outperforming all unimodal models. Therefore, multimodal fusion of imaging and clinical features may enhance early detection of lung cancer and reduce mortality rates.

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisor Dr. Amini for his continued help and support. He served as a constant source of guidance not just for this project but for my academic and professional career.

Secondly, I would like to thank Dr. Hsu and Luoting Zhuang for their direction in this project and for introducing me to the field of Biomedical Data Science.

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

Introduction

1.1 Lung Cancer Risk Modeling

Cancer is the second-leading cause of death in the United States. Among people under 85 years of age, it is the leading cause of death. Lung cancer death is the most common cancer death accounting for approximately 20% of all cancer deaths [1]. It is a condition where cells in the lung tissues grow uncontrollably and form a malignant tumor. It can be categorized according to stages which measure the degree of cancer spread. In the least severe, stage I, the cancer has not spread beyond the lungs to other parts of the body. In the most severe, stage IV, the cancer has spread to other parts of the body, such as the bones or brain [2]. Detection at an early stage is strongly associated with better patient outcomes. When detected during stage I, the 5-year survival rate of lung cancer is 90%. When detected during stage IV, the survival rate is less than 10% [3]. Because early diagnosis has such a large impact on patient survival, much work has been done on developing early lung cancer risk prediction models.

Lung cancer risk prediction models are tools designed to estimate the likelihood an individual will develop lung cancer. They can be used to help guide decisions about who should undergo medical screenings by identifying high risk individuals. They can also be used to inform public health interventions by identifying high risk populations [4]. Several types of lung cancer risk prediction models have been developed.

Clinical models work by considering several risk factors, such as demographic information and medical history, to calculate the likelihood of a patient developing lung cancer. These models use statistical techniques to interpret the relationships between these risk factors and outcomes based on historical data [4].

Imaging models use imaging data from medical scans such as low-dose computed tomography (CT) scans to detect diseases. In the case of lung cancer, they can be used to identify the location of growths in the lung nodules called nodules. Imaging models can also be used to categorize these nodules as being benign or malignant, that is cancerous or not cancerous. Current imaging models are largely deep-learning based [5].

This project makes use of two currently existing lung cancer risk models: the PLCOm2012 Model, a clinical model, and Sybil, an imaging-based model.

1.12 PLCOm2012 Model

The PLCOm2012 Model is a statistical lung cancer risk prediction model developed to improve the selection of individuals for lung cancer screening. It was first proposed by Dr. Martin C. Tammemägi and his team in [6]. It was developed using data from Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. It considers a broad range of risk factors including both demographic information and medical history and estimates the likelihood an individual will develop cancer within 6 years. The full list of risk factors is listed in the Table 1 below.

Risk Factor	Description
Age	Age of the individual
Race/Ethnicity	The individual's race/ethnicity
Education	Highest level of schooling
Body Mass Index (BMI)	A measure of body fat based on height and weight
Chronic Obstructive Pulmonary Disease (COPD)	Presence of COPD
Personal History of Cancer	Whether an individual has ever been diagnosed with cancer
Family History of Lung Cancer	Whether an individual has a first-degree relative diagnosed with lung cancer
Current Smoker	Whether an individual currently smokes cigarettes
Smoking Intensity	Number of packs of cigarettes an individual smokes in a day
Smoking Duration	Number of years an individual has smoked
Years Since Quitting	For former smokers, the number of years since quitting.

Table 1: PLCOm2012 risk factors

Using the above factors, the risk score is calculated as follows:

$$
\begin{aligned}
\text{Logit} = & -4.532506 \\
& + 0.0778868 \times \text{Age} \\
& + 0.3944778 \times \text{Black} \\
& - 0.7434744 \times \text{Hispanic} \\
& - 0.466585 \times \text{Asian} \\
& - 0.0812744 \times \text{Education Level} \\
& - 0.0274194 \times \text{BMI} \\
& + 0.3553063 \times \text{COPD} \\
& + 0.4589971 \times \text{Personal History of Cancer} \\
& + 0.587185 \times \text{Family History of Lung Cancer} \\
& + 0.2597431 \times \text{Current Smoker} \\
& - 1.822606 \times \log(\text{Smoking Intensity}) \\
& + 0.0317321 \times \text{Smoking Duration} \\
& - 0.0308572 \times \text{Years Since Quitting}
\end{aligned}
$$

$$\text{Risk Score} = \frac{e^{logit}}{1+e^{logit}}$$

We note that race could have either a positive or negative effect on the risk score. This is based on empirical evidence that lung cancer risk is not uniform across all racial and ethnic groups. Studies have shown that Black individuals tend to have higher rates of lung cancer compared to other racial groups, whereas Hispanic and Asian populations typically have lower lung cancer incidence [7]. Moreover, there is an inverse relationship between education and lung cancer risk: individuals with lower levels of education tend to have higher rates of lung cancer [8]. When factoring education into the above equation, it is measured in six ordinal levels: less than high-school graduate (level 1), high-school graduate (level 2), some training after high school (level 3), some college (level 4), college graduate (level 5), and postgraduate or professional degree (level 6) [6].

1.13 Sybil

Sybil is a deep-learning based imaging model developed by Massachusetts Institute of Technology, in collaboration with Massachusetts General Hospital (MGH) in 2022. It can predict an individual’s risk of developing lung cancer within 1 to 6 years from a single low-dose CT scan. Sybil differs from many other imaging models by not requiring manual nodule annotation or any additional clinical inputs, relying solely on the CT scan [9].

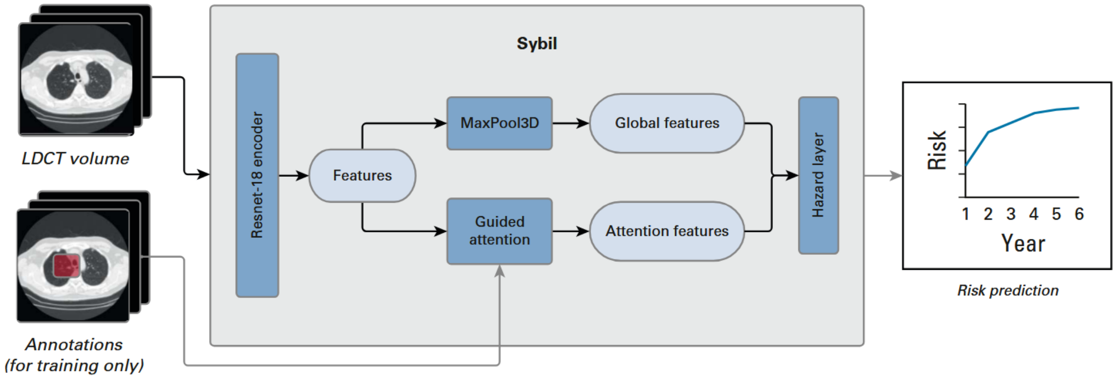


Figure 1: Architecture of Sybil

A diagram depicting the architecture of Sybil is shown in Figure 1 above. The model takes low-dose CT scans as input. The CT scans are then passed through a Res-Net-18 encoder, a deep convolutional neural network (CNN) originally designed for image recognition tasks. Resnet-18 is commonly used for image feature extraction due to its ability to learn hierarchical representations [10]. This encoder extracts features from the CT scan representing important patterns or structures in the scan, such as nodules or other lung abnormalities. The extracted features are then put through a MaxPool3D layer which reduces the dimensions while retaining key information. In addition to global features, Sybil uses a guided attention mechanism. This layer focuses on regions of interest within the scan. The output of this layer is a set of attention features, which are high relevance parts of the input data.

