Testing Statistical Models to Improve Screening of Lung Cancer

Elliot Burghardt: University of Iowa

Daren Kuwaye: University of Hawai'i at Mānoa

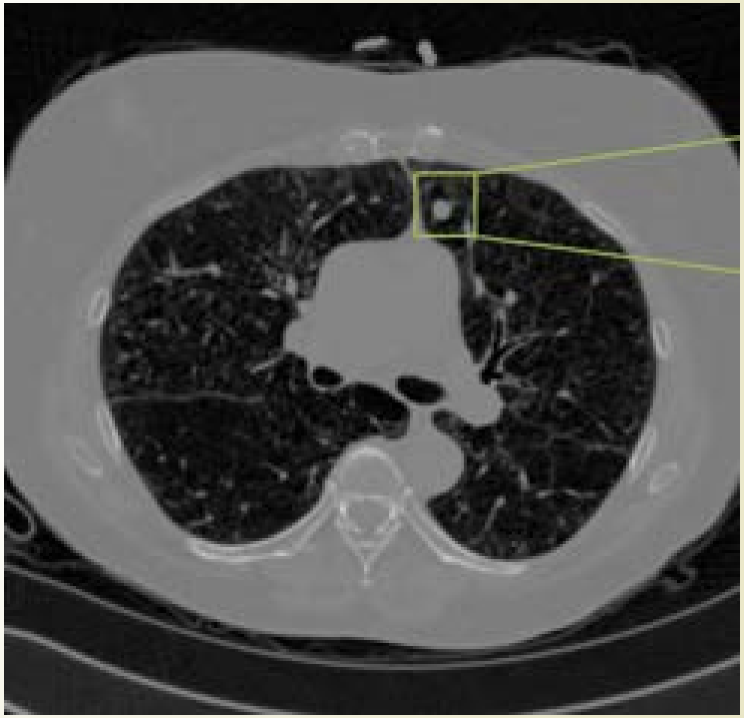
Iowa Summer Institute in Biostatistics - University of Iowa

Department of Biostatistics

Faculty Mentor: **Brian Smith, PhD**, Professor, Dept. of Biostatistics, University of Iowa

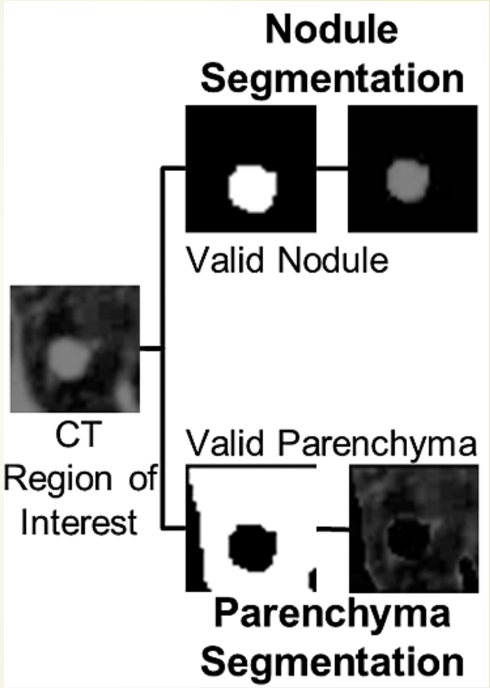
Background

- Over 1 in 4 cancer deaths in the US
- Early-stage detection improves prognosis
- CT Scans
- National Lung Screening Trial (NLST)
  - CT screening detects more early-stage cancers
  - CT Scans have a False Positive Rate of 96.4%
- False positives may require invasive procedures to resolve the diagnosis



Overview – Data Collection

- Radiomic features – quantified characteristics of tumor/nodule
- Process
  - Image segmentation – nodule and parenchyma
  - Feature extraction – summary statistics of the following:
    - Intensity
    - Shape
    - Border
    - Texture



Dilger et al.

