Endometriosis Early Detection

Daniel Moshe | Adi Haber

What Is Endometriosis?

- A chronic medical condition where endometrial tissue grows outside the uterus and adheres to other organs, mainly in the pelvic region.
- Main symptoms include chronic pain, infertility, fatigue, and sometimes even anxiety and depression.
- Endometriosis is prevalent mostly in women of reproductive age (15-49). Researchers say 10% of this population is affected - about 190 million patients worldwide.

Workshop Overview

Analyzing medical data found in UK Biobank using machine learning tools.

Research Question

Defining a research question regarding risk factors of a certain medical condition.

Creating a Cohort

Creating a cohort of patients relevant to the research question and cleaning it.

Extracting Features

Examining various medical studies and scientific articles to find features. Then extracting them from the UKB.

Creating a Model

Creating a machine learning model to get a prediction for the research question and repeating the process to improve it.

Our Research Question

What are the key risk factors associated with the development and diagnosis of Endometriosis?

Reasons for Choosing this Subject

- Endometriosis is under-researched, with significant gaps in understanding its causes, risk factors, and optimal diagnostic methods.
- The diagnosis process is long and tedious.
- There is a lack of awareness to this illness, both from the general population and medical professionals.

The Problem

Diagnosis dificulty

- Endometriosis is very hard to diagnose.
- Adhesions are hard to see using imaging.
- Diagnosis time averages at 7 years.

Effects of delay

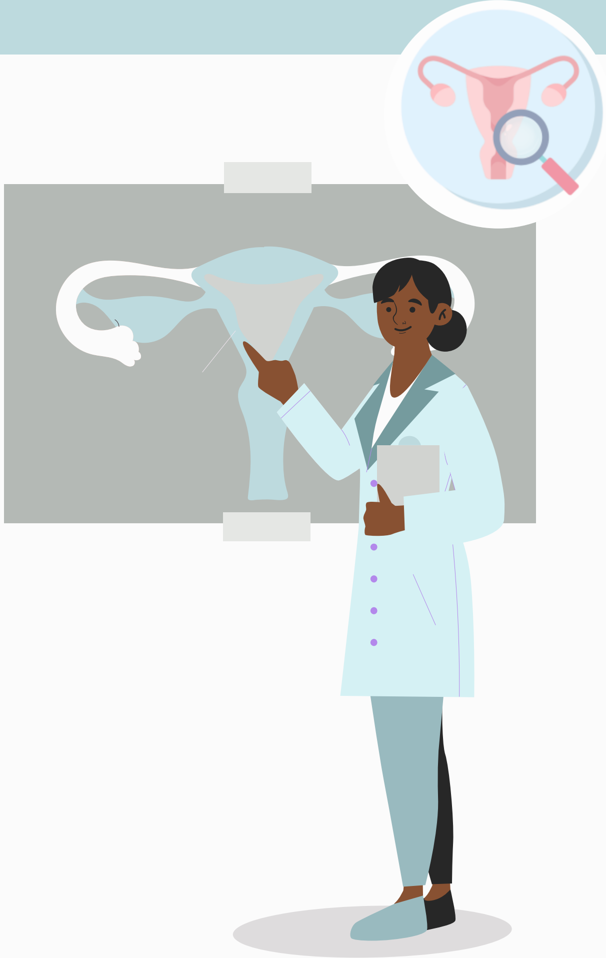
- Intensifies symptoms.
- Lowers quality of life.
- Causes incurable reproductive health challenges and infertility.

Current state

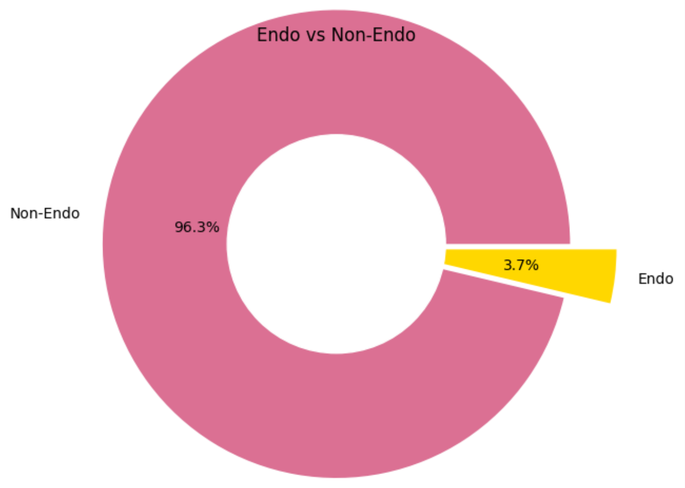
- Conventional diagnostic methods include invasive procedures.
- Doctors have subjective assessments.