In the final layer, the global and attention features are combined and passed through a hazard layer. A hazard layer is based on the concept of a hazard function which is used in survival analysis to predict the risk or an event occurring over a period of time [11]. The final output of the model is a risk prediction curve which represents the estimated risk of lung cancer over a 1-to-6-year period.

1.2 Thesis Statement and Main Contributions

This thesis investigates the application of multimodal machine learning frameworks to improve early lung cancer detection. The primary objective is to develop and evaluate multimodal machine learning algorithms that integrate both imaging features extracted from the lung cancer risk prediction model, Sybil, with the clinical features used in the PLCOm2012 statistical lung cancer risk prediction model. This approach aims to improve early-stage cancer detection, potentially reducing mortality rates.

Although previous research has explored the use of multimodal data for lung cancer prediction, this work is the first to combine Sybil imaging features with PLCO clinical data, and thus represents a novel approach.

CHAPTER 2

Literature Review

2.1 Current Work

There has been much work done in developing multimodal lung cancer risk models, both in predicting the risk of lung cancer development and in estimating lung cancer survival.

In [12], Ellen, et al. present an approach to predicting the survival of patients with non-small cell lung cancer (NSCLC) by integrating multiple biological and clinical data types. They developed a model that combined microRNA, mRNA, DNA methylation, long non-coding RNA (lncRNA), and clinical data. They hypothesized that the diversity of these data modalities would capture a holistic view of the patient's condition, allowing for better survival predictions.

In order to process this complex, high-dimensional data, the researchers made use of a denoising autoencoder. Autoencoders are a type of artificial neural network used to learn representations of input data by encoding the data into a low-dimensional format then reconstructing the original data. This process reduces dimensionality while retaining important information. Denoising autoencoders are trained to reconstruct the original data from a corrupted version, making them useful for removing noise from data [13].

This study also explored the effects of early versus late data integration approaches. In early integration, data from multiple modalities is combined into a single input representation before being processed by any machine learning model. In late integration, each data modality is processed separately using different models. The model outputs are then combined, typically right before the final prediction [14]. The study found that early integration consistently outperformed late integration and that models that combined multiple modalities outperformed single modality models.

Another work that leveraged a multimodal deep learning approach to predict the survival of patients with NSCLC is [15]. In this paper, Wu et al. propose a multimodal model named DeepMMSA which integrates CT images and clinical data to improve the accuracy of survival predictions.

DeepMMSA utilizes a 3D convolutional neural network (3D-ResNet) to extract features from the CT images. 3D-ResNets are a deep learning architecture specifically designed for tasks involving three-dimensional data, making them particularly effective for analyzing volumetric data like CT scans [16]. These extracted CT scan features were combined with clinical data which included age, histology (microscopic structure of the tissues), clinical TNM (Tumor size, lymph Node involvement, and the presence of Metastasis) stage, cancer stage, and gender. During this fusion stage, Batch Normalization was applied to adjust the mean and variance of the features since the different modalities had different scales. Batch normalization is a technique used in deep learning that improves neural network training by normalizing the inputs for each layer of the network [17].

The performance of DeepMMSA was evaluated on data from 422 NSCLC patients from The Cancer Imaging Archive (TCIA), the U.S. Cancer Institute's repository for cancer imaging and related information [18]. The experimental results demonstrated DeepMMSA outperformed conventional unimodal methods. Using C-index, a metric that evaluates how well a model can rank individuals according to their predicted risk [19], as a performance metric, it achieved the best performance of all the models tested.

In a recent work in predicting lung cancer risk, Gao et al. present a machine learning approach that integrated CT scans and clinical data elements [20]. The paper introduced a co-learning model trained on data from the National Lung Screening Trial (NLST) dataset and was externally validated using the Vanderbilt Lung Screening Program (VLSP) dataset. Co-leaning models are multimodal models that make use of the transfer of knowledge between modalities to improve performance [21].

The preprocessing of CT images involved techniques from a previous work on nodule detection by Liao et al. [22]. An attention-based multiple instance learning layer was used to aggregate the most informative regions from the CT scans. These features were then combined with clinical data based on the risk factors used in the PLCOm2012 risk model in an early fusion approach. This combined feature set when then used as input into multiple fully connected layers for final prediction.

The paper found that this co-learning model outperformed both a clinical only model and a CT scan only model on both the NLST dataset and the VLSP dataset.

CHAPTER 3

Methodology

3.1 Datasets

This project made use of two datasets: the National Lung Screening Trial (NLST) dataset and the UCLA Lung Cancer Screening dataset. The National Lung Screening Trial was a large, randomized trial that aimed to determine whether screening individuals at high risk of developing lung cancer using CT scans would reduce lung cancer mortality compared to traditional X-rays [23]. It was compiled by the Division of Cancer Prevention (DCP), a part of the National Cancer Institute (NCI). The UCLA Lung Cancer Screening dataset is a dataset compiled by UCLA for use in their lung cancer research efforts. Both datasets include both CT scans and clinical and demographic information for each individual patient.

The tabular portion of the UCLA dataset contains 100 features. These features can be described as falling into one of three categories: clinical, demographic, or geographic. The clinical columns include health information such as whether a patient had ever been diagnosed with certain diseases and information regarding those diseases. The demographic columns include information such as the age and ethnicity of the patient. The geographic columns include information about where the patient lived, such as the rates of poverty and the presence of food insecurity in a patient's area of residence.

The tabular portion of the NLST data set contained 304 columns. It contained much of the same information as the UCLA dataset regarding clinical and demographic information. However, the clinical information regarding the presence of lung cancer was much more detailed. It included several columns detailing the size, shape, and location of any nodules found in a patient. It also included columns detailing a patient's work history, including the potential exposure to any work-related hazards. Unlike the UCLA dataset, the NLST dataset did not include any geographic information.

3.12 CT Scan Selection

Both datasets included multiple CT scans for most patients. Many CT scans were taken the same day, but others were taken years apart. In selecting which scans to use for training, we picked the most recent scan for each patient while taking into consideration the scan parameters.

In CT scans, parameters are settings that define how the CT scan is performed and how the resulting images are processed. These parameters are usually selected according to diagnostic task [24]. For this project, we selected scans based on consistency across three parameters: orientation, convolution kernel, and slice thickness. Orientation is the plane or direction in which the images are taken. Orientation can be axial, that is, from top to bottom, coronal, that is, vertical from front to back, and sagittal, that is, from side to side. Slice thickness refers to the thickness of the tissue layer that each CT image captures. It is measured in millimeters and usually ranges from 0.5mm to 10mm. Convolution kernel is an algorithm applied to the raw data of a scan that can affect the smoothness or sharpness of an image.

For this project we selected scans that were taken using an axial orientation, since that is the type of orientation Sybil required. For convolution kernel we went with scans that were medium or sharp since these are better suited for detecting lung nodules [25]. For thickness, we selected the thinnest available. The majority were 1 mm thick, with a few as thick as 3.2 mm. An overview of the number of scans and how they were used in training is listed in the following figure.