Overview – Data Analysis

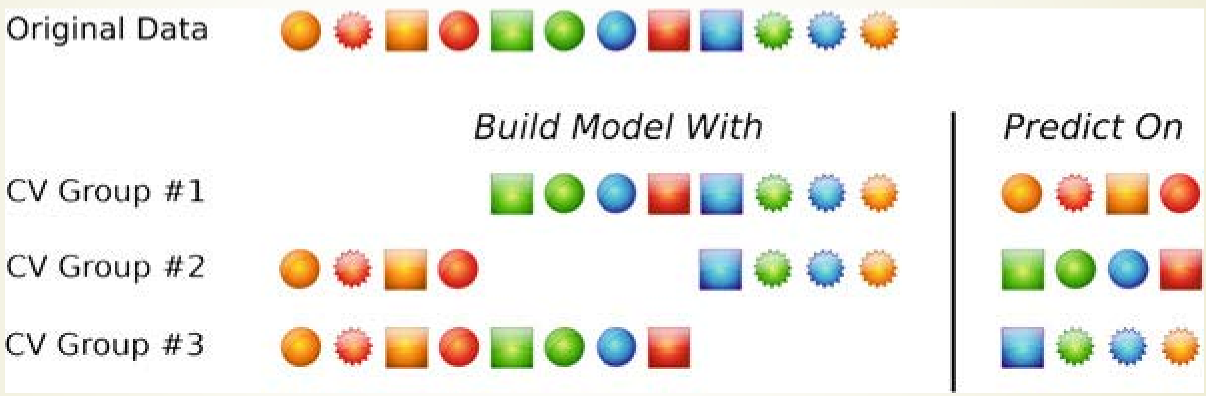
- Goal: Use radiomic features to improve classification of nodule
- Supervised machine learning
    - *Variables*
        - *Input:*144 radiomic variables and 2 clinical variables
        - *Output:* Cancer status - Malignant or Benign
    - 4 models
    - Use Cross Validation to estimate predictive performance
    - Compare the area under the ROC curve for each combination of tuning parameter(s)

Data Summary

Variable		Value
Number of Subjects		198 (100%)
	Benign	89 (44.9%)
	Malignant	109 (55.1%)
Clinical Variables		8
	Age (years)	Mean = 59.93 sd = 13.77
	Pack Years	Mean = 26.39 sd = 29.11
Radiomic Variables		144

Cross Validation (CV)

- Used to estimate predictive performance
- Process (3-Fold CV):

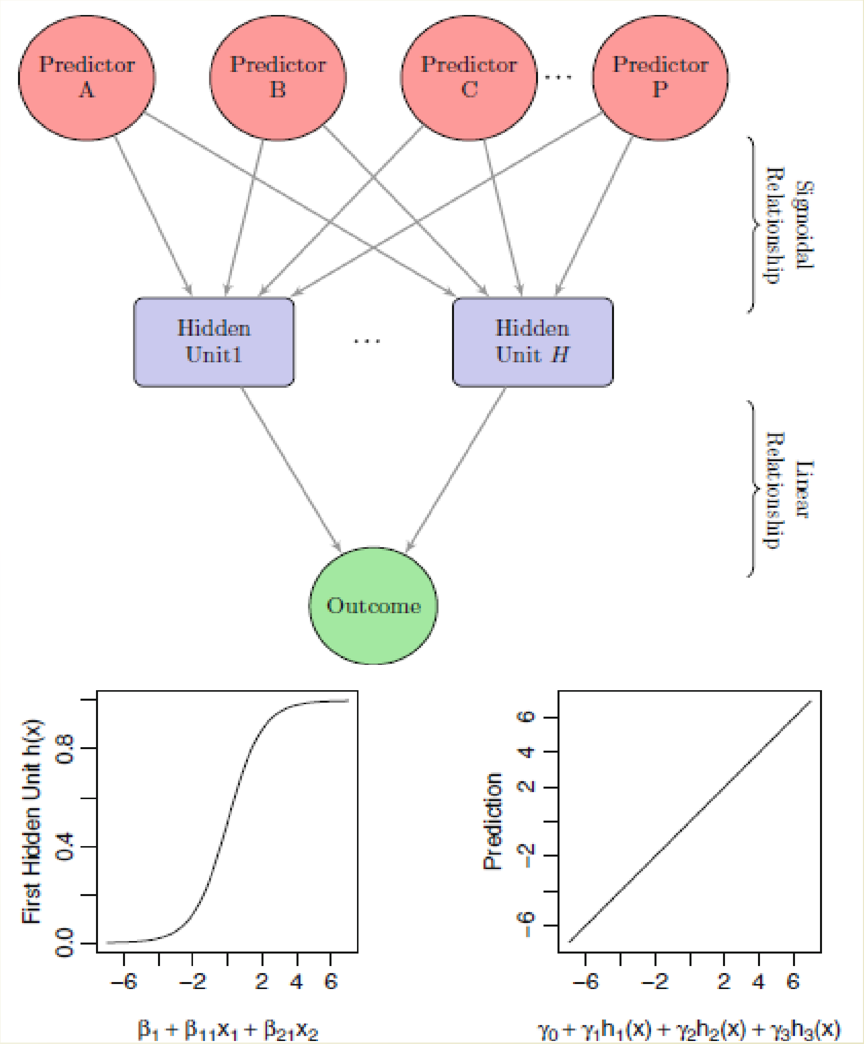


- Protects against "over-fitting" a model
- To improve estimation, we chose to use 10-Fold CV repeated 10 times

