**Endometriosis early detection**

Daniel Moshe, Adi Haber

A blue glowing uterus

Description automatically generated

# Table of Contents

[Table of Contents](#_Toc176244362)

[Background 3](#_Toc176244363)

[Problem Statement 4](#_Toc176244364)

[Proposed Solution 4](#_Toc176244365)

[Introduction 4](#_Toc176244366)

[Machine Learning 4](#_Toc176244367)

[Endometriosis 5](#_Toc176244368)

[UK BioBank [6] 5](#_Toc176244369)

[Data Preprocessing 5](#_Toc176244370)

[Stage one: Feature Selection 5](#_Toc176244371)

[Stage Two: Feature Extraction 11](#_Toc176244372)

[Stage Three: Analyzing the Data 13](#_Toc176244373)

[Stage Four: Handling Categorical features 16](#_Toc176244374)

[Stage Five: Feature Engineering 18](#_Toc176244375)

[Data Imputations and Cleansing 20](#_Toc176244376)

[Removed Features 20](#_Toc176244377)

[Imputing Features 20](#_Toc176244378)

[Model Selection 22](#_Toc176244379)

[Model Selection Overview 22](#_Toc176244380)

[Models Tested 22](#_Toc176244381)

[Model Selection Class 23](#_Toc176244382)

[Model Selection results 24](#_Toc176244383)

[Tuning Chosen model 25](#_Toc176244384)

[Result Analysis 26](#_Toc176244385)

[Model Limitations and Challenges 27](#_Toc176244386)

[Conclusions 29](#_Toc176244387)

[Acknowledgments 29](#_Toc176244388)

[Bibliography 30](#_Toc176244389)

# Background

Endometriosis, a chronic inflammatory condition, primarily manifests through symptoms such as pain and infertility [1]. It occurs when tissue resembling the uterine lining grows outside the uterus, adhering to pelvic organs, and occasionally other areas of the body. This abnormal adhesion triggers inflammation and the formation of scar tissue, resulting in debilitating pain and, in some instances, infertility. Endometriosis predominantly affects women of reproductive age, with research suggesting that approximately 5-10% of this demographic, totalling around 180 million individuals globally, are affected [2].

Main known indicators of the endometriosis include:

* Pelvic pain and/or lower abdominal pain
* Painful menstrual cramps
* Abnormal menstrual bleeding pattern (either by amount or irregularity)
* Family history [3]
* Infertility [4]

Endometriosis can be classified to 4 stages, based on the size and depth of the adhesions:

* Stage I: minimal
* Stage II: mild
* Stage III: moderate
* Stage IV: severe

The symptoms of endometriosis do not correlate with the stages, meaning Stage IV patients can have no symptoms and stage I patients can have severe symptoms.

Diagnosing endometriosis presents a challenge since adhesions are not always detectable through imaging techniques like ultrasound or MRI. Typically, a definitive diagnosis necessitates undergoing laparoscopic surgery [3] and a subsequent biopsy.

Problem Statement  
Our primary objective revolves around the prompt identification of endometriosis. Remarkably, 60% of women dealing with endometriosis navigate consultations with three or more clinicians before receiving a diagnosis, leading to an average delay of seven years before definitive diagnosis [5]. This prolonged delay intensifies symptoms, lowers overall quality of life, and contributes to enduring reproductive health challenges. Conventional diagnostic methods, predominantly reliant on invasive procedures and subjective assessments, further complicate the diagnostic process.

# Proposed Solution

This project endeavors to aid in diagnosing endometriosis by analyzing patient data. We will collect data on endometriosis and healthy patients from the UK Biobank and select a group of features (symptoms and risk factors) from which we will try to detect the existence of endometriosis. Our primary goal is to build the optimal machine-learning model for accurate endometriosis detection based on the features we found.

# Introduction

## Machine Learning

Machine learning, a subset of artificial intelligence, revolutionizes medical research by extracting insights from vast datasets to enhance diagnostic accuracy, treatment efficacy, patient outcomes, and identifying risk factors. There are two primary subcategories of machine learning - supervised and unsupervised learning.

Supervised learning algorithms use labeled data to train models to predict outcomes or classify instances, offering valuable insights into disease detection and prognosis.

Unsupervised learning techniques uncover hidden patterns within unlabeled data, enabling researchers to identify unexplored disease subtypes or biomarkers.