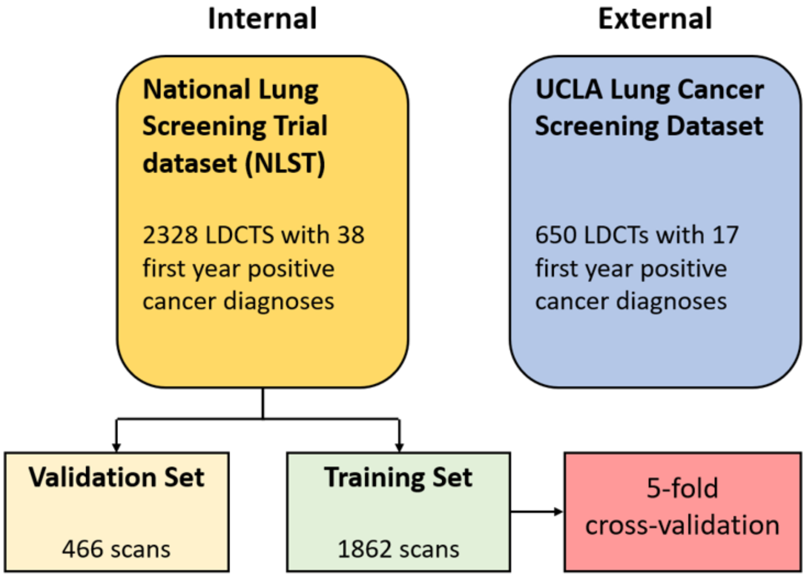

Internal
External
National Lung Screening Trial dataset (NLST)
2328 LDCTS with 38 first year positive cancer diagnoses
UCLA Lung Cancer Screening Dataset
650 LDCTs with 17 first year positive cancer diagnoses
Validation Set
466 scans
Training Set
1862 scans
5-fold cross-validation

Figure 2: Number of scans and positive cases for both datasets

The NLST and UCLA datasets were used as internal and external validation sets, respectively. The NLST dataset consisted of CT scans for 2328 people, 38 of which were diagnosed with lung cancer within one year of a scan being taken. This represents a positive case rate of 1.63%. This dataset was split into a training set and validation set using an 80/20 split. Because the dataset was imbalanced, StratifiedShuffleSplit from the Sci-kit learn was used to perform the split. This ensures that the positive case rate is preserved after splitting. Models were trained and tuned on this training set using 5-fold cross validation, again using StratifiedShuffleSplit to maintain class balance between the folds. After training and tuning, the performance of the best scoring model was evaluated on the validation set. The performance of this best performing model was then evaluated on the UCLA dataset which served as an external validation set. An external validation set is a dataset that is completely independent of the data used to train and validate a machine learning model. It is usually collected from a different source and can be used to assess the generalizability of the model [26].

3.2 Data Preparation

This project made use of two data modalities: clinical and imaging. The differences in these types of data necessitated separate preprocessing steps. The pre-processing steps are listed in the following figure.

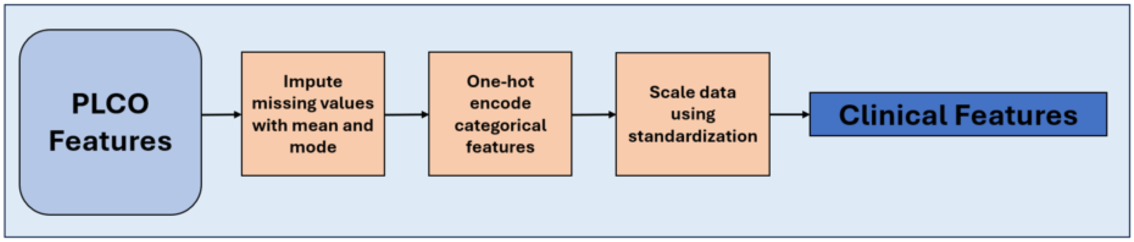
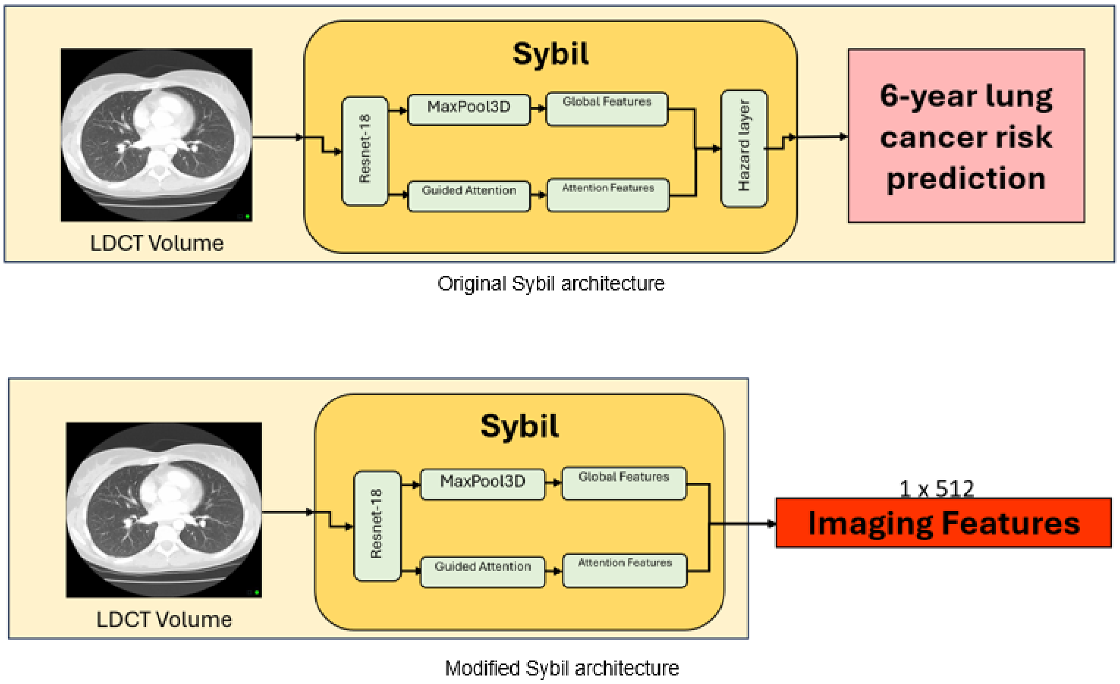

PLCO Features
Impute missing values with mean and mode
One-hot encode categorical features
Scale data using standardization
Clinical Features

Figure 3: Clinical feature pre-processing pipeline

In order to prepare the clinical data for use in model training, we first needed to impute missing data. We did so by filling in the missing data with the mean and mode for each numerical and categorical feature, respectively. Next, we had to encode the categorical data into numerical data. We did so by using one-hot-encoding. Finally, we scaled the data using the standard scalar included in the Sci-kit learn library.



Original Sybil architecture

Modified Sybil architecture

Figure 4: Imaging feature pre-processing pipeline

In order to make the imaging data usable for model training, we made use of Sybil as a feature extractor. In the original architecture, a CT scan input into Sybil is passed through multiple layers that extract and pool the important features of the scan. These extracted features are then fed into a hazard layer for lung cancer risk prediction. As part of this project, Sybil's architecture was modified, removing the final layer. Now instead of outputting a prediction, Sybil would output the extracted imaging features as a 1x512 feature vector. By doing this, Sybil was converted into an imaging feature extractor.