Proposed Solution

- Diagnosing Endometriosis by analysing UK Biobank patient data.
- Developing a machine-learning model for precise Endometriosis detection using these features extracted from the UK Biobank.



Exploring our Data



~270,000

Female patients in the biobank

~10,000

Patients diagnosed with Endometriosis

Sparse Features

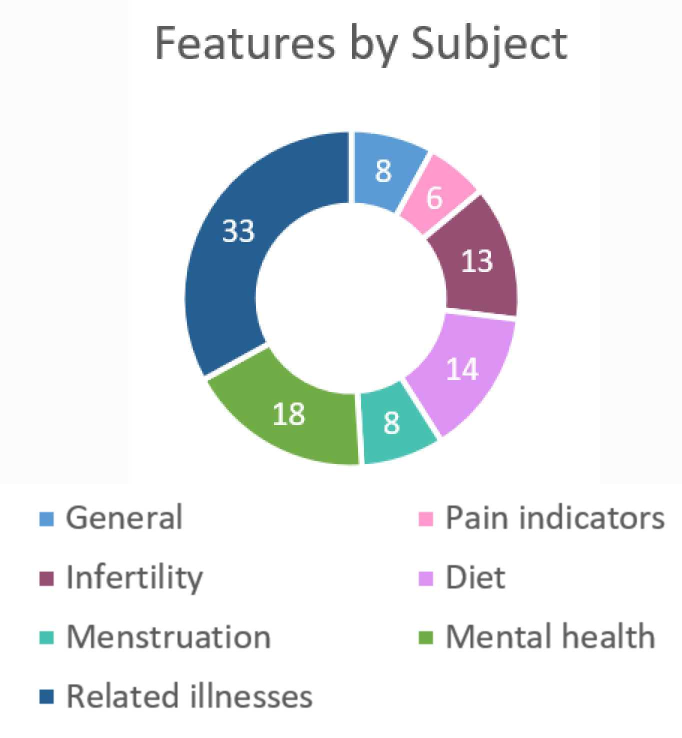
Many features are sparse, making feature extraction challenging.

Mixed Data Types

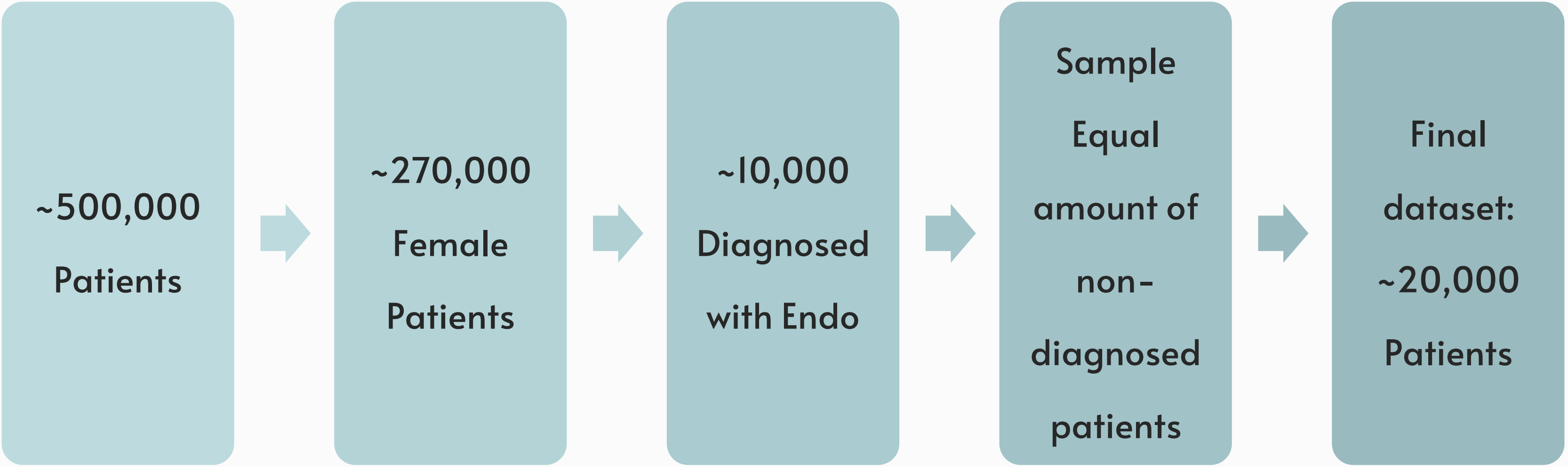
UKB contains both categorical and numerical data, requiring separate processing.