Deep learning, a subset of machine learning, utilizes neural networks with multiple layers to automatically extract complex features from raw data, paving the way for advanced image analysis, genomic sequencing, and complex medical issues.

With the integration of these machine learning paradigms, medical researchers unlock unprecedented opportunities to unravel the complexities of diseases, revolutionizing healthcare delivery.

## Endometriosis

Endometriosis, a prevalent chronic gynecological condition reliant on estrogen, concerns the presence of uterine endometrial tissue outside its normal cavity. This disorder is characterized by the presence of endometrial tissue outside the uterus, leading to pelvic pain and fertility issues.

## UK BioBank [6]

UK Biobank is a large-scale biomedical database and research resource, containing in-depth, de-identified genetic and health information from half a million UK participants. The database, which is regularly augmented with additional data, is globally accessible to approved researchers and scientists undertaking vital research into the most common and life-threatening diseases. UK Biobank’s research resource is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.

# Data Preprocessing

## Stage one: Feature Selection

The first step of our research was to find the features we would like to try and detect endometriosis by. We have sorted the features into groups:

**General Features**

The general features include:

* Sex/Genetic sex
* Age at recruitment
* Year of birth
* Weight/BMI

As endometriosis mostly effects women, we will use the sex data field to filter female patients.

We chose to include weight\BMI as features because we have found several research papers stating there is an inverse relationship between BMI and the risk of endometriosis, meaning endometriosis is more commonly associated with lower BMI [7].

The year of birth and age at recruitment features will help us infer which of the other features are relevant to our goal. Endometriosis mainly effects women of reproductive age, meaning data of patients over this age group may not be relevant.

**Pain Indicators**

One of the most recognizable symptoms of endometriosis is reported to be unusually painful periods, and pelvic pain in general. Research shows that 45% of patients with chronic pelvic pain also have endometriosis [8]. Other pain related symptoms that are usually associated with endometriosis include:

* Lower abdomen pain
* Back pain
* Hip pain
* Recurring Headaches and migraines
* Genital organ pain

We have added these symptoms, including pelvic pain and menstruation pain as features.

**Infertility**

Research shows that 30% of patients with infertility have endometriosis. Researchers believe the two are connected, meaning endometriosis might cause fertility issues in some of the patients, depending on the severity (stage) of the endometriosis [8].

Therefore, we expect that a diagnosis of infertility might also indicate a risk of endometriosis.

**Hormonal Contraception**

Hormonal birth control is one of the most common treatments for endometriosis. The hormonal contraceptive regulates estrogen levels. Endometriosis is estrogen-dependent therefore endometriosis patients tend to have excess estrogen. To stop estrogen production, meaning decrease the growth of endometrial-like tissue, Hormonal birth control is usually suggested as a treatment.

With this information in mind, we expect an inverse relationship between risk of endometriosis and using hormonal contraceptive. We also expect that a large portion of the patients diagnosed with endometriosis might also be taking a hormonal contraceptive.

**Menstruation**

In most patients with endometriosis, we can see an abnormal menstrual bleeding pattern, whether it is in amount (excessive bleeding for a longer duration than usual, meaning the menstrual cycle is shorter than 27 days) or in irregularity [1].

Moreover, the age of a patient at menarche seems to have a connection to endometriosis. Patients that started menstruating before the age of 12 have a higher risk of endometriosis.

The features we have decided to select in this category include:

* Age at menarche
* length of menstrual cycle
* menstrual pain/cramps

**Diet**

Some research papers show a connection between endometriosis and nutrition. It is shown that increased dietary fat consumption and low fiber consumption might increase estrogen concentrations, therefore increasing the risk of endometriosis. The same goes for red meat and dairy consumption [9].

On the contrary, vitamin D, C and E, and antioxidants in general are associated with lower risk of endometriosis, as they are anti-inflammatory [9].

The features we have selected to select for the stated dietary preferences for endometriosis:

* For dietary fat intake, we have selected fat consumption.
* For fiber consumption, we chose to look at fresh fruit and vegetables intake.
* For red meat consumption, we chose beef an processed meet intake.
* For dairy consumption, we chose lactose intake and cheese intake.
* For vitamins C, D, E we chose the corresponding vitamin intake features.

**Pregnancy**

Research shows a relationship between endometriosis and pregnancy complications. It has been suggested that endometriosis might change the uterine environment by causing progesterone resistance in the endometrium. This difficulties in embryo development and implantation [10].