Once feature extraction had been performed, principal component analysis (PCA) was applied to the imaging feature set to reduce its dimension while still preserving as much of the imaging information as possible. Using Sci-kit Learn's built in PCA function, we were able to reduce the dimensionality of the dataset while maintaining 95% of the variance, or original information.

3.3 Model Architecture

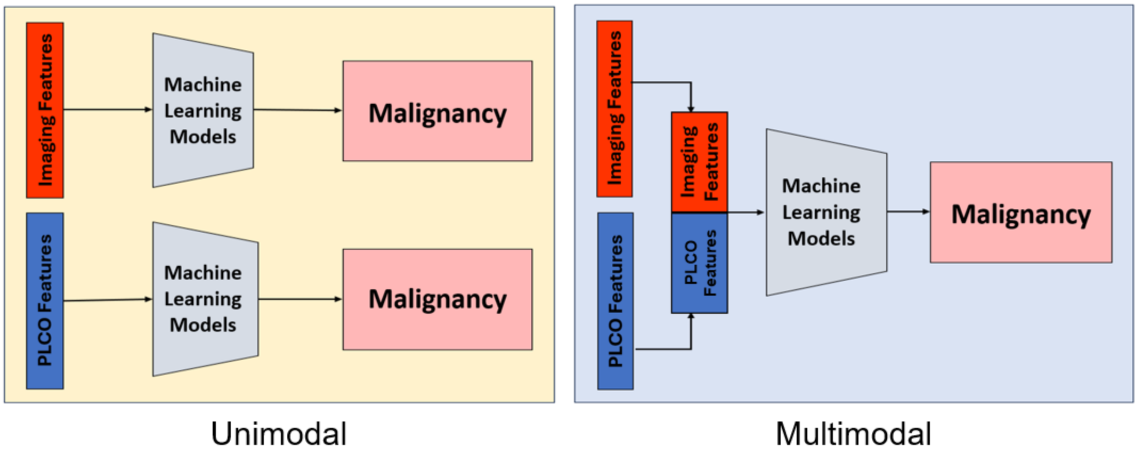


Figure 5: Unimodal and Multimodal architectures

Using the clinical and imaging features, we took two approaches to training the machine learning models: a unimodal approach and a multi-modal approach. In the unimodal approach, machine learning models would be trained on only one type of data, that is only the imaging data, or only the clinical data. In the multimodal approach, the imaging data and the clinical data would be concatenated into a single dataset and machine learning models would be trained on this combined dataset. Four machine learning models were used: logistic regression, random forest, XGBoost, and support vector classifier.

3.4 Machine Learning Models

3.41 Logistic Regression

Logistic Regression is a statistical model that predicts the probability of a binary outcome, relying on one or more predicator variables [27]. It works by transforming the linear combination of input features through the logistic, or sigmoid, function which outputs a real-values number in the range of [0,1]. This number represents a probability. The logistic function is defined as:

$$P(y = 1) = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn)}}$$

where:

- P(y = 1) is the probability of a positive class
- x_1, x_2,…, x_n are the input features
- β_0, β_1,…, β_n are the coefficients of the model

The model is trained by maximizing the likelihood of the observed data.

3.42 Random Forest

Random forest is a learning method that can be used for classification or regression. It works by combining the output of multiple decision trees during training in what is known as a forest. Each decision tree in the forest is trained on a different random subset of the training features in a process known as bootstrap aggregating, or bagging. To make a prediction, each tree produces its own prediction, and the final prediction is made based on a majority vote. The randomness in constructing the individual trees makes random forest resistant to overfitting [28].

3.43 XGBoost

XGBoost or eXtreme Gradient Boosting is a predictive model that implements gradient boosted decision trees. It is similar to random forest in that they are both models built on combining multiple decision trees. The difference is that with XGBoost, every time a decision tree is trained, the weights of the training sample are adjusted based on the training deviation of the model. That is, the weights of correct classification samples are decreased, and the weights of misclassified samples are increased. Each subsequent tree learns from the deviation of the previous tree. An accurate model can then be constructed by combining the decision trees [29].

XGBoost's loss function is given by:

$$L(\theta) = \sum_{i=1}^{n} \mathrm{l}(\mathrm{y_i}, \overline{y_i}) + \sum_{k=1}^{K} \Omega(\mathrm{f_k})$$

where:

- $\mathrm{l}(\mathrm{y_i}, \overline{y_i})$ is the loss function (mean squared error for regression or log loss for classification) between the true label $\mathrm{y_i}$ and the predicted label $\overline{y_i}$
- $\Omega(\mathrm{f_k})$ is the regularization term controlling the complexity of each tree $\mathrm{f_k}$
- K is the number of trees

3.44 Support Vector Classifier

Support Vector Classifier (SVC) is a machine learning model used for classification tasks. It is a variation of the Support Vector Machine (SVM) model that aims to find the optimal decision boundary, known as a hyperplane, that best separates the data points of different classes in a feature space. The goal of training the SVC model is to maximize the margin between the hyperplane and the nearest data points of each class, which are called support vectors [30].

SVC training minimizes the following objective function:

$$\frac{min}{w,b}\frac{1}{2}\ \| w \|^2 + C\sum_{i=1}^{n} max\,(0,1 - y_{\boldsymbol{i}}(w \cdot x_i + b))$$

where:

- w is the weight vector that defines the orientation of the hyperplane
- b is the bias term
- $y_{\boldsymbol{i}}$ represents the true label of the i-th data point
- x_i is the feature vector of the i-th data point
- C is a regularization parameter

3.5 Performance Metrics

This project made use of two metrics to assess the performance of the models: Area Under the Receiver Operating Characteristic Curve (AUROC) and Area under the Precision Recall Curve (AUPRC). A Receiver Operating Characteristic (ROC) curve plots sensitivity against specificity, that is the true positive rate against the false positive rate. A Precision Recall curve plots precision against recall, which is another name for sensitivity. Sensitivity is a measurement of the proportion of positive cases that are correctly identified by the model. It represents how well a model can detect a positive class. Specificity is a measurement of the proportion of negative cases that are correctly identified by the model. It represents how well a model can detect a negative class. Precision measures how many of positive predictions that are correct.

The equations to calculate each are given by the following:

- Sensitivity (recall) = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$
- Precision = $\frac{TP}{TP+FP}$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

AUROC measures a model's ability to distinguish between classes, giving equal weight to both positive and negative classes. A higher AUROC indicates the model is better at distinguishing between classes. A perfect classifier would have an AUROC of 1.0, and a random classifier would have an AUROC of 0.5 [31].

AUPRC measures how well a model can identify the positive class. A model that can accurately identify true positives while minimizing false positives will achieve a higher AUPRC. This makes AUPRC more suited to imbalanced datasets since it describes a model's ability to identify the minority class. The AUPRC of a perfect classifier would be 1.0 and the AUPRC of a random model would be equal to the proportion of positive cases in the dataset [32].

3.6 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that simplifies complex datasets by transforming them into a lower dimensional space while preserving as much information as possible. It works by identifying the directions, or axes, of greatest variance within the data. These axes are called principal components. Once identified, the original data is projected onto these principal components, creating a lower-dimensional representation of the data that retains most of the important information [33].