Kuhn and Johnson, p. 71

Model 4 - Artificial Neural Network

- Thought of as a "black box" inspired by the brain
- Tuning Parameter: number of hidden units
- Hard to interpret
- ROC = 0.79



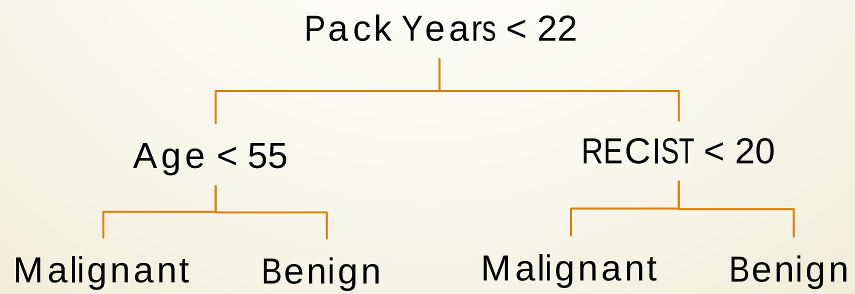
Kuhn and Johnson, p. 142

Model 3– Partial Least Squares

- Linear regression model with fewer variables
  - Orthogonal linear combinations of predictor variables
  - Dimensions are reduced
- Tuning Parameter: number of components
- Hard to interpret
- Continuous outcomes…
- ROC= 0.80

Model 2 – Stochastic Gradient Boosting

- Uses many binary trees
- Final decision based on majority rule
  - (Ties broken at random)
- Variable selection at each node
- Tuning parameters: number of trees, height of tree
- ROC = 0.83



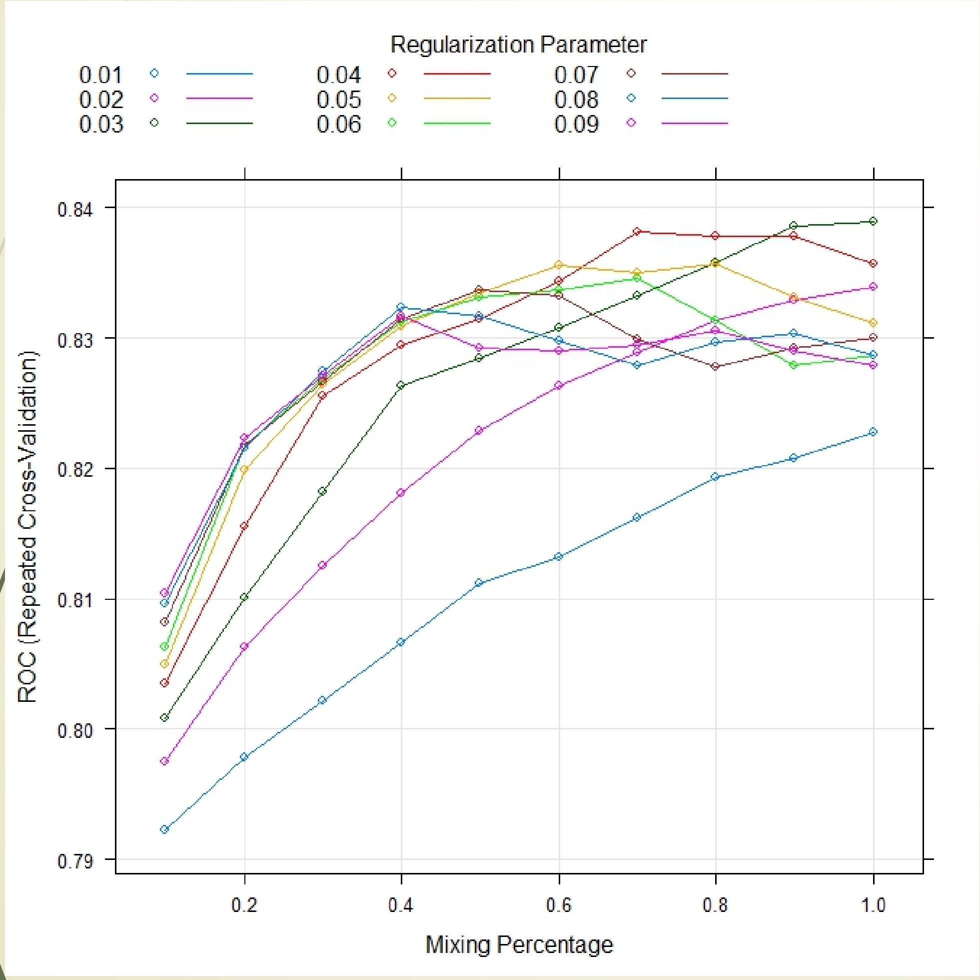
Model 1 – Elastic Net Penalized Logistic Regression