We have decided to select as features some of the most common pregnancy complications:

* General complications
* Preterm delivery
* Spontaneous miscarriage/abortion
* Stillbirth
* placenta praevia (vaginal bleeding during pregnancy)

Moreover, many research papers indicate a relationship between delivery type and endometriosis: the rate of delivery by caesarean sections is higher in endometriosis patients. We have added delivery by caesarean section as one of our features as well.

**Mental state**

As stated above, pain is one of the most recognizable symptoms of endometriosis. The presence of pain significantly influences the psychological and social well-being of individuals with endometriosis, affecting their overall quality of life [12].

Mental symptoms seem to not get medical attention regarding endometriosis detection. We wanted to see if it is prevalent in endometriosis patients, so we added features regarding patient anxiety and depression.

**Related diseases**

Many research papers show a variety of diseases related to endometriosis. We have decided to choose some of the recurring ones, as well as some we were curious about.

The diseases we chose as features:

* Irritable bowel syndrome (IBS)
* Fatigue
* Anemia
* Lupus
* Fibromyalgia
* Hypothyroidism
* Allergies
* Blood clots
* Asthma
* Overian dysfunction
* Ovarian cancer
* uterine cancer
* cervical cancer
* Breast cancer
* Melanoma

After researching every feature group, we created a CSV file with all the features we selected from the UK Biobank showcase and their corresponding codes in the dataset. The goal of creating this CSV was to use it as input for the next step, planning the feature extraction.

Stage Two: Feature Extraction

As a part of our research for this project, we decided to search for a python library used to extract features from the UK Biobank. We were surprised we did not find a generic library that takes a csv of the features and generated a pandas data frame. So, we have decided to create one.

The UK Biobank data is organized in CSV files, in which the columns correspond to the different data fields available. The data (at least the part of the data we have access to) is organized in three main CSV files, and each file has a “fields” file that describes which fields are present in it.

The library we created is in the file “parse\_database.py” which is in the “Dataset” folder. In this file we created the class “UKBDatasetCreator” that is goal is to get a list of features by their code from the UKB, and create a pandas dataframe from the selected features.

We have also created a utility script that provides various functions and configurations to support data extraction and other parts of your project, which is called “Utils.py” and is in the “Model” folder.

**Utilities in this file:**

* Logging: The script sets up a logging configuration with a specific format for all of our project files.
* Load feature names: the function “init()” loads feature codes and corresponding name from a csv we created, to be used for converting feature code to readable name.
* Converting feature to code: given a keyword from a feature, the function “feature\_to\_code()” returns the corresponding code.
* Converting code to feature: given a feature code, the function “code\_to\_feature()” returns the corresponding feature name.
* Changing a feature name: some of the features have long and non-indicative names. The function “change\_feature\_name()” allows us to change the name of a feature in our dataset.
* Plotting and printing: the rest of the functions focus on plotting and printing for the notebooks where the rest of the code is in.

## Stage Three: Analyzing the Data

We Started by analyzing the gender distribution in our dataset, to ensure its relevance to the condition under study. Given that endometriosis mostly affects female patients, we identified and subsequently removed all male entries from our dataset. This left us with about 275,000 records, containing only female patient data.

A graph of patients by gender

Description automatically generatedFigure 1: female to male patient ratio in the dataset

At this stage, we excluded male patients from our dataset.

Next, we checked the prevalence of Endometriosis-affected women in the dataset. There are 10,171 patients that were diagnosed with endometriosis.

This revealed that the prevalence of Endometriosis-affected women in our dataset is approximately 4%. This distribution differs from existing research on the prevalence of endometriosis in the general population, at around 10%. This discrepancy might be because all women recruited were aged 40–69, meaning they were in the reproductive age range (15-49, as defined by the World Health Organization) in the late 90’s – early 2000’s, when there was less awareness of Endometriosis. Moreover the 10% distribution in the general population is an estimate based on sample tests, meaning even if 10% of the population has endometriosis, not all Endometriosis patients know they have it or are diagnosed with it. Despite this, we took several measures to minimize the effect of this discrepancy on our analysis, as described below.

A pink circle with yellow text

Description automatically generatedFigure 2: prevalence of Endometriosis in initial dataset.