Feature Selection

- Created a list of 100 features by feature id, from the UK Biobank Showcase.
- Features include:
  - General data
  - Related diseases
  - Mental health
  - Infertility and pregnancy difficulties
  - Pain indicators
  - Diet
  - Menstruation



Data Filtering



~500,000 Patients

~270,000 Female Patients

~10,000 Diagnosed with Endo

Sample Equal amount of non-diagnosed patients

Final dataset: ~20,000 Patients

Data Preprocessing

Categorical Features

Mainly including dates and icd-10 codes. We converted them to one-hot encoding.

Feature Engineering

Added new features based on raw data, like number of icd-10 diagnoses and estrogen exposure.

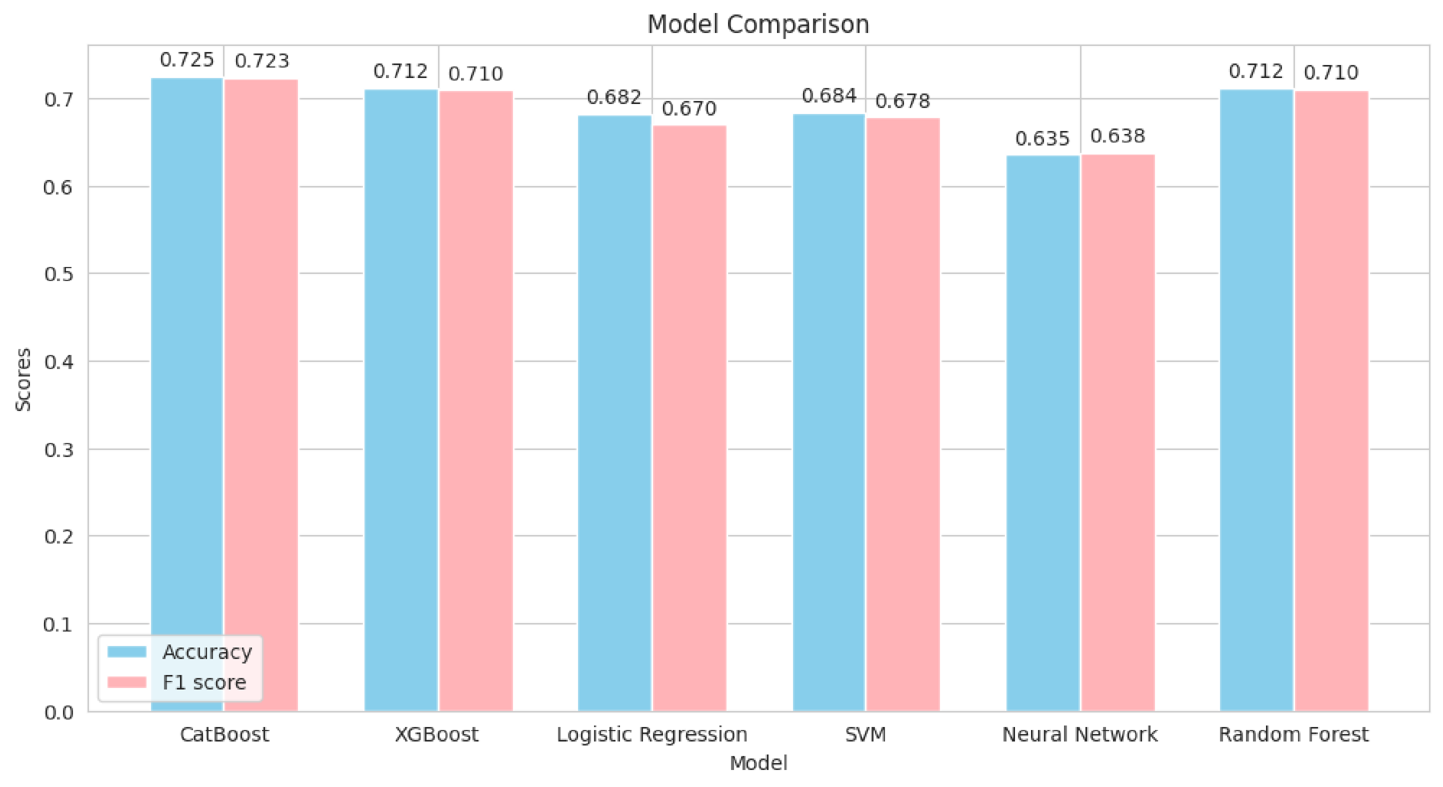
Feature Correlations

Analyzed feature correlations and eliminated highly correlated features to reduce redundancy and improve model performance.

