Machine Learning algorithms for the detection and localization of surgically resectable cancers

Chris Wilhelm, Jayden Kunwar,
Bryce Thalheimer, Arthur Beyer

Link to original paper

Introduction

- Early cancer diagnosis and treatment can drastically alter the outcome of the disease
- **CancerSeek Algorithm** uses DNA and protein analysis on over 1800 blood samples from healthy individuals and cancer patients
    - Created a logistic regression to predict whether a sample was cancerous using a scored number from DNA analysis and 8 protein levels
- Our goal is to **improve the prediction algorithms** by building our own models
    - We will use an **optimized subset of data** from the dataset (additional protein levels, individual characteristics)
    - We will compare **various classification methods** (logistic regression, neural networks, random forests)

01

DATA

Description of dataset, features, and relevant visualizations

02

METHODS

Model progress: logistic regression, neural networks, and AdaBoost + random forest

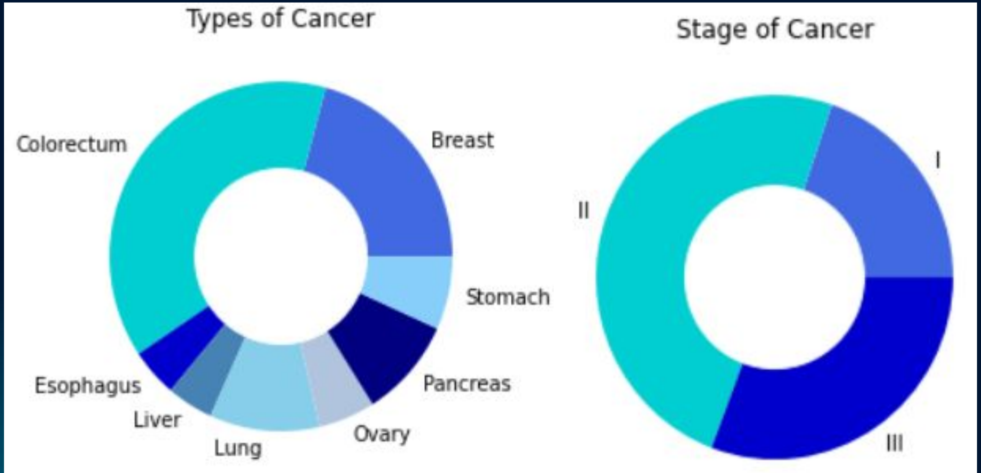
03

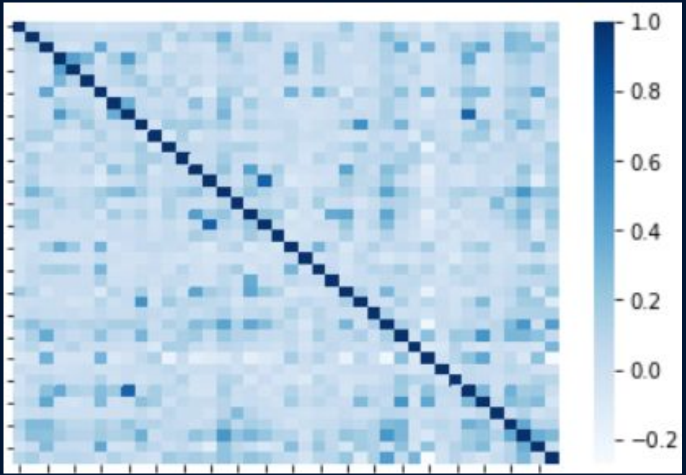
DELIVERABLES

Overall progress and plans for the next project update

Data

- 1005 cancer patients, 812 healthy patients
- **Diverse** data (8 types of cancer, 3 stages of cancer, 41 protein biomarkers)





Data + Datasets

Biomarker Data

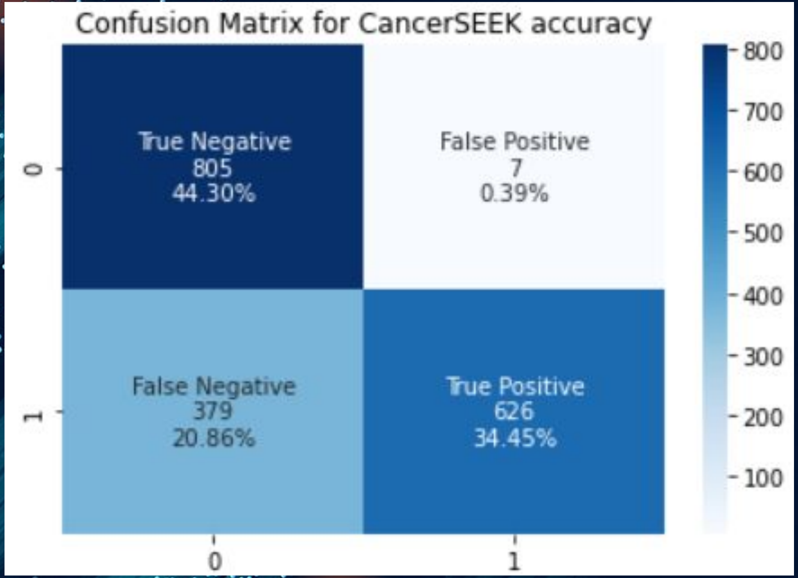
- 41 common protein levels that can be assayed and have known potential link to cancerous phenomena
- Circulating Tumor Mutant DNA Score (score of mutant driver genes from tumors found in plasma)