To try and balance the rate of Endometriosis patients, we have decided to create a balanced dataset with 50% diagnosed and undiagnosed ratio. We sampled the test and train sets randomly.

A diagram of a patient's health

Description automatically generated

Figure 3: Process of data filtering

Next, we checked the average diagnosis age. To calculate the average diagnosis, we used the features: date of first diagnosis, date of birth. We discovered that on average, Biobank patients were diagnosed at the age of 42. In the general population, an average woman will be diagnosed in her 30s. This poses a limitation of our data and consequently a future limitation of our model, as the model’s train set includes patients mostly diagnosed after the current average age of diagnosis.

A graph of age at diagnosis

Description automatically generated  
Figure 4: diagnosis age counts

Then we wanted to see how sparse our features are. There are ! features with more than 50% missing values, meaning !% of our features are very sparse. After many tests, we have decided to remove features with more than 90% missing values.

Figure 5: Sparseness of features

## Stage Four: Handling Categorical features

In our dataset, we identified several categorical features, which mainly consist of dates and ICD-10 disease codes. These features are crucial as they capture the presence or absence of specific diagnoses, or the timing of certain events related to a patient's health.

Given the nature of these categorical features, particularly that they represent the dates of illness reports or specific ICD-10 codes, we decided to encode them using one-hot encoding, as we wanted the information on if a patient has the certain illness or not, and not when he has reported he has it or what is the icd-10 code of the illness.

**What is One-Hot Encoding?**

One-hot encoding is a method used to transform categorical variables into a format that can be provided to machine learning algorithms to improve predictions.

One-hot encoding converts categorical variables into a binary (0 or 1) matrix, where each unique value in a categorical feature is represented by a separate column. For instance, if we have a feature representing disease codes with values like "A01," "B02," and "C03," one-hot encoding would create three new binary columns, each corresponding to one of these codes. If a patient has a diagnosis of "A01," the respective column will have a value of 1, while the others will be 0.

**Why One-Hot Encoding for Our Dataset?**

The main reason we chose one-hot encoding for these features is that we were primarily interested in capturing whether a patient had a particular diagnosis, rather than the specific date it was reported or the exact disease code. By one-hot encoding the dates and ICD-10 codes, we transformed these categorical features into a binary format that simply indicates the presence or absence of a condition. This approach allowed us to focus on whether a patient has been diagnosed with a particular illness, which is crucial for the predictive models we are building for early detection of endometriosis.

**Analysis of One-Hot Encoding**

Some of the features were combined for encoding, for example: “[Date K50 first reported (crohn's disease)](https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=131626)”, “[Date K59 first reported (other functional intestinal disorders)](https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=131640)“, “[Date K52 first reported (other non-infective gastro-enteritis and colitis)](https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=131630)”, “Date K51 first reported (ulcerative colitis)” were combined to the feature “has gastro conditions” as the specific gastrointestinal illness is not as critical to our model.

A graph with numbers and text

Description automatically generated with medium confidence

Figure 6: Categorical features diagnosis rate

## Stage Five: Feature Engineering

Feature engineering is used in a machine learning project to improve model accuracy and efficiency. By transforming existing features, we can create new features we think will better the predictive power of our model.

**Endometriosis diagnosis feature**

The first feature transformation we did was to create a feature that will represent endometriosis diagnosis, and the source of the diagnosis. The UK Biobank has two separate features for this purpose, and they use different data encodings. The feature we determine endometriosis patients by is “[Date N80 first reported (endometriosis)](https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=132122)”. The data for this feature is dates (with a few special date encodings). We wanted to create a nominal feature that will include not only the diagnosis (has endometriosis or not), but also the source of diagnosis (medical diagnosis or self-diagnosis). We created a feature called “endo diagnosis”, a nominal feature with the following encoding:

* 0 – not diagnosed with endometriosis.
* 1 – endometriosis medical diagnosis.
* 2 – endometriosis self-diagnosis.

We decided to save the ‘date first reported’ feature to use it to gauge information about the patients age and the relevancy of other features to the diagnosis.

