No Data, Sum Problem

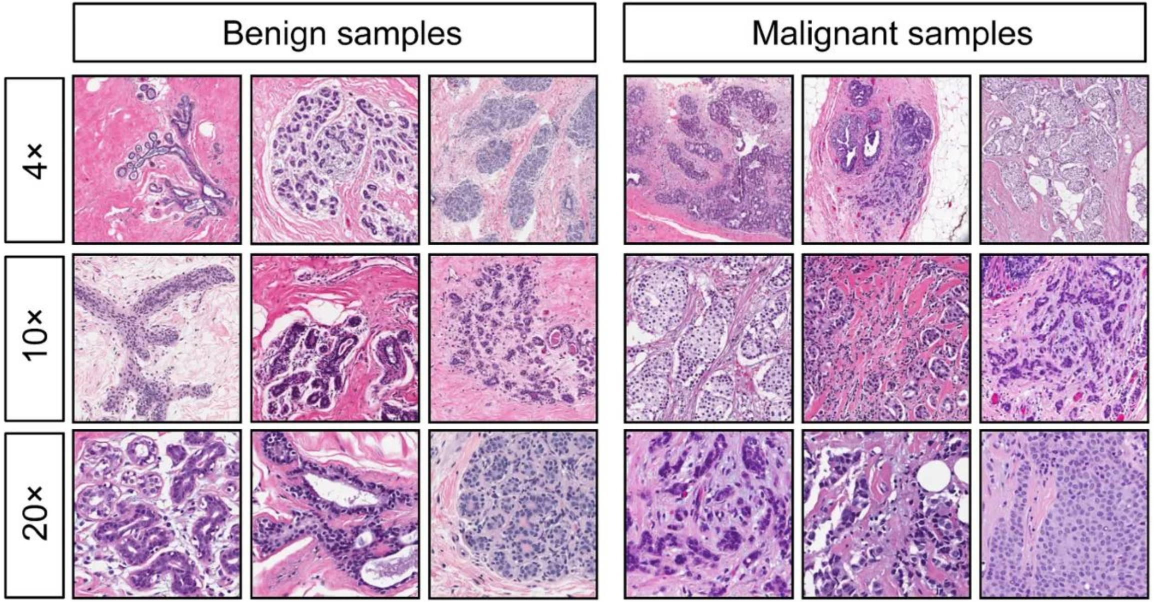
“Breast Cancer Classification Edition”

KGML 2025
University of Michigan

The problem statement

- Lifetime chance of being diagnosed with breast cancer for females is 1 in 8[1]
- Important to catch early (at localized stage), 99% survival rate if caught before spread[1]
- Recommendation is screening above age 40[2], meaning there are many scans of the population with a small proportion malignant (cancer detection rate is 5.1 per 1000 scans[3])
- **Need an effective way to correctly detect malignant tumors from data**