- Binomial model is represented by
  - $\log\frac{\Pr(Diagnosis=1\,|X=x)}{\Pr(Diagnosis=0\,|X=x)} = \beta_0 + \beta^T x$
    - G = {0, 1} where 0 is Benign and 1 is Malignant
    - X is vector of input variables
    - β is vector of coefficients
- Objective function

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\{ -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \left(\beta_0 + x_i^T\beta\right) - \log\left(1 + e^{(\beta_0 + x_i^T\beta)}\right)\right] + \lambda\left[\underbrace{(1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\frac{1}{2}\sum_{j=1}^{p}\left|\beta_j\right|}_{\text{Ridge vs Lasso}}\right]\right\}$$

Variability vs Bias

Elastic Net Penalized Logistic Regression – Optimization



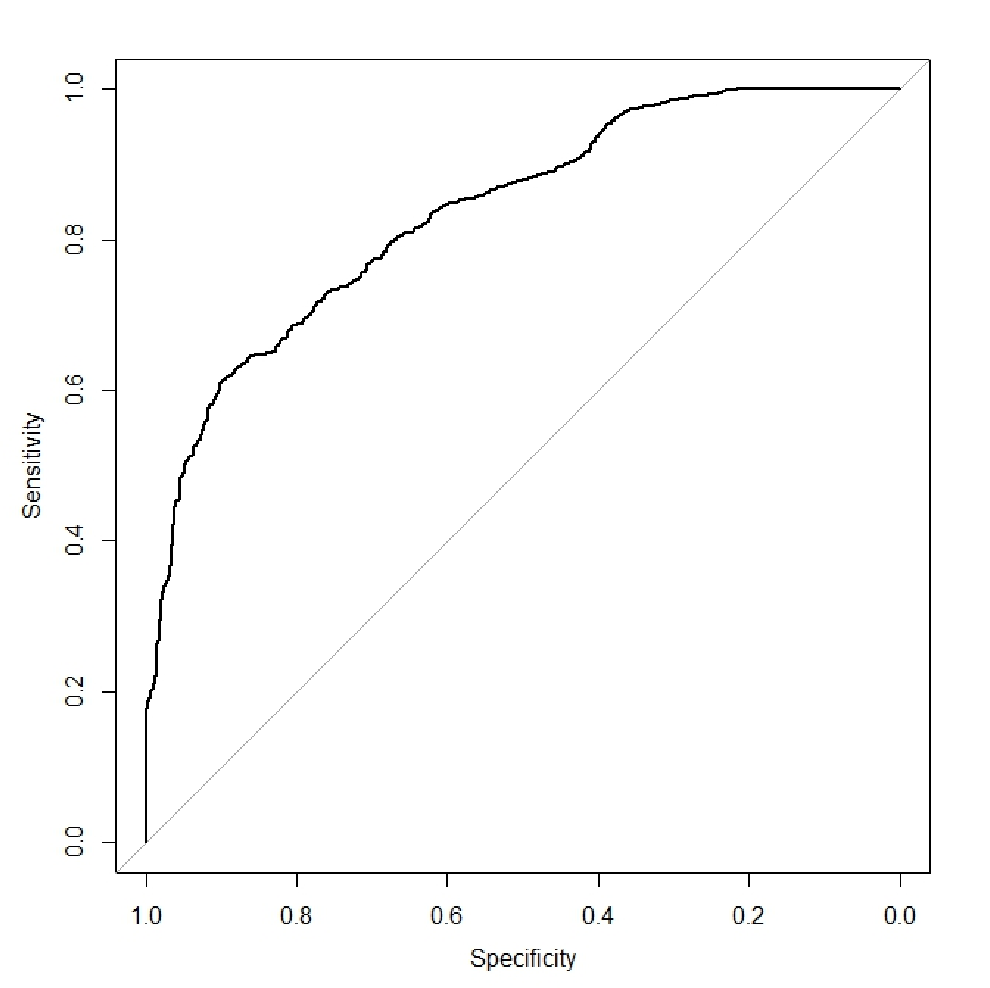
Tuning parameters

- Mixing percentage(α)
- Regularization parameter(λ)

Optimal Performance

- α= 0.94
- λ= 0.03
- ROC = 0.84

Elastic Net Penalized Logistic Regression – Optimization



Tuning parameters

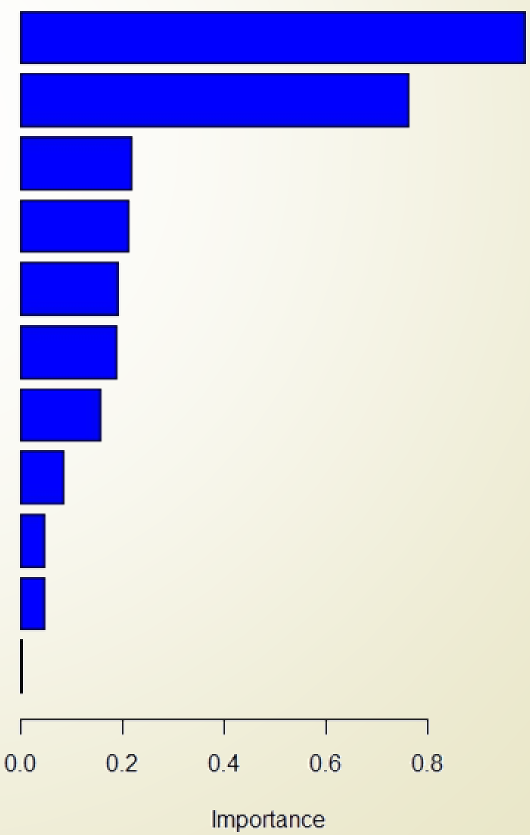
- Mixing percentage(α)
- Regularization parameter(λ)