3.7 Shapley Additive Explanations

Shapley Additive Explanations (SHAP) is a method for explaining individual predictions of machine learning models by calculating the contribution of each feature to the model's prediction. It is based on Shapley values, a concept from game theory which involves assigning each player a value that represents their contribution to the outcome.

In SHAP, each feature is treated as a “player” that contributes to the model’s prediction. The Shapley value for each feature represents how much of that feature contributes to the difference between the model’s actual prediction and the expected prediction [34].

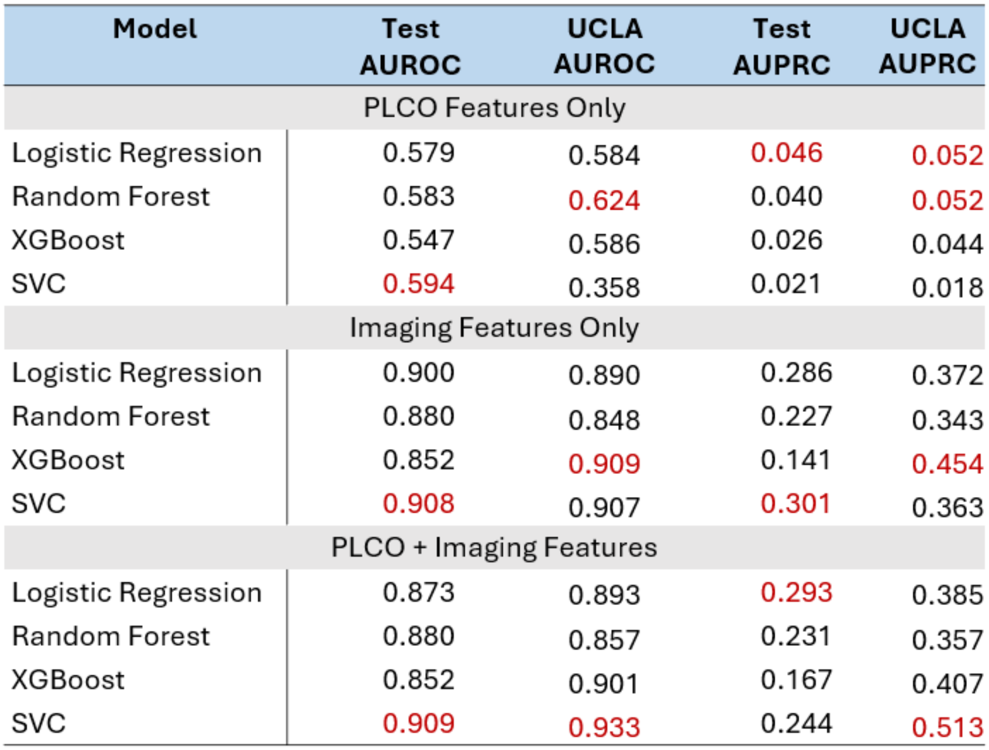
CHAPTER 4

Results

This project trained several models using different sets of features to predict the risk of an individual developing lung cancer within one year. Logistic regression, random forest, XGBoost, and SVC models were trained on the NLST dataset and evaluated on the held-out NLST test set and UCLA dataset using AUROC and AUPRC as performance metrics. The three combinations of features used were:

1. Only PLCO clinical features
2. Only extracted imaging features
3. PLCO clinical and extracted imaging features together

The results of this training are listed in the table below.



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	0.363
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

Table 2: Performance of models trained on different sets of data

For models trained on only PLCO clinical features, performance was poor across all models. The highest scoring models with regards to AUROC were SVC and Random Forest with scores of 0.594 and 0.624 on the NLST test set and UCLA set, respectively. Because an AUROC score of 0.50 is equivalent to a random model, these scores indicate the models had little predictive power. The highest scoring models with regards to AUPRC were Logistic regression and both logistic regression and random forest with scores of 0.046 and 0.052 on the NLST test set and UCLA set, respectively. These scores are little better than the positive rates of 0.016 and 0.026 of the NLST and UCLA sets, further confirming the lack of predictive power of the models.

When trained on only imaging features, all models demonstrated a substantial improvement in both AUROC and AUPRC. The highest scoring models with regards to AUROC were SVC and XGBoost with scores of 0.908 and a comparable 0.909 on the NLST test set and UCLA set, respectively. The highest scoring models with regards to AUPRC were again SVC and XGBoost with scores of 0.301 and 0.454 on the NLST test set and UCLA set, respectively.

With the exception of XGBoost, the models trained on both PLCO and imaging features demonstrated an improvement of most of their scores over the same models trained on only imaging features. Moreover, most of the best performing models were among those trained on both PLCO and imaging features. The highest scoring model with regards to AUROC was SVC scores of 0.909 and 0.933 on the NLST test set and UCLA set, respectively. These scores are the highest AUROC amongst all models and combination of features. The highest scoring models with regards to AUPRC were logistic regression and SVC with scores of 0.293 and 0.513 on the NLST test set and

UCLA set, respectively. SVC's AUPRC was the highest among all models and combination of features on the UCLA set, though logistic regression's AUPRC on the NLST set was lower than the SVC model trained on only imaging features. Because SVC trained on both PLCO and imaging data achieved the best score in three out of the four metrics compared, we consider this the best performing model. For this reason, we look more closely at the SVC models and their performance on the UCLA dataset.

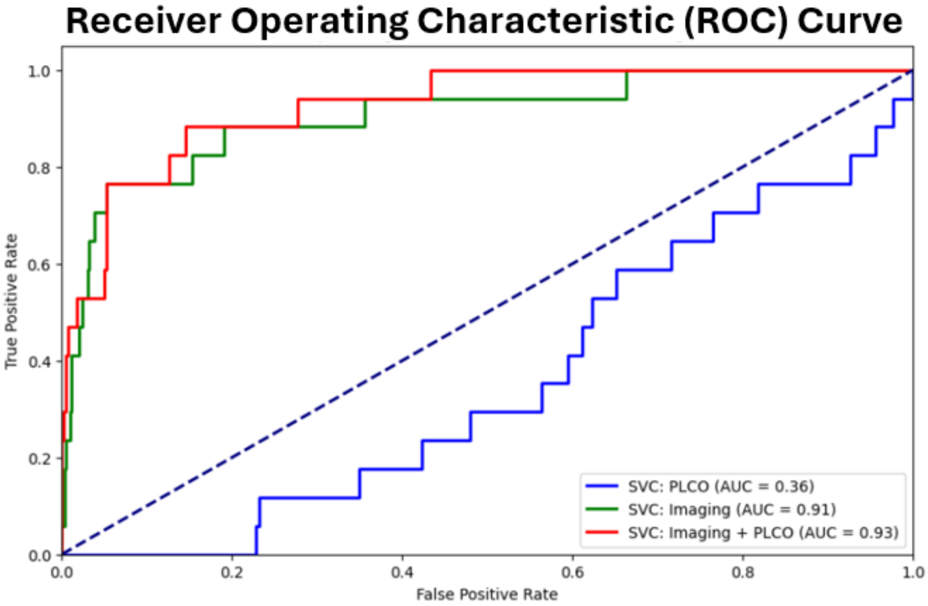


Figure 6: Receiver Operating Characteristic curves for various SVC models

In the above figure we see the Receiver Operating Characteristic (ROC) curve which visualizes the performance for the SVC models trained on different sets of features. Specifically, it represents the models' abilities to distinguish between true positives and false positives. A true positive is an individual that the model predicted would develop lung cancer within one year that did develop lung cancer within one year.