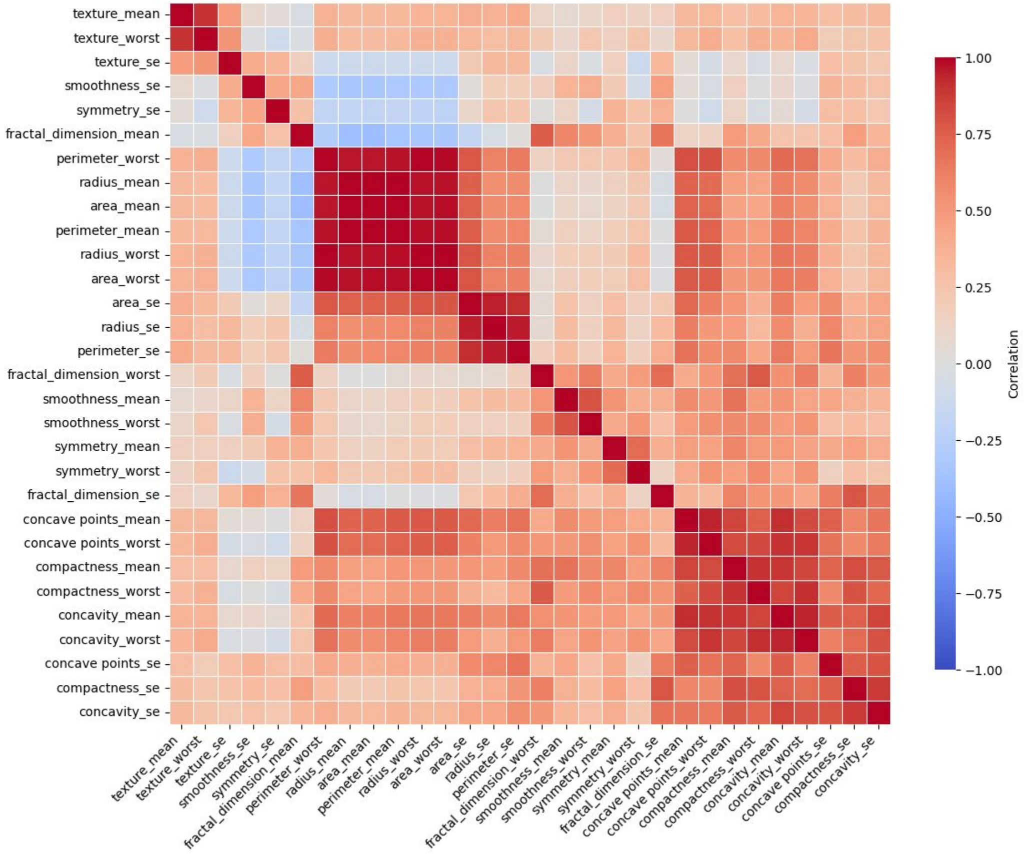
The Data Set



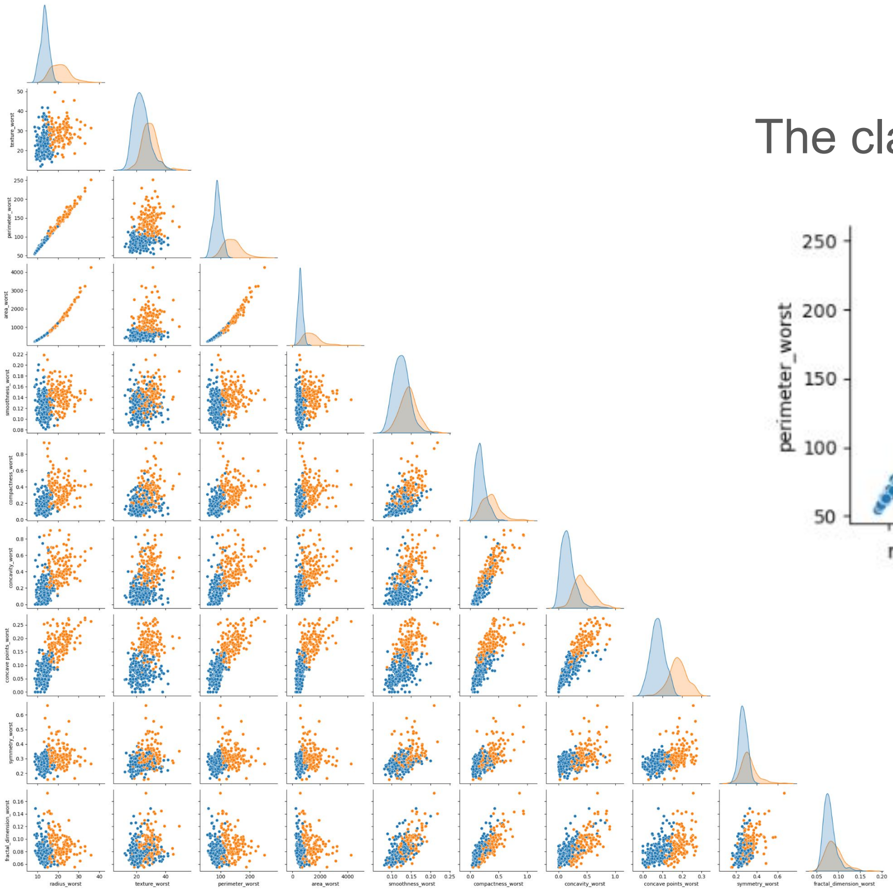
Measurements of cell properties from biopsy (Fine Needle Aspiration)

Features include mean, standard error on mean and worst case measurements.