Optimal Performance

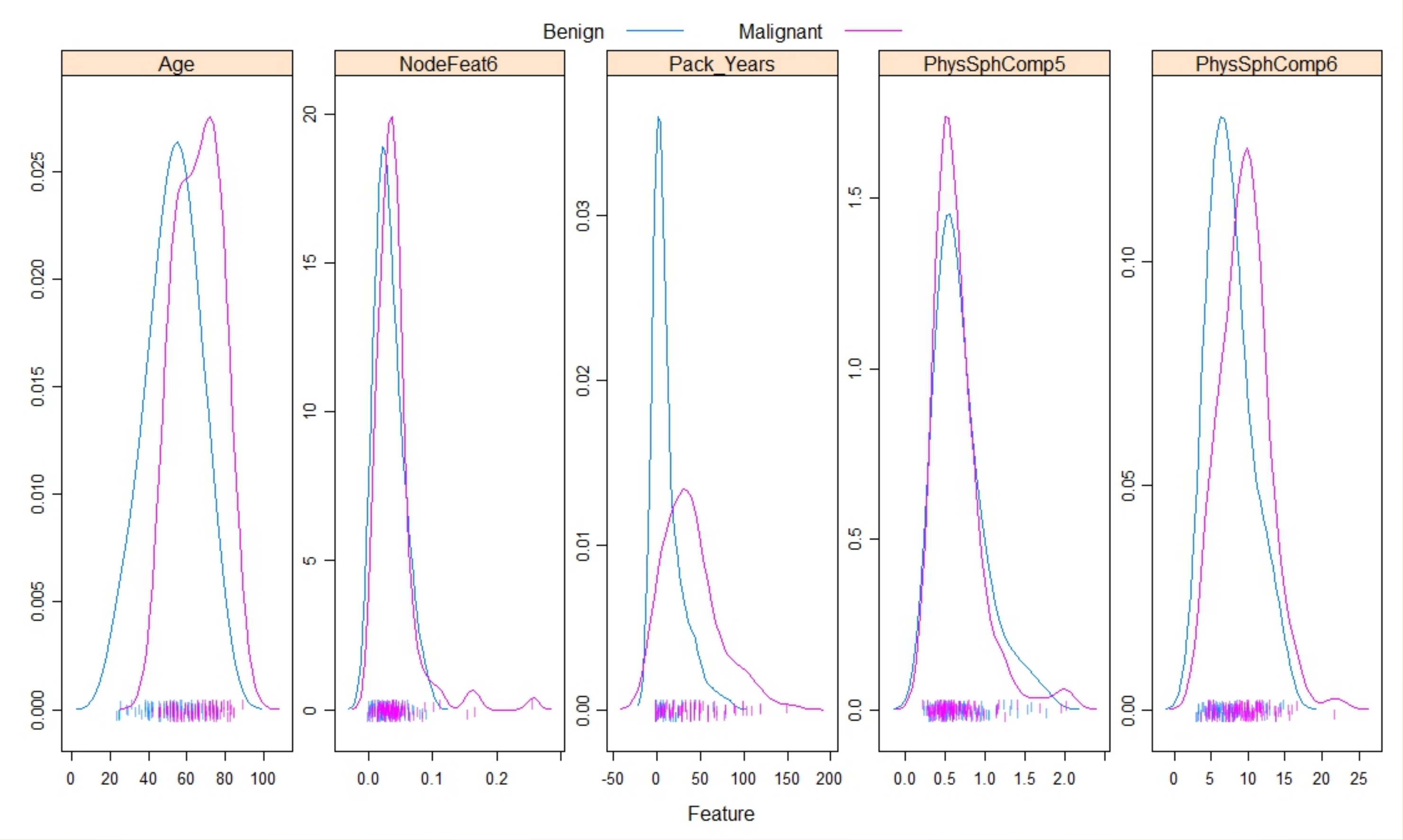
- α= 0.94
- λ= 0.03
- ROC = 0.84

Elastic Net Penalized Logistic Regression – Equation

$$\log \frac{\Pr(Diagnosis = 1 \,|\, X = x)}{\Pr(Diagnosis = 0 \,|\, X = x)}$$
$$= 0.299$$
$$+ 0.993 PackYears$$
$$+ 0.764 Age$$
$$- 0.217 PhysSphComp5$$
$$+ 0.213 NodeFeat6$$
$$+ 0.191 PhysSphComp6$$
$$- 0.189 PhysSphComp3$$
$$+ 0.157 X2DKurtNod3$$
$$+ 0.085 NodeFeat7$$
$$+ 0.048 X2DVarSurrTiss5$$
$$+ 0.002 NodeFeat3$$



Elastic Net Penalized Logistic Regression – Variables



Summary

- Models were based on 146 measurements from 198 subjects at the University of Iowa Hospital
  - Clinical variables had a large impact
  - Both nodule and parenchyma features had an impact
- All of our models had similar performance despite design differences
  - ROC between 0.79 and 0.84
  - Approach from uninterpretable black box to a collection of binary trees to logistic regression
- Elastic net model performance
  - 😃 Reduced false positive rate (23.6%)
  - 😞 At the expense of sensitivity (70.6%)

Future Work

- Set a threshold for false negative then minimize the false positive
- Study the impact of changing the population on the performance of this model
  - Adults aged 55-80 with a history of smoking
  - Multicenter
    - Across US vs. global
    - Beyond academic medical institutions
- Use model to differentiate between types of lung cancer
  - Histology-based
  - Molecular subsets

References

1. Dilger S, Uthoff J, Judisch A, Hammond E, Mott S, Smith B et al. Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. Journal of Medical Imaging. 2015;2(4):041004.
2. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer Science+Business Media; 2013.

Acknowledgement

Instructors

- Brian Smith, PhD
- Gideon Zamba, PhD
- Lauren Sager

Collaborators

- Samantha Dilger, PhD
- Johanna Uthoff
- Alexandra Judisch
- Emily Hammond
- Sarah Mott
- John Newell, Jr.
- Eric Hoffman, PhD
- Jessica Sieren, MD

Funding

- SIB Program sponsored by the National Heart Lung and Blood Institute (NHLBI) HL131467
- MSTP sponsored by the National Institute of General Medical Sciences (NIGMS) 5 T32 GM007337

The University of Iowa
College of Public Health

NIH
National Heart, Lung, and Blood Institute

NIH
National Institute of General Medical Sciences

Questions