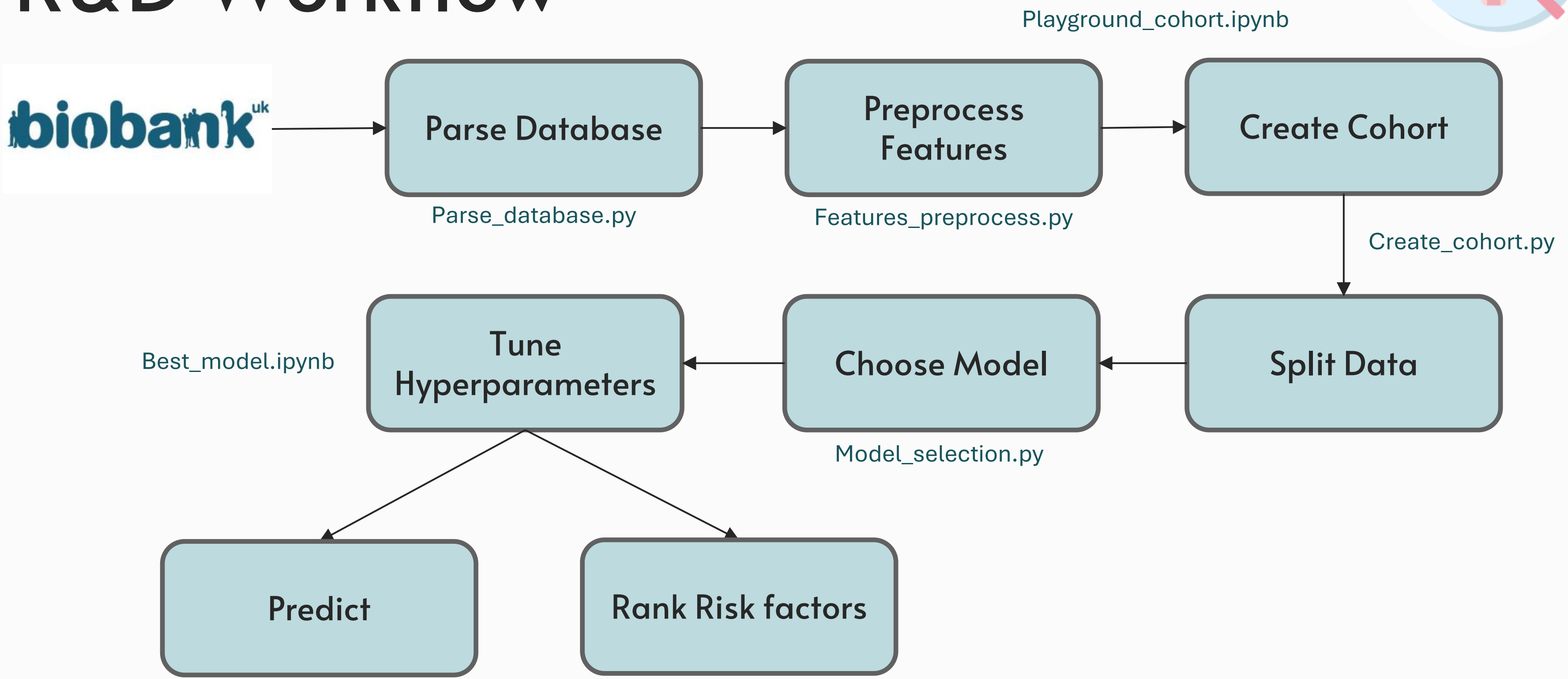
Missing Data

Removed features with more than 90% missing data. Remaining missing values were addressed through mean imputation.

Model Selection



R&D Workflow



biobank uk

Parse Database

Parse_database.py

Preprocess Features

Features_preprocess.py

Create Cohort

Playground_cohort.ipynb

Create_cohort.py

Split Data

Choose Model

Model_selection.py

Tune Hyperparameters

Best_model.ipynb

Predict

Rank Risk factors

Dilemmas and Challenges

Remote Code

There was no easy way to run interactive python on the remote server, so we used Jupyter remote kernels with SSH tunneling.