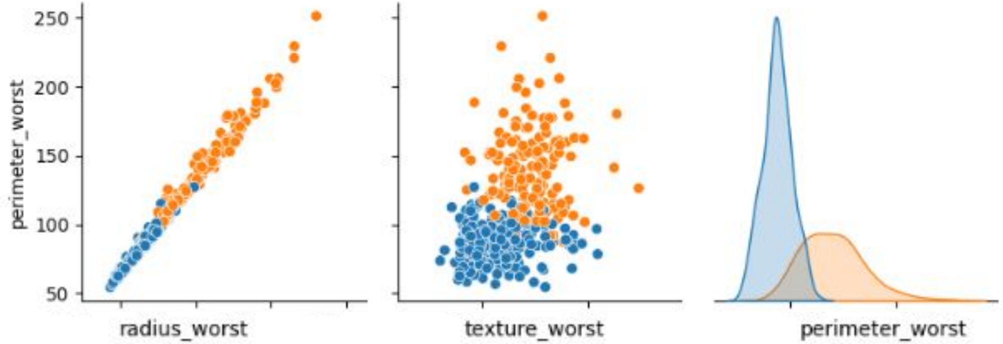
Many Correlated Features



Pair-plots for “worst” Measurements



The classes separate with only **one or two** features!



Cross Validated, Optimized Linear SVM

We trained a 10-fold cross-validated, hyperparameter optimized linear SVM:

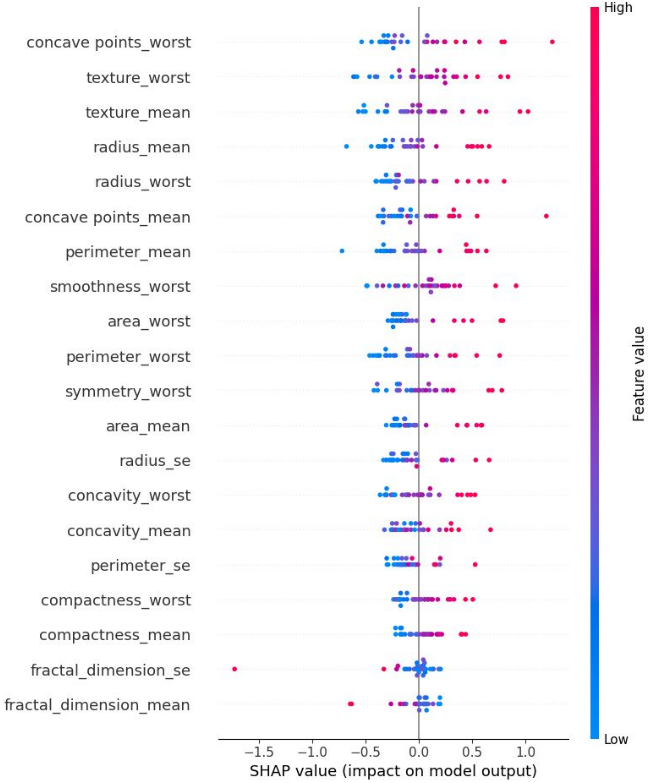
Test Accuracy: **97.59%** +- 3.07 (95-CI)

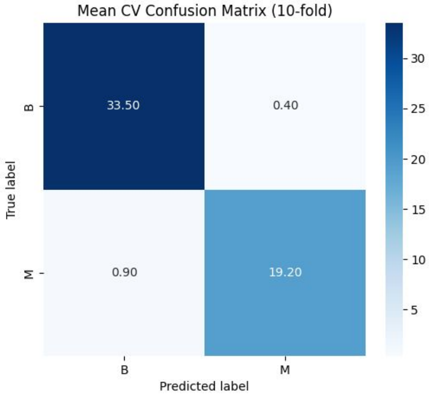
Classification report (across CV folds):

	Precision	Recall	F1
B	0.97	0.99	0.98
M	0.98	0.96	0.97

Feature importance was assessed by computing SHAP values.

