Detecting Cancer From Blood Tests

Team: Protein Profiles

Parinaz Fathi, Simeiyun (May) Liu, Nihan Akis-Man, Cerise Chen

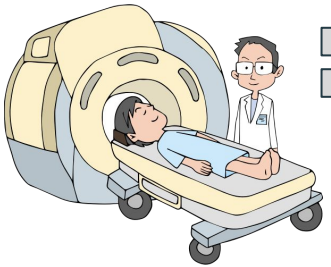
Erdos Data Science Bootcamp, Fall 2024

Introduction

Cancer: one of the leading causes of death worldwide

Early detection → better outcomes

Crosby et al., Science, 2022, DOI 10.1126/science.aay9040



= $ $ $

Diagnosis can be costly and may not happen until symptoms are detected

Cancer alters the levels of some proteins in blood

Our goal: detect cancer based on blood test results

Our Datasets

Pancancer Dataset
(n = 1477)
1463 proteins

Acute myeloid leukemia (AML) (n = 50),
Chronic lymphocytic leukemia (CLL) (n = 48)
Diffuse large B-cell lymphoma (DLBCL) (n = 55)
Myeloma (n = 38)
Blood cancer (n = 191)
Colorectal cancer (n = 221)
Lung cancer (n = 268)
Glioma (n = 145)
Breast cancer (n = 152)
Cervical cancer (n = 102)
Endometrial cancer (n = 101)
Ovarian cancer (n = 134)
Prostate cancer (n = 163)

Alvez et al., 2023, Nature Communications

Esophageal Cancer
(n = 91)
92 proteins

Gao et al., 2024, Journal for Immunotherapy of Cancer

Hodgkins Lymphoma
(n = 54)
92 proteins

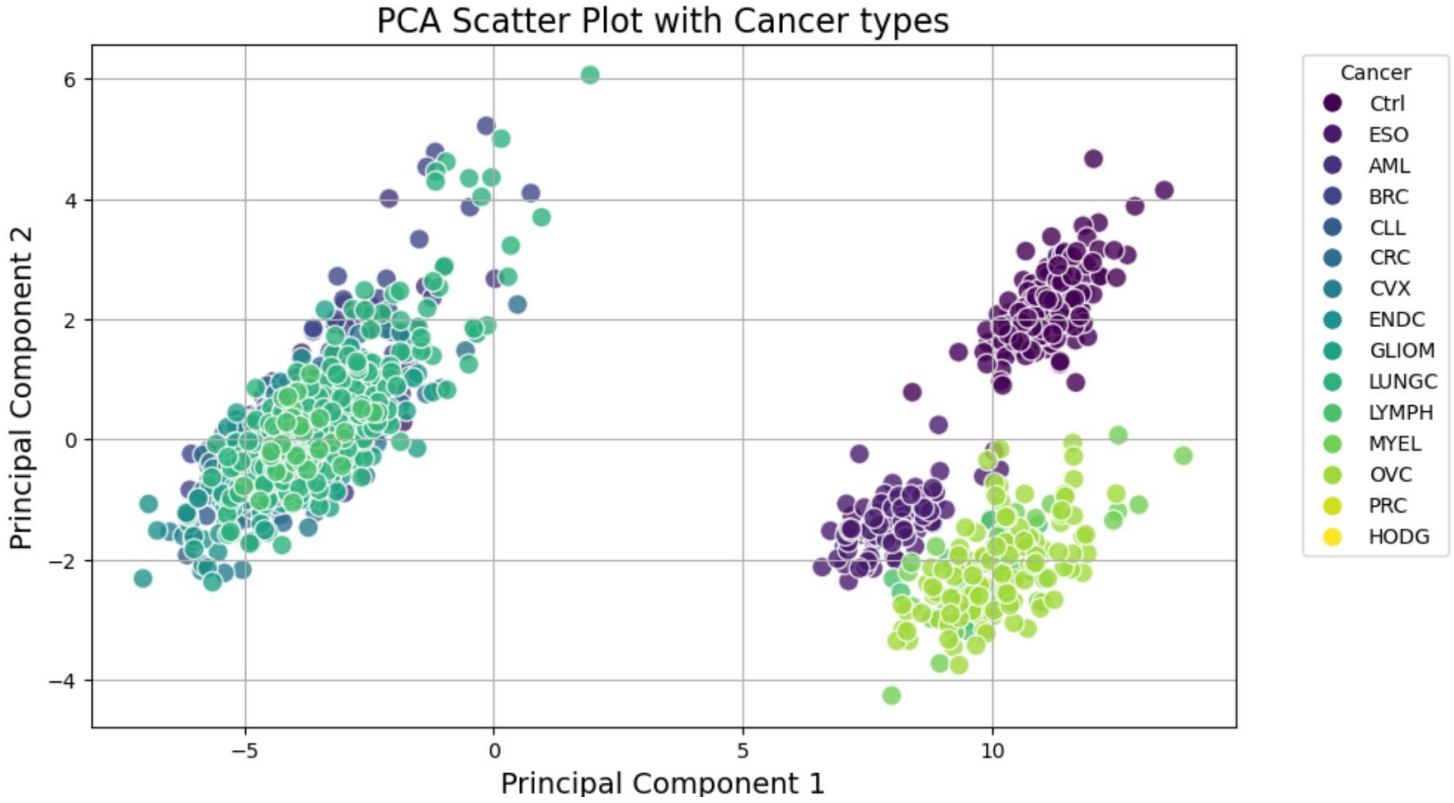
Gonzalez-Kozlova et al., 2024,
Cancer Research Communications