**Number of ICD-10 diagnoses**

Endometriosis is very hard to diagnose, and its symptoms are commonly mistaken for those of other medical conditions. Moreover, the Endometriosis diagnosis process is usually very long, taking 7-8 years on average, meaning many other illnesses can be diagnosed in this time. Knowing this, we assume that Endometriosis patients will have more IDC-10 diagnoses than non-Endometriosis patients. For this reason, we decided to add the number of ICD-10 diagnosis each patient has as a feature. We inferred this feature from the ‘hesin\_diag’ file, which contains icd-10 codes of different diagnosis each patient has. We grouped the data by patient id and added the number of diagnoses each patient has to our main dataset.

**Estrogen exposure**

One of the most common risk factors of Endometriosis is prolonged estrogen exposure [3]. Thus, we wanted to include estrogen exposure as a feature, with the assumption that longer estrogen exposure will correlate with an Endometriosis diagnosis.

We calculated Estrogen exposure (in years) by reducing the age of menarche (age at first menstrual cycle) from age of menopause (age after last menstrual cycle).

# Data Imputations and Cleansing

The UK Biobank data is very sparse, containing numerous NaN values. For instance, one feature we wanted to analyse was whether the patient has anaemia, to explore a potential correlation between anaemia and endometriosis. However, upon extracting this feature, we found that only 256 women had documented data on anaemia diagnosis. During the data imputation and cleansing stage, we have decided whether to retain such features and impute the missing values or exclude these features.

## Removed Features

As stated before, after many tests we have decided to remove features with more than 90% missing values, as we saw that even after imputations, these features did not contribute to the model.

## Imputing Features

For the remaining features with missing values, we implemented feature imputation to fill in these gaps. Imputation is a technique used to replace missing data with substituted values, allowing us to maintain the integrity of the dataset and ensure that our machine learning models can be trained effectively. We explored several imputation methods:

**Mean Imputation:**

This method replaces missing values with the mean (average) of the observed values for that feature. It’s a straightforward approach that assumes the missing data is randomly distributed.

We found that mean imputation worked best for our dataset, likely because it provided a balanced substitute that didn’t skew the data distribution.

**Median Imputation:**

In this method, missing values are replaced with the median of the observed values. The median is the middle value in a sorted list and is less sensitive to outliers than the mean.

While median imputation is useful for datasets with skewed distributions, it didn’t perform as well as mean imputation in our case.

**Most Frequent Imputation:**

Also known as mode imputation, this method fills in missing values with the most frequently occurring value in the feature. This approach is particularly useful for categorical features.

This method was less effective for our dataset, which contains many continuous variables, making it less suitable for our needs.

**K-Nearest Neighbours (KNN) Imputation:**

KNN imputation replaces missing values based on the values of the k-nearest neighbours in the dataset. It’s a more sophisticated method that considers the relationships between different features.

Although KNN imputation is powerful, it is also computationally intensive and didn’t consistently outperform mean imputation in our testing.

After evaluating these methods, we determined that mean imputation provided the best results for our specific dataset. It offered a good balance between simplicity and effectiveness, ensuring that the imputed values didn’t introduce bias or distort the underlying data patterns. This choice allowed us to maintain a consistent approach across features, contributing to more reliable model performance.

# Model Selection

## Model Selection Overview

We are planning to try 3 main groups of models:

**Boosting Models:** Like XGBoost and CatBoost, which are proven to be good candidates for processing tabular data and can work with sparse data (data that contains many NaN values). As The UKB data is very sparse, we hope that these models will be able to generalize the data and give a good prediction.

**Classic Machine Learning Models:** Like logistic regression and SVM, on the imputed data.

**Deep Learning Models:** We wanted to try and see if a Neural network can identify complex patterns in the data, as Endometriosis is a condition with multiple different causes and symptoms, which might be hard for a traditional ML model to generalize.

## Models Tested

**Boosting Models:**

* XGBoost (Extreme Gradient Boosting): A powerful ensemble method that uses gradient boosting techniques to create a series of models that correct the errors of previous models. It is well-known for its performance in various machine learning competitions.
* CatBoost (Categorical Boosting): Specifically designed to handle categorical data more effectively without requiring extensive preprocessing. CatBoost automates the treatment of categorical features, which is particularly useful for datasets like ours.

**Classic Machine Learning Models:**

* Logistic Regression: A statistical model that predicts the probability of a binary outcome. It’s simple yet effective for linearly separable data.
* Support Vector Machine (SVM): A model that finds the optimal hyperplane to separate different classes in the feature space. SVMs are powerful for high-dimensional datasets.
* Random Forest: An ensemble learning method that creates multiple decision trees and combines their outputs to improve accuracy and reduce overfitting.