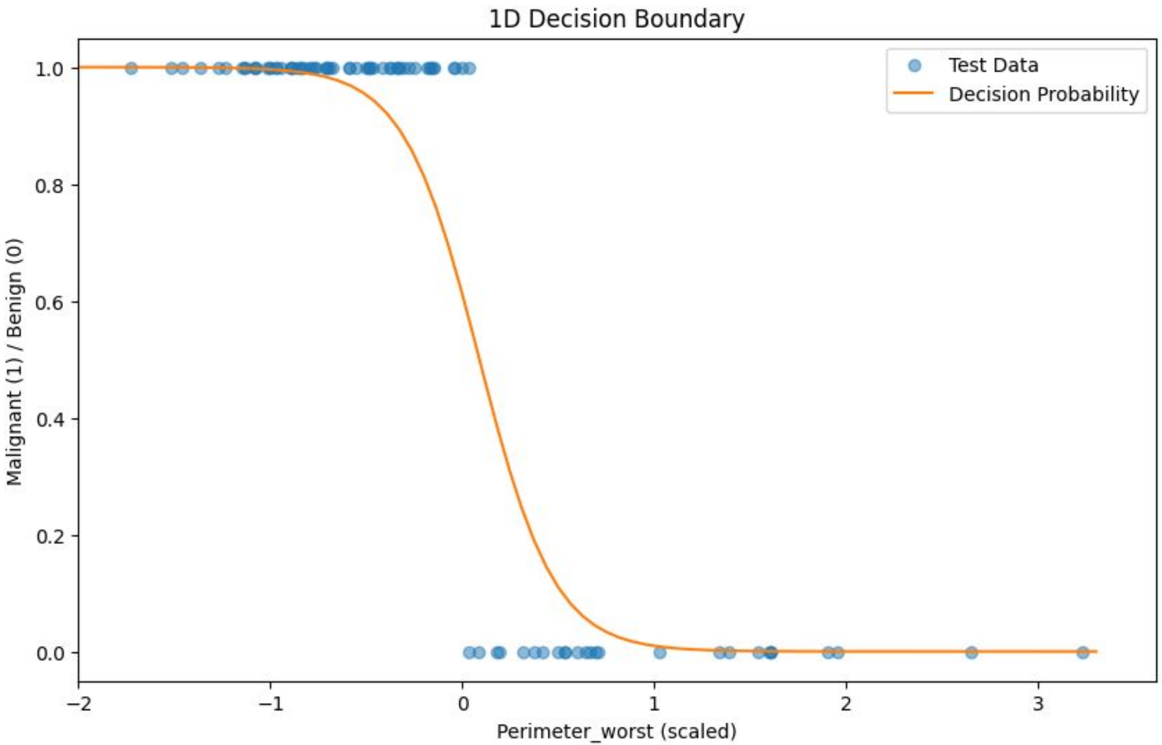
The top features were correlated.



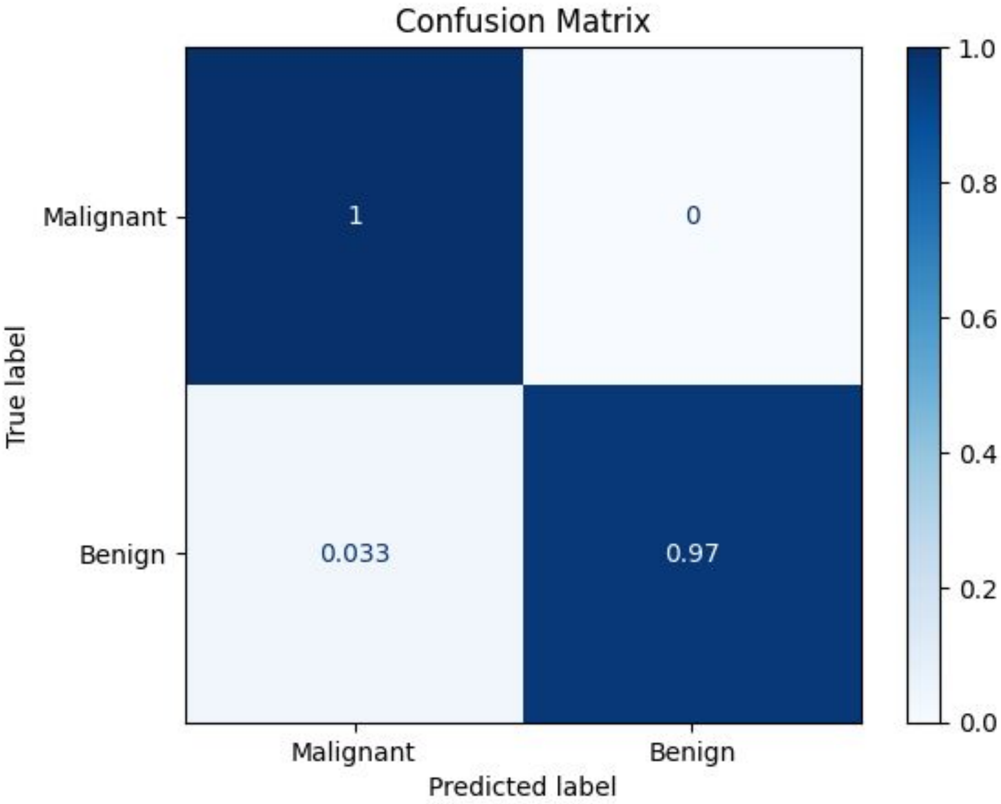


Linear Model with Logistic Regression:

Class Prob = sigmoid(-5.1 x perimeter_worst + 0.45)



Linear Model with Logistic Regression:



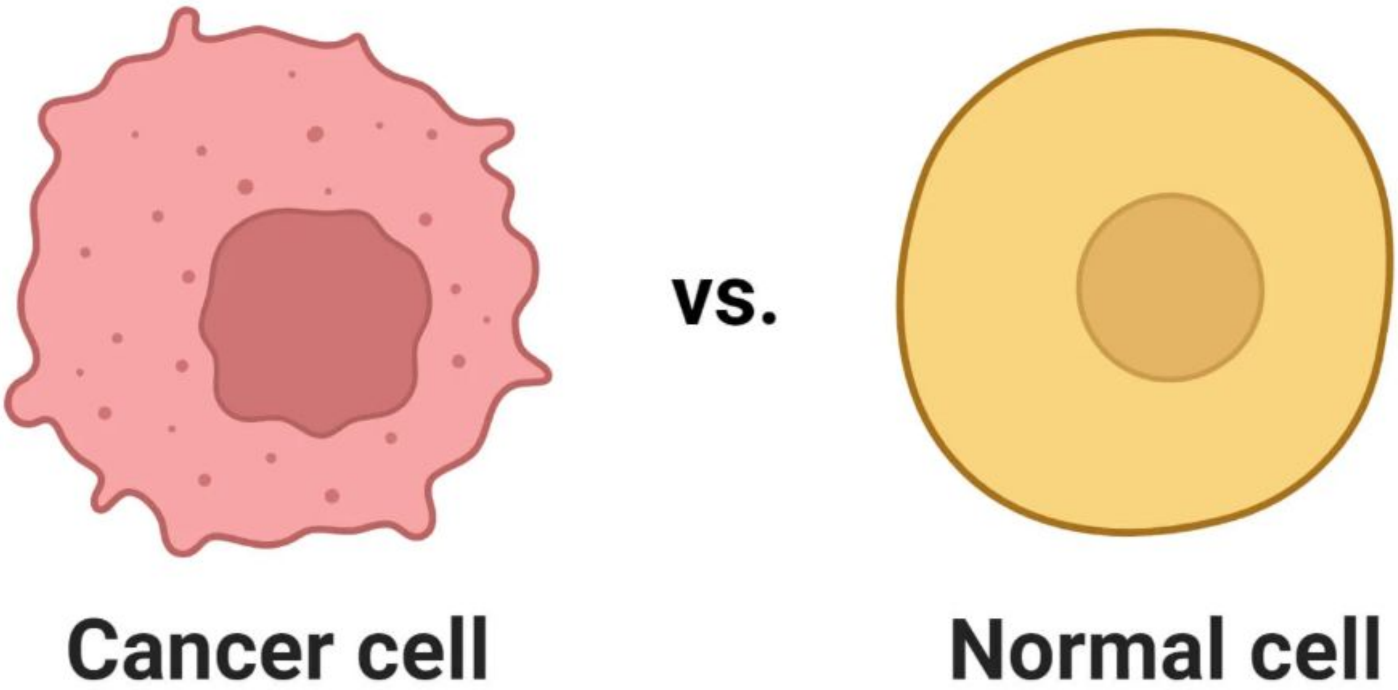
Accuracy: 97.67%

ROC AUC: 0.9994

*matrix values normalized by truths

Why Does it Work so well?

Knowledge Guided Feature Selection (KGFS)



Food for Thought

- Is accuracy all what we want?
  - Confusion matrix is more meaningful but we need to balance out false/true/positive/negatives
- A data challenge without hidden test sets is problematic
  - We inadvertently optimize for test data
- Real world is slightly (or heavily) OOD
- How these can be used for decision making?
  - Classifiers need to be calibrated
- Do we even need fancy ML?
  - Patient putting trust on an algorithm vs a doctor