A false positive is an individual that the model predicted would develop lung cancer within one year that did not develop lung cancer within one year.

The blue line represents the model trained only on PLCO features. This model performed poorly, by far worse than the three. Its AUC of 0.36 is worse than the score of a model that produces random guesses. The model trained on only imaging features, represented by the green line, performed significantly better. The models trained on both PLCO and imaging features, represented by the red line, performed the best, slightly beating out the model trained on only imaging features. For this model, the addition of the imaging features seemingly had an additive effect on the performance of the model.

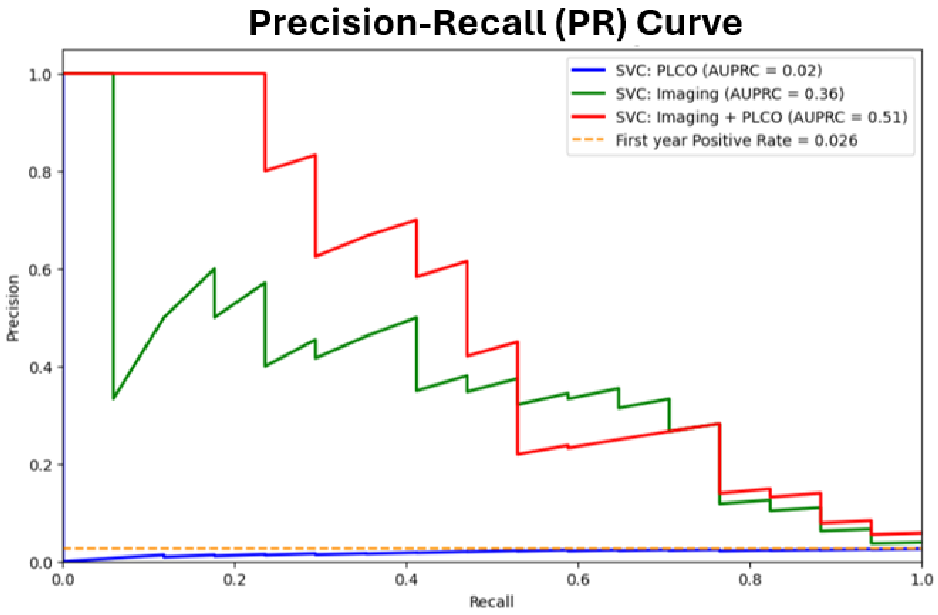


Figure 7: Precision-Recall curves for various SVC models

In the above figure we see the Precision Recall (PR) curve for the three SVC models. The curve plots the model's precision (the ratio of true positives to the total number predicted as positive) against recall (the ratio of true positives to the total number of actual positives). The yellow dotted line represents the first-year positive rate and a baseline performance for the models. We see a pattern similar to that in the ROC curve. The model trained on only the PLCO features, represented by the blue line, performed the worst of all the models, its AUPRC comparable to guessing. The model trained on imaging features, represented by the green line, performed significantly better. The model trained on both the imaging features and the clinical features, represented by the red line, performed the best of the models. Again, it seems the inclusion of the PLCO features has an additive effect on the model performance, albeit a small one. To understand the effects of the individual clinical features on model predictions, we consider the SHAP values.

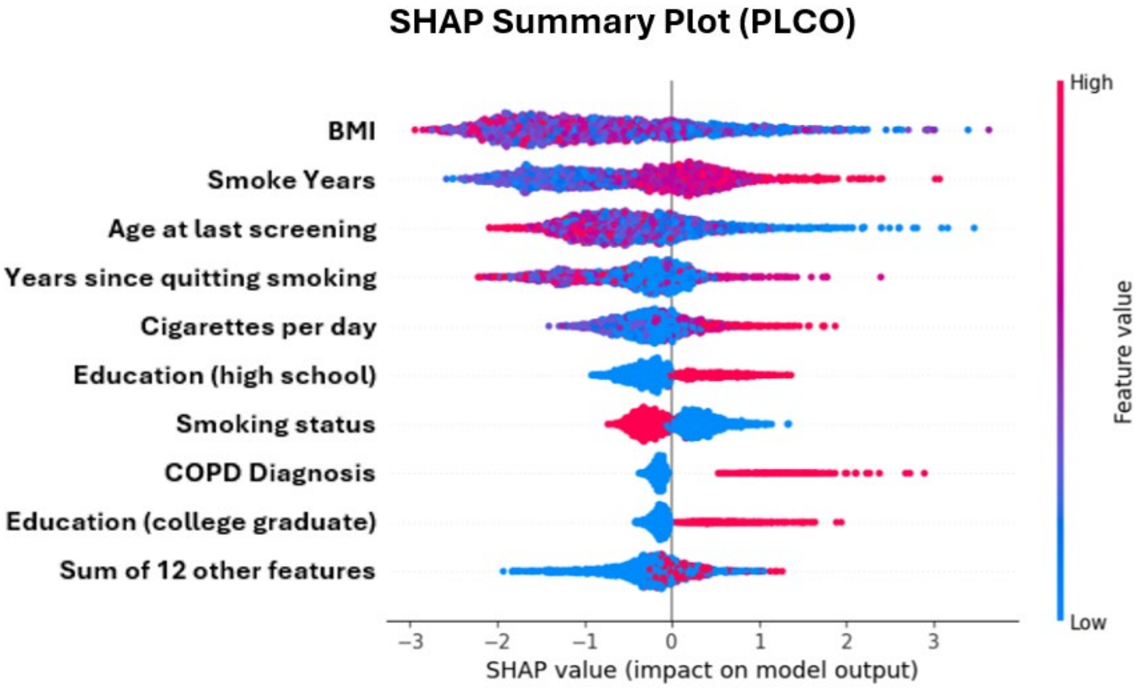


Figure 8: SHAP values for SVC model trained on only PLCO features

The above figure visualizes the importance of different features on the predictions made by the SVC model trained on only the PLCO features, providing insight into how each feature contributes to the model's prediction. Features are ranked by most important to least important, with the most important at the top. Blue coloring represents a low feature value and red a high feature value. The SHAP values on the x-axis show whether a feature increases or decreases risk prediction.

Because BMI is listed on the top of the figure, that means it is the feature most influential on the model's prediction. The blue dots, representing a low BMI value, having a large positive SHAP value indicate that an individual having a low BMI influenced the model to increase the likelihood of predicting lung cancer development. Similarly, the red dots, representing a high BMI value, having negative SHAP values indicate that an individual having a high BMI influenced the model to decrease the likelihood of predicting lung cancer development.

The two next most influential features were the number of years smoked and the age of the individual at their last cancer screening. A high value for the number of years smoked had a positive effect on prediction, and a low value, a negative effect. For the age of the individual at last screening, a low value had a positive effect on prediction and a high value, a negative effect.