**Deep Learning Models:**

* MLP (Multilayer Perceptron): A type of feedforward artificial neural network that consists of multiple layers of nodes. MLPs can capture complex patterns in data, making them suitable for a wide range of tasks.

## Model Selection Class

Our goal was to create a class that would allow us to easily train and compare the performance of various models on their dataset and identify the best-performing model for their specific use case. The class handles the training, evaluation, and model selection process, allowing us to focus on the higher-level task of model selection and comparison, rather than focus on the details of training individual models.

The `ModelSelector` class provides a convenient way to train and evaluate multiple machine learning models on a given dataset, and automatically selects the best-performing model based on accuracy and F1 score.

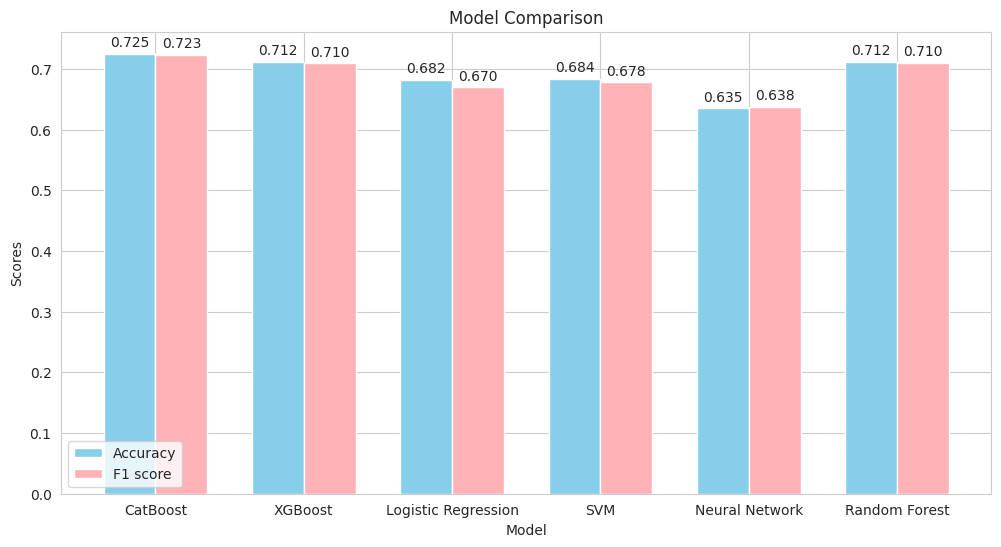
**How the class works:**

1. `\_\_init\_\_()`: The constructor initializes a dictionary of models to be trained, including XGBoost, CatBoost, Logistic Regression, SVM, MLP Classifier, and Random Forest. It also initializes variables to store the training results and the best model.
2. `train\_models()`: This method takes the training and testing data as input, and trains each of the models in the `models` dictionary. For each model, it prints the model name, fits the model to the training data, makes predictions on the test data, and saves the accuracy and F1 score in the `results` dictionary.
3. `save\_results()`: This helper method stores the accuracy and F1 score for a given model in the `results` dictionary.
4. `update\_best\_model()`: This method checks if the current model's accuracy is better than the best model seen so far. If so, it updates the `best\_model`, `best\_score`, and `best\_model\_name` variables.
5. `print\_results()`: This method prints the accuracy and F1 score for each trained model.
6. `plot\_results()`: This method creates a bar plot to visualize the accuracy and F1 score for each trained model.

The main benefit of this `ModelSelector` class is that it provides a standardized way to train and evaluate multiple machine learning models on a dataset, and automatically selects the best-performing model based on the specified metrics (in this case, accuracy and F1 score).

## Model Selection results

After running the model selection, results show CatBoost was selected as the final model for our project. Its ability to natively handle categorical variables, combined with its superior performance metrics, made it the ideal choice for our project.

Figure 7: Model selection results

|  |
| --- |
|  |
|  |  |

## Tuning Chosen model

After selecting CatBoost as the best-performing model in our project, we focused on hyperparameter tuning to optimize its performance further. Hyperparameter tuning involves adjusting the model’s parameters that are not learned from the data but are set before the training process begins. The goal is to find the combination of hyperparameters that results in the best performance of the model.

Code for hyperparameter tuning is found in “best\_estimator.py”

We used grid search to systematically explore the best hyperparameters and chose to use cross validation instead of a separate validation set.

