GENE EXPRESSION FOR

PREDICTING ENDOMETRIOSIS

NAVIGATING NOISY DATA…

" Often dismissed as 'women's troubles', endometriosis affects one woman in 10 of reproductive age, yet a lack of research and funding means sufferers can live in severe pain, unable to work or socialise.

Sarah Boseley, Jessica Glenza *and* **Helen Davidson**

The Guardian

Agenda

Problem Statement

Data Source:
- ~28k genes, 408 samples (115 patients, 53 controls) from Turku dataset

Motivation:
- Identify most predictive gene features to better understand and detect endometriosis

Key Question:
- Can we maintain or improve endometriosis prediction accuracy using Random Forest or XGBoost after drastically reducing features from over 28k genes to a small, most-predictive subset - and what does this subset reveal about the disease?

Goal:
- Reveal top predictive genes to inspire therapeutic targets and insights into disease biology

PATCHES OF ENDOMETRIOSIS
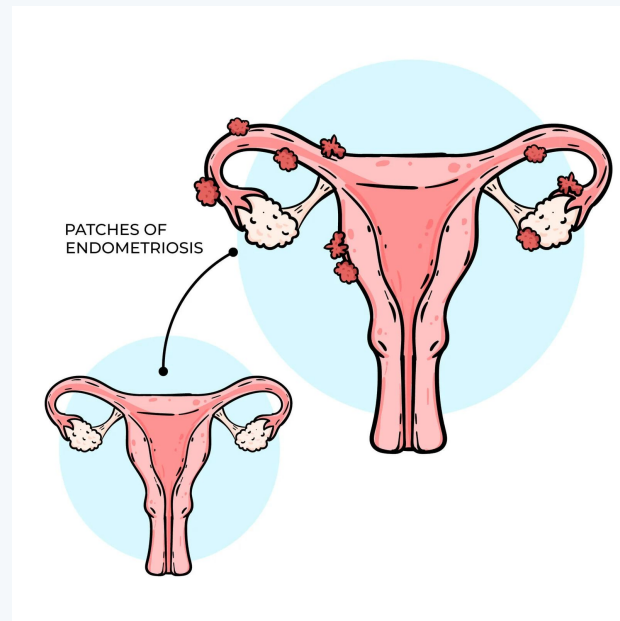
Visualization & EDA [1]

- **Data Integrity:** No missing values; duplicates averaged
- **Class Imbalance:** More diseased than healthy samples
- **Data Structure:** Transposed so samples = rows, genes = columns; labels derived from sample names
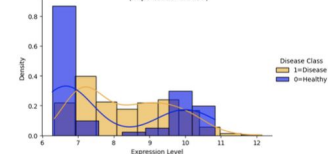
Healthy Control (disease colum = 0)	Disease (disease colum = 1)
CP: peritoneum samples from healthy control **CE:** endometrium sample from healthy control	**PeLR:** red peritoneal endometriotic lesion **DiEIn:** intestinal endometriotic lesions **PeLW:** white peritoneal endometriotic lesion **REV:** deep rectovaginal endometriotic lesions **DiEB:** deep endometriotic lesions in the bladder **PE:** endometrium sample from endometriosis patient **SuL:** sacrouterine ligament endometriotic lesions **PeLB:** black peritoneal endometriotic lesion **OMA:** ovarian endometrioma samples **PP:** peritoneum samples from patients

Visualization & EDA [2]

Histogram of top 3 most predictive genes across the 2 classes	Boxplots of top 3 most predictive genes across the 2 classes
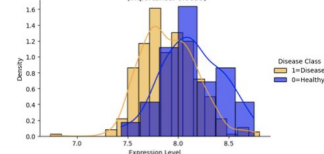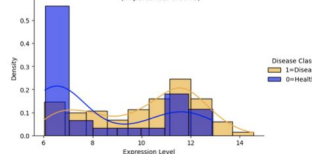
Visualization & EDA [3]

Modeling Overview

```
┌──────────────────┐      ┌──────────────────────┐      ┌──────────────────────┐
│ Data Cleaning    │ ───> │ Train Baseline Model │ ───> │ Train same Random    │
│ & EDA            │      │ (Random Forest) w/   │      │ Forest again w/ just │
│                  │      │ all 28k features     │      │ top 5 features       │
└──────────────────┘      └──────────────────────┘      └──────────────────────┘
                                                                    │
                                                                    v
┌──────────────────┐      ┌──────────────────────┐      ┌──────────────────────┐
│ RFE 2 (feature   │ <─── │ Train Fine-Tuned     │ <─── │ RFE 1 (feature       │
│ selection)       │      │ Random Forest w/     │      │ selection) w/        │
│ w/ XGBoost       │      │ features from RFE 1  │      │ Random Forest        │
│                  │      │                      │      │ Estimator            │
└──────────────────┘      └──────────────────────┘      └──────────────────────┘
         │
         v
┌──────────────────┐
│ Train Fine-Tuned │
│ XGBoost w/       │
│ features from    │
│ RFE 2            │
└──────────────────┘
```

Training Details

Problems

Solutions

HIGHLY DIMENSIONAL

The features of this data set are all the known genes – That's over 28,000 columns!