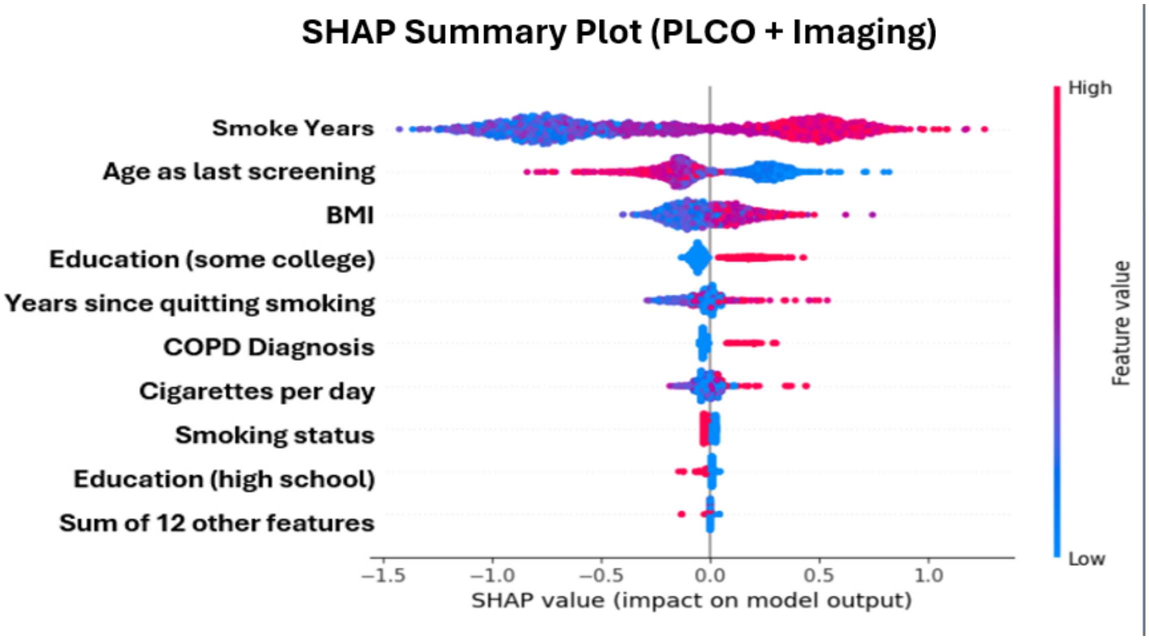


Figure 9: Clinical feature SHAP values for SVC model trained on both Clinical and PLCO features

The above figure visualizes the importance of different features on the predictions made by the SVC model trained on both the PLCO and imaging features. The imaging features were omitted from the chart to better assess which clinical features were the most influential. It should be noted that the imaging features were in general more influential than the clinical features.

Similar to the model trained on only PLCO features, the three most influential clinical features to the model trained on both clinical and imaging features were the number of years smoked, the age of the individual at their last cancer screening, and BMI. Although the three most influential features were the same, their order of influence was different. Moreover, the influence of BMI was inverted in this model. For this model, a high value for BMI had a positive effect on prediction, and a low value, a negative effect, the opposite of the other model.

CHAPTER 5

Discussion

Several models were trained on three sets of features: only PLCO clinical features, only imaging features extracted using Sybil, and both clinical and imaging features together.

The models trained on clinical features alone performed poorly, achieving performances comparable to guessing. This suggests that the clinical features used did not capture enough discriminative information for accurate one-year lung cancer prediction. This could be due to the twelve features used representing too small of a feature set or that the individual features were not informative enough.

The models trained on just imaging features performed significantly better. These results highlight the strength of imaging data, particularly in capturing physical characteristics of lung tissues that may not be evident through clinical data alone.

Models trained on both clinical and imaging features together in general performed slightly better than the models trained on just imaging features. The best performing model was the SVC model trained on both clinical and imaging features, though the improvement was not as substantial as expected. This suggests that while there is some complementary information in the clinical data, its contribution to the model's performance was limited.

Several potential factors may explain these observations. First, the clinical features used in this study may lack the information required to predict early lung cancer risk. Second, the small improvements observed could be a result of the imaging data already possessing much of the relevant information contained in the clinical features.

CHAPTER 6

Conclusion

The goal of this project was to develop multimodal machine learning algorithms that integrated both imaging features and clinical features to enhance one year lung cancer risk prediction. It was hypothesized that by combining a diverse set of features and modalities, the models would be able to provide a more holistic, and therefore accurate, lung cancer risk assessment.

While combining clinical features with imaging features did provide an increase in performance, the modest improvement suggests that the clinical features had limited predictive power in comparison to imaging data. This suggests a benefit in leveraging multiple modalities but also underscores the need for more informative clinical features to achieve more substantial gains in predictive accuracy.

CHAPTER 7

Future Work

There are many possible avenues to extend this work both in experimentation with the current methodology and in exploration of new approaches.

When deciding on which clinical features to use, we based our decision on an existing model, the PLCOm2012 model. Although a well-regarded model, its feature set did not provide a strong basis for training predictive machine learning models. By incorporating a more robust set of clinical features, model performance might be improved. We could consider other lung cancer risk prediction models such as the Spitz model that also considers occupational hazards or the Hoggart model that incorporates genetic factors [35].

In the data pre-processing steps implemented in this project, simple techniques were used to establish a baseline level of performance. Using more advanced imputation and encoding techniques could improve model performance. Imputing missing data using K-Nearest Neighbors or Multivariate Imputation by Chained Equations could be a better approach than simply using the mean and mode. Certain clinical features in the PLCO risk model were recognized as having an ordinal effect on the risk of developing lung cancer. Different ethnicities and different educational attainments contributed differently to the PLCO model's prediction. By implementing ordinal encoding instead of one-hot encoding, these relationships could be better preserved.

A simple approach to combining the data was again implemented to establish a baseline level of performance. Experimenting with more effective methods of combining the clinical and imaging data could lead to the development of better performing algorithms. This could involve more complex architectures such as deep learning models or ensemble models that combine predictions from several models. Exploring various early fusion techniques that would combine the clinical and imaging data before any model training and late fusion techniques that would involve processing the data separately and then combing the at a later stage could provide insights into the most effective ways to merge these different data modalities.

REFERENCES

[1] R. L. Siegel, A. N. Giaquinto, and A. Jemal, “Cancer statistics, 2024,” *CA. Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, Jan. 2024, doi: 10.3322/caac.21820.

[2] M. Mustafa, Ar. J. Azizi, El. IIIzam, A. Nazirah, S. Sharifa, and Sa. Abbas, “Lung Cancer: Risk Factors, Management, And Prognosis,” *IOSR J. Dent. Med. Sci.*, vol. 15, no. 10, pp. 94–101, Oct. 2016, doi: 10.9790/0853-15100494101.

[3] J. Ning *et al.*, “Early diagnosis of lung cancer: which is the optimal choice?,” *Aging*, vol. 13, no. 4, pp. 6214–6227, Feb. 2021, doi: 10.18632/aging.202504.

[4] M. R. Spitz *et al.*, “A Risk Model for Prediction of Lung Cancer,” *JNCI J. Natl. Cancer Inst.*, vol. 99, no. 9, pp. 715–726, May 2007, doi: 10.1093/jnci/djk153.

[5] A. R. Wahab Sait, “Lung Cancer Detection Model Using Deep Learning Technique,” *Appl. Sci.*, vol. 13, no. 22, p. 12510, Nov. 2023, doi: 10.3390/app132212510.