**What is Cross-Validation?**

Cross-validation is a technique used to assess how well a model generalizes to an independent dataset. In k-fold cross-validation, the data is divided into k subsets (folds). The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, each time with a different fold as the test set, and the results are averaged to provide a robust estimate of the model's performance.

**Why Cross-Validation Instead of a Validation Set?**

* Efficient Use of Data: In our dataset, splitting the data into separate training, validation, and test sets would reduce the amount of data available for training. By using cross-validation, we ensured that every data point was used both for training and validation, maximizing the dataset's utility.
* Better Performance Estimates: Cross-validation provides a more reliable estimate of a model's performance because it reduces the variance that can arise from the train-test split. It averages the model's performance across multiple folds, giving a more stable and generalizable measure metrics.

# Result Analysis

# Model Limitations and Challenges

In our research, we confront several notable constraints that shape the scope and reliability of our findings

**Age of Patients**

while utilizing the UK Biobank dataset provides valuable insights, its demographic skew towards women averaging 50 years old presents a limitation. Given that our objective is to assist in diagnosing endometriosis in younger women, the dataset may not fully represent the nuances of the condition in this demographic.

**Age of Endometriosis Diagnosis**

We encountered a discrepancy between the average age of endometriosis diagnosis in the biobank data and the true age at diagnosis observed in the general population. Endometriosis is typically diagnosed in women in their late 20s to early 30s. This is the age range where symptoms often become more prominent and severe, leading to a diagnosis. In contrast, the average diagnosis age in our biobank data was significantly older, at 42 years. This discrepancy suggests that the biobank data might not accurately represent the current patient population. This age discrepancy could mean that the data may not fully capture the characteristics of the current patient population, particularly younger women who are now being diagnosed earlier. Our might have limited generalizability to the broader population, particularly for predicting endometriosis in younger women who are now more commonly diagnosed in their late 20s or early 30s.

**Lack of diversity in dataset**

Another critical limitation of our study is the lack of population diversity within the UK Biobank (UKB) dataset. This issue impacts the generalizability of our findings across different ethnic and racial groups. The UK Biobank is predominantly composed of individuals from a white British background. This lack of diversity means that the dataset does not fully represent the broader, multi-ethnic population that exists within the UK or globally. Thus, our model might inherit and perpetuate biases present in the dataset, which can lead to less accurate predictions and poorer performance when applied to more diverse populations.

**Balanced Dataset**

One of the intentional limitations in our study design is the use of a balanced dataset, where 50% of the patients have endometriosis and 50% do not. This contrasts with the general population, where it is estimated that only about 10% of women have endometriosis. While this discrepancy exists, our decision to use a balanced dataset was driven by specific research goals and the intended application of our model. This decision was made to focus on an "artificial problem" that would allow us to explore and develop machine learning techniques without being influenced by the natural class imbalance present in the general population.

Moreover, our model is intended to be used in clinical settings by health professionals. The 10% prevalence figure represents the general population, but in clinical settings, this number is likely higher because patients seeking medical care often do so due to symptoms or concerns. Thus, the prevalence of endometriosis among patients being evaluated by medical professionals is likely greater than 10%, and the 50/50 approach is likely better fitted to clinical settings.

In summary, our study faced several limitations that could impact the generalizability and applicability of our findings. The discrepancy in average diagnosis age and the lack of population diversity in the UK Biobank data suggest that our model may not fully represent the broader population. Additionally, the use of an artificially balanced dataset, while useful for research purposes, deviates from the true prevalence of endometriosis in the general population. Despite these challenges, the decisions made were intentional and aimed at advancing our research while ensuring relevance in clinical settings. Future work should focus on addressing these limitations to improve the model's accuracy and fairness when applied to more diverse and representative populations.

# Conclusions

Our study has significant implications for both the awareness and diagnosis of endometriosis. By developing and validating a machine learning model, we aim to contribute to the early detection of this often-overlooked condition.   
Our work has a dual impact: First, it helps raise awareness of endometriosis among healthcare professionals and the public, emphasizing the critical need for timely diagnosis. Second, it has the potential to shorten the time to diagnosis, leading to earlier interventions and better patient outcomes.