→

RFE

Recursive Feature Elimination! Its, more computationally expensive compared to PCA but maintains the interpretability of our findings!

IMBALANCED

Dataset contains 62 health samples in data set vs 330 endometriosis.

→

CUSTOM SCORER & SMOTE

Resampled data set using SMOTE and optimized for model macro average f1-score using a custom scorer while fine tuning model parameters.

OVERFITTING

Original data set only had 400 observations which is significantly less than the number of features which can lead to overfitting.

→

HYPERPARAMETER TUNING

Used GridSearchCV to explore the best number of estimators, best max depth, the minimum samples need to split or be considered a leaf in the tree.

Results & Inferences [1]

Model	Train Macro F1 Score	Train PC-Recall AUC	Test Macro F1 Score	Test PC-Recall AUC	Minority Class (0) Recall	Minority Class (0) Precision	Minority Class (0) F1 Score
Baseline RF (28k+ features)	1.00	1.00	0.73	0.99	0.33	1.00	0.50
Improved Baseline RF (Top 5 Predictors)	1.00	1.00	0.71	0.99	0.56	0.45	0.50
Fine-tuned RF w/ RFE 1	0.91	0.98	0.64	0.95	0.22	0.67	0.33
Fine-tuned XGBoost w/ RFE 2	1.00	1.00	0.80	0.99	0.89	0.53	0.67

Results & Inferences [2]

Top 1 & 2:

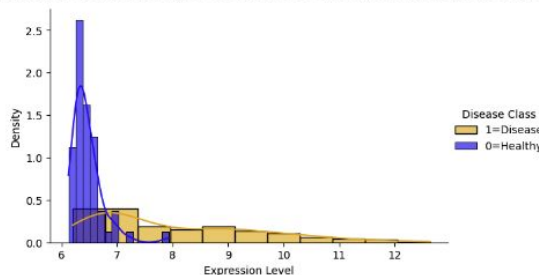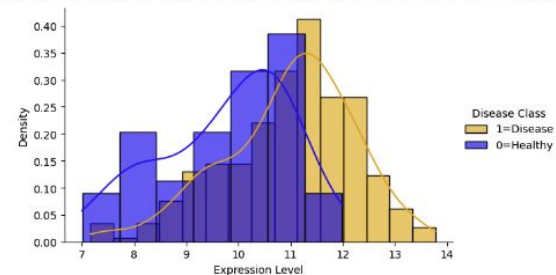Normalized distribution of gene ITGB1BP2_ILMN_1671871 expression by disease status

Normalized distribution of gene TPK1_ILMN_1804629 expression by disease status
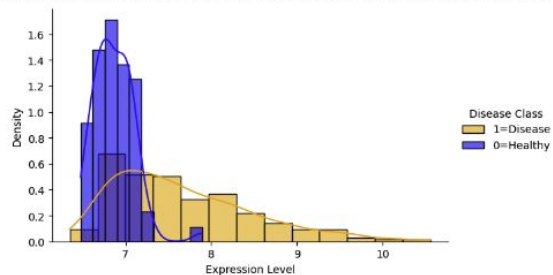
Top 3 & 4:

Normalized distribution of gene SCG2_ILMN_1703178 expression by disease status

Normalized distribution of gene WDR92_ILMN_1662065 expression by disease status
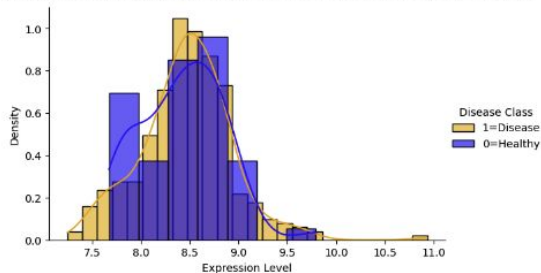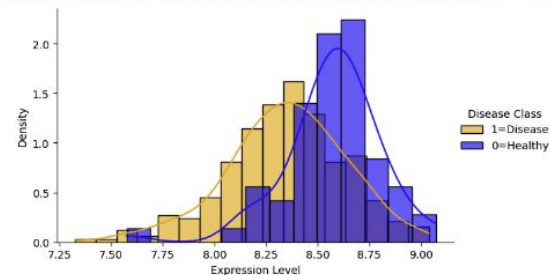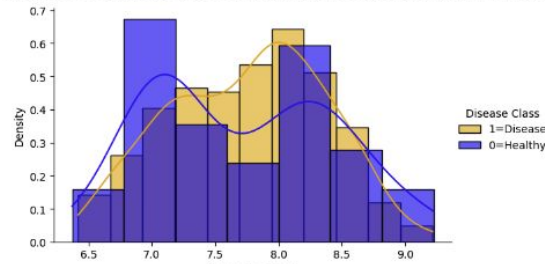
Top 5 & 6:

Normalized distribution of gene COL16A1_ILMN_1684554 expression by disease status

Normalized distribution of gene CD40_ILMN_1779257 expression by disease status

Next Steps

Limitations:
- Highly-dimensional, noisy data
- Relatively small dataset, only ~15% healthy samples

Potential steps:
- Other models to try: SVM

Thank you!