Data Imputation

After choosing to use CatBoost, we realized there is no improvement to results after imputation, so we skipped it in the final model.

Generic Code

We aimed to write generic code and classes that are adaptable for use in similar projects, whether for biobank usage or general ML usage.

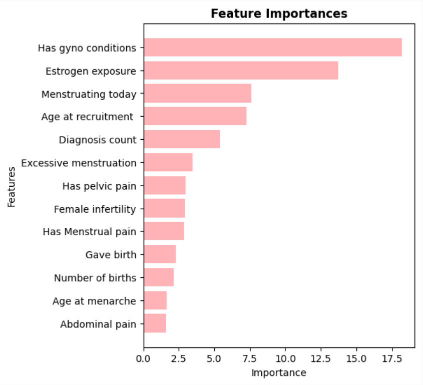
Validation Set

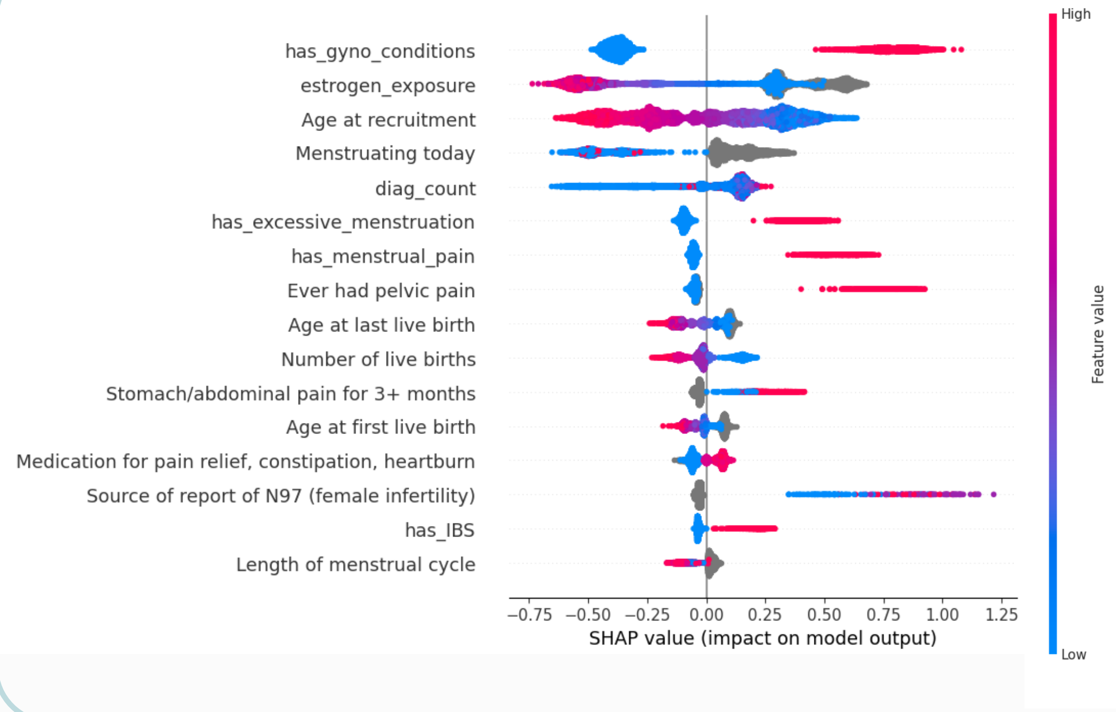
We considered using a validation set for tuning model hyperparameters, but opted for cross-validation, as our dataset is limited in size.

Result Analysis

Predictions:

- Accuracy: 0.73
- F1 Score: 0.72





Model Limitations

Ages of Patients

- Endometriosis is usually discovered at late 20s or early 30s.
- Biobank average diagnosis age is 42.
- Data might not truly represent patients today.

Diversity of Patients

- Biobank does not contain a diverse population.
- Mainly comprised of white British population.

Balanced dataset

- We chose a 50% ratio of diagnosis in our cohort.
- In the general population, 10% have endo.
- In healthcare settings, ratio probably differs.

Conclusions

Impact

- Raising awareness to Endometriosis.
- Potentially reducing diagnosis time and improving patient outcomes.

Application

- Model can be applied in clinical settings to assist healthcare professionals in diagnosing Endometriosis earlier.

Future Work

- Exploring the model's application to other related conditions.
- Expanding the dataset to include more diverse populations.

Thank you

Endometriosis Early Detection | Daniel Moshe, Adi Haber