In terms of practical application, our model can be integrated into clinical settings, serving as a valuable tool for healthcare professionals. By assisting in the early identification of endometriosis, the model could play a crucial role in reducing the diagnostic delay that so many patients currently experience, ultimately improving their quality of life.

Looking ahead, future work will focus on several key areas. We plan to explore the model's application to other related gynecological conditions, which could benefit from similar early detection approaches. Additionally, we recognize the importance of expanding our dataset to include a more diverse population. Doing so will enhance the model's generalizability and ensure that it is applicable to a broader range of patients, making it a more powerful tool in the fight against endometriosis and related conditions.

# Acknowledgments

We thank Adi Shraibman, Dorit Shweiki, and Yonatan Bilu for their mentorship and support throughout the project.  
We thank The Academic College of Tel Aviv-Yaffo for providing us access to the UK BioBank data.  
We thank Grammarly and ChatGPT for grammar and tone suggestions.

# Bibliography

[1] Wang, P.-H., Yang, S.-T., Chang, W.-H., Liu, C.-H., Lee, F.-K., & Lee, W.-L. (2022). Endometriosis: Part I. Basic concept. Taiwanese Journal of Obstetrics and Gynecology, 61(6), 927–934. <https://doi.org/10.1016/j.tjog.2022.08.002>

[2] Zondervan, K. T., Becker, C. M., Koga, K., Missmer, S. A., Taylor, R. N., & Viganò, P. (2018). Endometriosis. Nature reviews. Disease primers, 4(1), 9. <https://doi.org/10.1038/s41572-018-0008-5>

[3] Blass, I., Sahar, T., Shraibman, A., Ofer, D., Rappoport, N., & Linial, M. (2022). Revisiting the Risk Factors for Endometriosis: A Machine Learning Approach. Journal of Personalized Medicine, 12(7), 1114–1114. <https://doi.org/10.3390/jpm12071114>

[4] Tom Gunnar Tanbo, & Péter Fedorcsák. (2017). Endometriosis‐associated infertility: aspects of pathophysiological mechanisms and treatment options. Acta Obstetricia et Gynecologica Scandinavica, 96(6), 659–667. <https://doi.org/10.1111/aogs.13082>

[5] Horne, A. W., & Missmer, S. A. (2022). Pathophysiology, diagnosis, and management of endometriosis. BMJ (Clinical research ed.), 379, e070750. <https://doi.org/10.1136/bmj-2022-070750>

[6] UK BioBank (2015). Ukbiobank.ac.uk. <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us>

[7] Hong, J., & Yi, K. W. (2022). What is the link between endometriosis and adiposity?. Obstetrics & gynecology science, 65(3), 227–233. <https://doi.org/10.5468/ogs.21343>

[8] Mehedintu, C., Plotogea, M. N., Ionescu, S., & Antonovici, M. (2014). Endometriosis still a challenge. Journal of medicine and life, 7(3), 349–357. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4233437>

[9] Barnard, N. D., Holtz, D. N., Schmidt, N., Kolipaka, S., Hata, E., Sutton, M., Znayenko-Miller, T., Hazen, N. D., Cobb, C., & Kahleova, H. (2023). Nutrition in the prevention and treatment of endometriosis: A review. Frontiers in nutrition, 10, 1089891. <https://doi.org/10.3389/fnut.2023.1089891>

[10] Farland, L. V., Prescott, J., Sasamoto, N., Tobias, D. K., Gaskins, A. J., Stuart, J. J., Carusi, D. A., Chavarro, J. E., Horne, A. W., Rich-Edwards, J. W., & Missmer, S. A. (2019). Endometriosis and Risk of Adverse Pregnancy Outcomes. Obstetrics and gynecology, 134(3), 527–536. <https://doi.org/10.1097/AOG.0000000000003410>

[11] Afshari-Stasiak, S., Andrysiewicz, S., Andrysiewicz, A., Bamberska, J., Szubert, M. (2023). Rate of caesarean sections is higher in endometriosis patients – experience of single tertiary obstetric clinic in the light of epidemiologic data. J Pre Clin Clin Res., 17(2), 56-61. <https://doi.org/10.26444/jpccr/163481>

[12] Szypłowska, M., Tarkowski, R., & Kułak, K. (2023). The impact of endometriosis on depressive and anxiety symptoms and quality of life: a systematic review. Frontiers in public health, 11, 1230303. <https://doi.org/10.3389/fpubh.2023.1230303>