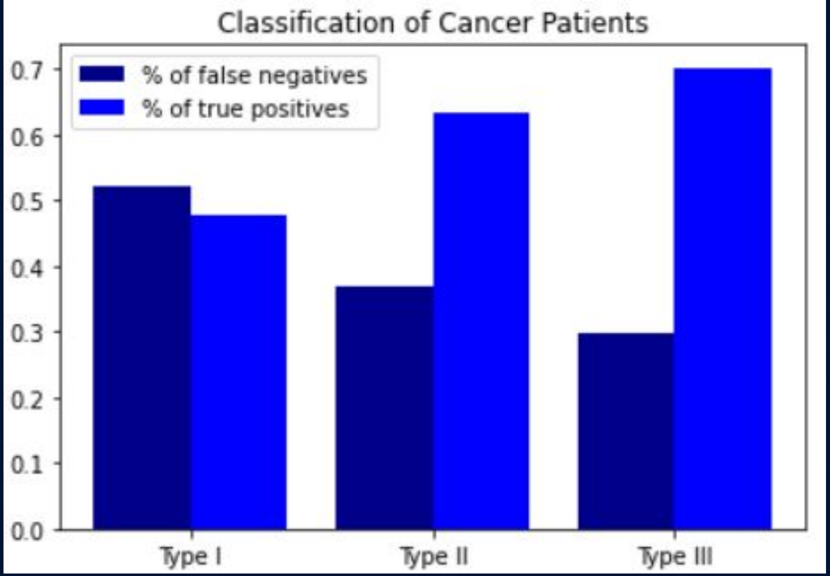
Datasets

- Demographic + Biomarker
  - Age, sex
  - All biomarkers
- Biomarker
  - Just biomarkers
- Replicated Dataset from Paper
  - Subset of 8 proteins and ctDNA score

CancerSEEK Baseline

- CancerSEEK accurately predicts patients **78.8%** of the time
- Specificity of **99.6%** and sensitivity of **79.1%**
- The accuracy rate for CancerSEEK is proportional to the cancer stage

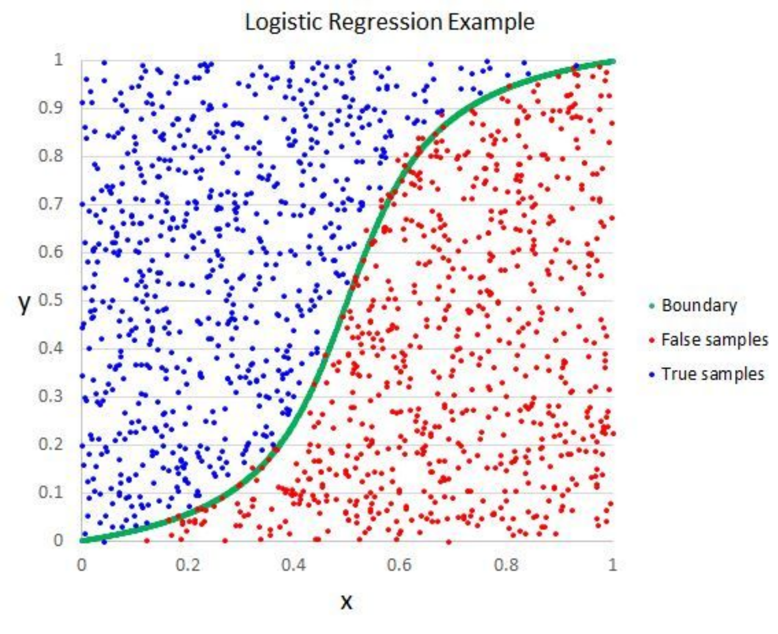




Logistic Regression Model

- Trained with Cross Validation
- Created 3 models:
  - Demographic and Bio Data
  - Bio Data
  - Replicated Data from the original study

- Future Possibilities:
  - Shift the decision boundary
  - Try different kinds of regression (linear)