[6] M. C. Tammemägi *et al.*, “Selection Criteria for Lung-Cancer Screening,” *N. Engl. J. Med.*, vol. 368, no. 8, pp. 728–736, Feb. 2013, doi: 10.1056/NEJMoa1211776.

[7] M. Wafa, A. Oussama Khoudeir, I. A. Ruhban, A. M. Saad, M. J. Al-Husseini, and M. M. Gad, “Racial disparities in lung cancer incidence and mortality over the last two decades; a Population-Based Study,” *Ann. Oncol.*, vol. 30, p. vi107, Oct. 2019, doi: 10.1093/annonc/mdz338.074.

[8] J. D. Albano *et al.*, “Cancer Mortality in the United States by Education Level and Race,” *JNCI J. Natl. Cancer Inst.*, vol. 99, no. 18, pp. 1384–1394, Sep. 2007, doi: 10.1093/jnci/djm127.

[9] P. G. Mikhael *et al.*, “Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk From a Single Low-Dose Chest Computed Tomography,” *J. Clin. Oncol.*, vol. 41, no. 12, pp. 2191–2200, Apr. 2023, doi: 10.1200/JCO.22.01345.

[10] C. Wang, “Federated Learning with ResNet-18 for Medical Image Diagnosis,” in *Proceedings of the 2023 8th International Conference on Multimedia Systems and Signal Processing*, Shenzhen China: ACM, May 2023, pp. 34–39. doi: 10.1145/3613917.3613922.

[11] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *Br. J. Cancer*, vol. 89, no. 2, pp. 232–238, Jul. 2003, doi: 10.1038/sj.bjc.6601118.

[12] J. G. Ellen, E. Jacob, N. Nikolaou, and N. Markuzon, “Autoencoder-based multimodal prediction of non-small cell lung cancer survival,” *Sci. Rep.*, vol. 13, no. 1, p. 15761, Sep. 2023, doi: 10.1038/s41598-023-42365-x.

[13] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, “Autoencoders and their applications in machine learning: a survey,” *Artif. Intell. Rev.*, vol. 57, no. 2, p. 28, Feb. 2024, doi: 10.1007/s10462-023-10662-6.

[14] S. Vieira, W. H. L. Pinaya, R. Garcia-Dias, and A. Mechelli, “Multimodal integration,” in *Machine Learning*, Elsevier, 2020, pp. 283–305. doi: 10.1016/B978-0-12-815739-8.00016-X.

[15] Y. Wu, J. Ma, X. Huang, S. H. Ling, and S. W. Su, “DeepMMSA: A Novel Multimodal Deep Learning Method for Non-small Cell Lung Cancer Survival Analysis,” Jun. 12, 2021, *arXiv*: arXiv:2106.06744. Accessed: Oct. 15, 2024. [Online]. Available: http://arxiv.org/abs/2106.06744

[16] K. Hara, H. Kataoka, and Y. Satoh, “Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition,” in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice: IEEE, Oct. 2017, pp. 3154–3160. doi: 10.1109/ICCVW.2017.373.

[17] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, “Understanding Batch Normalization,” Nov. 30, 2018, *arXiv*: arXiv:1806.02375. Accessed: Oct. 15, 2024. [Online]. Available: http://arxiv.org/abs/1806.02375

[18] F. Prior *et al.*, “The public cancer radiology imaging collections of The Cancer Imaging Archive,” *Sci. Data*, vol. 4, no. 1, p. 170124, Sep. 2017, doi: 10.1038/sdata.2017.124.

[19] A. Nowak-Vila, K. Elgui, and G. Robin, “A Statistical Learning Take on the Concordance Index for Survival Analysis,” 2023, *arXiv*. doi: 10.48550/ARXIV.2302.12059.

[20] R. Gao *et al.*, “Cancer Risk Estimation Combining Lung Screening CT with Clinical Data Elements,” *Radiol. Artif. Intell.*, vol. 3, no. 6, p. e210032, Nov. 2021, doi: 10.1148/ryai.2021210032.

[21] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, “Multimodal Co-learning: Challenges, applications with datasets, recent advances and future directions,” *Inf. Fusion*, vol. 81, pp. 203–239, May 2022, doi: 10.1016/j.inffus.2021.12.003.

[22] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, “Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3484–3495, Nov. 2019, doi: 10.1109/TNNLS.2019.2892409.

[23] National Lung Screening Trial Research Team, “The National Lung Screening Trial: Overview and Study Design,” *Radiology*, vol. 258, no. 1, pp. 243–253, Jan. 2011, doi: 10.1148/radiol.10091808.

[24] S. P. Raman, M. Mahesh, R. V. Blasko, and E. K. Fishman, “CT Scan Parameters and Radiation Dose: Practical Advice for Radiologists,” *J. Am. Coll. Radiol.*, vol. 10, no. 11, pp. 840–846, Nov. 2013, doi: 10.1016/j.jacr.2013.05.032.

[25] S. J. Swensen *et al.*, “Lung Cancer Screening with CT: Mayo Clinic Experience,” *Radiology*, vol. 226, no. 3, pp. 756–761, Mar. 2003, doi: 10.1148/radiol.2263020036.

[26] J. M. G. Taylor, D. P. Ankerst, and R. R. Andridge, “Validation of Biomarker-Based Risk Prediction Models,” *Clin. Cancer Res.*, vol. 14, no. 19, pp. 5977–5983, Oct. 2008, doi: 10.1158/1078-0432.CCR-07-4534.

[27] J. C. Stoltzfus, “Logistic Regression: A Brief Primer,” *Acad. Emerg. Med.*, vol. 18, no. 10, pp. 1099–1104, Oct. 2011, doi: 10.1111/j.1553-2712.2011.01185.x.

[28] S. Buschjäger and K. Morik, “There is no Double-Descent in Random Forests,” Nov. 08, 2021, *arXiv*: arXiv:2111.04409. Accessed: Oct. 29, 2024. [Online]. Available: http://arxiv.org/abs/2111.04409

[29] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[30] A. Mammone, M. Turchi, and N. Cristianini, “Support vector machines,” *WIREs Comput. Stat.*, vol. 1, no. 3, pp. 283–289, Nov. 2009, doi: 10.1002/wics.49.

[31] M. A. Kohn and T. B. Newman, “The walking man approach to interpreting the receiver operating characteristic curve and area under the receiver operating characteristic curve,” *J. Clin. Epidemiol.*, vol. 162, pp. 182–186, Oct. 2023, doi: 10.1016/j.jclinepi.2023.07.020.

[32] M. B. A. McDermott, L. H. Hansen, H. Zhang, G. Angelotti, and J. Gallifant, “A Closer Look at AUROC and AUPRC under Class Imbalance,” 2024, *arXiv*. doi: 10.48550/ARXIV.2401.06091.

[33] H. Abdi and L. J. Williams, “Principal component analysis,” *WIREs Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[34] D. Bowen and L. Ungar, “Generalized SHAP: Generating multiple types of explanations in machine learning,” Jun. 15, 2020, *arXiv*: arXiv:2006.07155. Accessed: Oct. 29, 2024. [Online]. Available: http://arxiv.org/abs/2006.07155

[35] A. M. D’Amelio *et al.*, “Comparison of discriminatory power and accuracy of three lung cancer risk models,” *Br. J. Cancer*, vol. 103, no. 3, pp. 423–429, Jul. 2010, doi: 10.1038/sj.bjc.6605759.