Southern German Population-Based Cohort
(n = 170)
728 proteins

Petrera et al., 2020, Journal of Proteome Research

49 proteins in common across all of these data sets

Exploratory Data Analysis



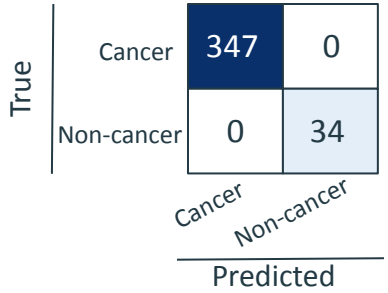
Objective 1: Can we tell whether someone has cancer?

Cancer
(n = 1622)

Not Cancer
(n = 170)

This is a binary classification problem with class imbalance

Train-test split (test = 20 %) with stratification + 5-fold cross-validation



Model	Accuracy	ROC AUC	F-1 Score
Logistic Regression	1	1	1
k-Nearest Neighbors (k=5)	1	1	1

Testing Results
Accuracy: 1
ROC AUC: 1
F-1 Score: 1

Objective 2: If we know someone has cancer, can we determine what type?

Cancer
(n = 1477)

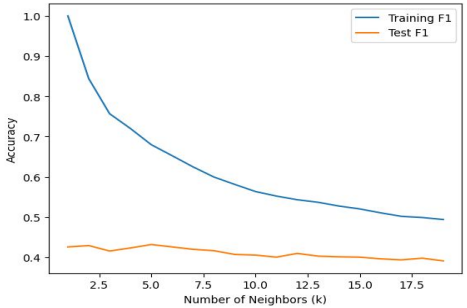
What type
out of 9?

This is a multiclass classification problem with class imbalance

Train-test split (test = 20 %) with stratification + 5-fold cross-validation + penalty

Model	Accuracy	ROC AUC	F-1 Score
Logistic Regression	0.7663	0.9571	0.7498
k-Nearest Neighbors(k = 17)	0.3937	0.8184	0.4030
Random Forest	0.6528	0.9189	0.6536
Extra Trees	0.6096	0.9029	0.6119
XGBoost	0.7138	0.9471	0.7123
Multinomial Regression	0.5932	0.9122	0.5920

Testing Results
Accuracy: 0.7365
ROC AUC: 0.9534
F-1 Score: 0.7194



Objective 3: How can we use the fewest proteins possible to distinguish between different cancer types

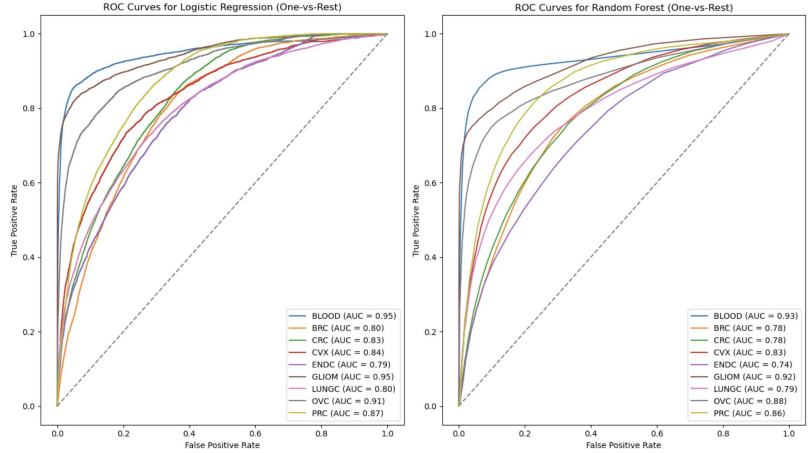
Cancer
(n = 1477)

What type out of 9 with minimal # of proteins?

Train-test split (test = 20 %) with stratification + 5-fold cross-validation + SelectKBest to choose the top 200 proteins + penalty

	Training			Testing		
	Accuracy	ROC AUC	F-1 Score	Accuracy	ROC AUC	F-1 Score
1463 Proteins	0.7654	0.9625	0.7662	0.7567	0.9596	0.7519
200 KBest Proteins	0.7197	0.9461	0.7208	0.7331	0.9486	0.7324

ROC AUC (One-vs-Rest) - Logistic Regression: 0.858
ROC AUC (One-vs-Rest) - Random Forest: 0.835



Conclusions and Future Work

Cancer
(n = 1622)

Not Cancer
(n = 170)

49 proteins
Logistic Regression
100 % Accuracy

Cancer
(n = 1477)
What type
out of 9?

1463 proteins
Logistic Regression
ROC AUC = 0.9534

200 K-Best proteins
Logistic Regression
ROC AUC = 0.9486

Future Work: Validate with larger datasets and other cancer types

Acknowledgements: Dohoon Kim (Team Mentor), Erdos Instructor and Advisor Steven and Alec