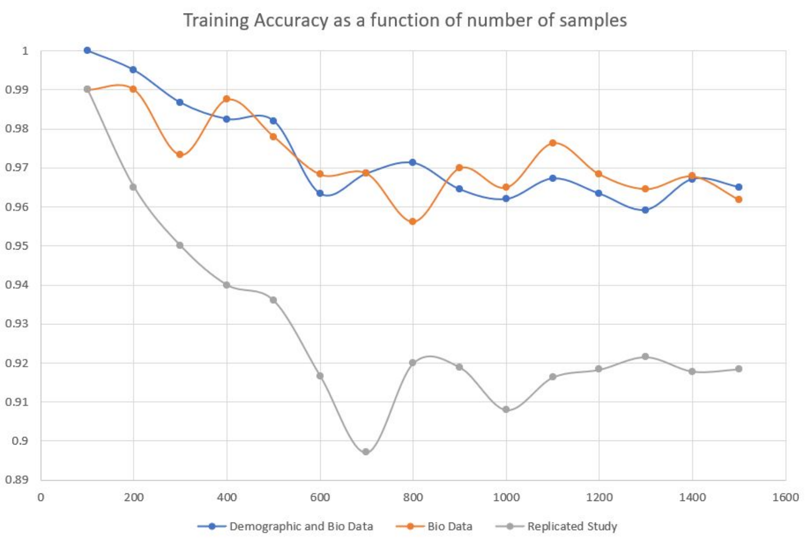


Logistic Regression Results

N = 100 Test Data	Accuracy	Specificity N = 769	Sensitivity N = 948
Demographic and Bio Data	82.9%	84.4%	82.6%
Biological Data	65.1%	31.0%	92.6%
Replicated Study Data	82.9%	84.3%	82.6%

Random Forest Model

- Performed **Cross Validation** to find the best model
  - Grid Cross Validation to tune the hyper parameters with our Train and Dev sets
- To be further explored:
  - Investigating the decision path for interpretability
  - Analyzing the parameters to see which aspects impact the decisions
  - Improving specificity for RF model



Random Forest Test Accuracy

N = 100 Test Data	Accuracy	Specificity N = 769	Sensitivity N = 948
Demographic and Bio Data	96.4%	94.8%	97.5%
Biological Data	96.3%	95.2%	97.2%
Replicated Study Data	91.5%	92.8%	90.4%

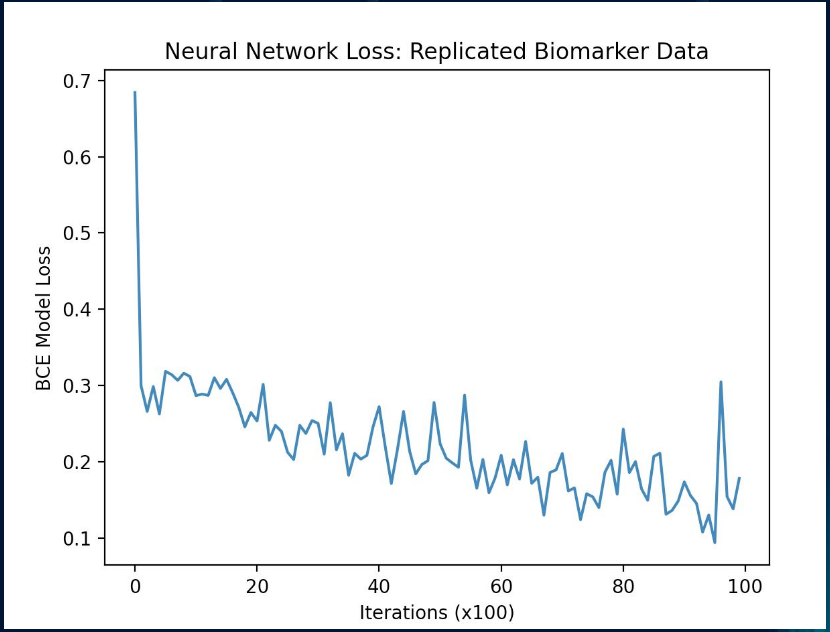
Neural Network Model

Model Structure

- 2-Hidden Layer NN with Dropout
  - ReLu and PReLu Activation
- Hidden Layer Widths = 100
- LR = .005
- Dropout Rate = 0.2
- Epochs = 4000

Next Steps

- Further hyperparameter tuning (structure, rates, etc.)



Neural Network

N = 600 Test Data	Accuracy	Specificity (n = 274)	Sensitivity (n = 326)
Demographic and Bio Data	88.5%	99.3%	79.4%
Bio Data	86.0%	98.2%	75.8&
Replicated Study Data	75.7%	97.8%	57.1%

Model Evaluation

- Logistic Regression
  - Baseline
  - Worst accuracy, worst specificity
- Random Forest
  - Best performing model
  - High accuracy (96%), specificity O.K. (94%) but lower than clinical-grade specificity requirements for diagnostics
- Neural Networks
  - Good performance overall
  - Very high specificity, can improve on sensitivity

Next Steps

- Continue tuning models
  - Change size of neural network
  - Tune hyper-parameters (e.g. shift logistic regression)
- Potentially try linear regression
- Interpretability analysis
  - How did our models go wrong?
  - How can they be improved?

Update on Deliverables

- Must accomplish
  - Logistic Regression ✔
  - Neural Network ✔
  - AdaBoost, Random Forests ✔
- Expect to accomplish
  - Advanced Data Pre-Processing ✔
  - Interpretability Analysis
  - Multi-Class Classification of Cancer Type
- Would like to accomplish
  - High Specificity ✔
  - Additional Data
  - Cancer Localization

Feedback?