A Machine Learning Approach for Early Detection of Endometriosis Based on Symptom Experience

T. Lecourtois[1]

[1] Mines Saint Etienne

...

March 2024

Abstract

Endometriosis, a persistent gynecological condition impacting 5 to 10% of women in their childbearing years, frequently faces delayed identification, occurring 6 to 10 years following the emergence of symptoms. This paper endeavors to develop a user-friendly self-diagnosis tool that can estimate the likelihood of endometriosis solely based on reported symptoms. Employing machine learning techniques, we trained predictive models using questionnaire responses from two cohorts : those who had received a diagnosis and those who remained undiagnosed. This investigation aims to expedite the diagnostic process by directing women with a heightened probability of endometriosis toward additional medical evaluations.

Keywords

Endometriosis, Symptom-based diagnosis, Machine Learning, Feature selection

1 Introduction

Current diagnostic approaches for endometriosis are both invasive and costly, necessitating the exploration of non-invasive screening tools. While previous studies have investigated biomarkers, genomic data, and patient-reported symptoms, none have fully supplanted the need for laparoscopy.

This paper underscores the utilization of machine learning (ML) to craft a self-diagnostic tool centered solely around patient-reported symptoms. The objective is to develop a user-friendly model tailored for women in the initial stages of medical examination, offering an early indication of their potential endometriosis likelihood. The study pinpoints a set of 24 symptoms most effective in predicting endometriosis, achieving notable sensitivity (0.93) and specificity (0.93) on holdout data. The overarching goal is to expedite the diagnostic process and shed light on the significance of various symptoms in predicting endometriosis.

1.1 Related Work

In reviewing prior studies, several investigations have addressed the challenge of early detection of endometriosis, exploring various aspects ranging from biomarkers [1] to genomic models [2] and patient-reported symptoms [3]. Previous approaches often relied on invasive and costly methods [4], such as laparoscopy, to confirm the diagnosis of endometriosis. Despite progress in identifying specific biomarkers, these methods have yet to eliminate the need for invasive interventions.

2 Material and Methods

We applied several ML algorithms to train multiple endometriosis prediction models. Specifically, we applied decision trees, Random Forest and Logistic Regression. Besides generating predictions, these models also provide an importance analysis feature, which can be used to identify and remove non-contributing features from future surveys. Model performance was evaluated using common ML metrics : accuracy, sensitivity (recall), specificity, precision, F1-score, area under the ROC curve (AUC) and Matthew Coorelation Coefficient. To ensure significance of the results, we used a ten-fold cross-validation procedure.

2.1 Machine Learning Algorithms

Machine learning (ML) algorithms are instrumental in constructing accurate and reliable models for the early detection of endometriosis based on symptom experience. In this subsection, we provide a brief overview of three key ML algorithms utilized in our study : logistic regression, decision trees, and random forest.

2.1.1 Logistic Regression

Logistic regression is a powerful statistical method employed for binary classification tasks. In the context of endometriosis detection, logistic regression models the probability that an individual has endometriosis, leveraging a set of input features derived from reported symptoms. The algorithm estimates coefficients for each feature, providing valuable insights into the impact of specific symptoms on the likelihood of endometriosis.

2.1.2 Decision Trees

Decision trees are versatile non-linear models widely used for classification tasks. In our study, decision trees are employed to capture intricate relationships between symptoms and the likelihood of endometriosis. These tree-like structures recursively partition the dataset based on the

Metric	Mean	Std
Recall	0.9108	0.0437
Specificity	0.9234	0.0591
Precision	0.9318	0.0486
F1-score	0.9196	0.0268
Accuracy	0.9167	0.0291
AUC	0.9171	0.0297

TABLE 1 – Performance Metrics for Logistic Regression

most informative features, offering a transparent representation of the decision-making process.

Metric	Mean	Std
Recall	0.8919	0.0401
Specificity	0.8580	0.0672
Precision	0.8763	0.0497
F1-score	0.8827	0.0300
Accuracy	0.8757	0.0339
AUC	0.8750	0.0349

TABLE 2 – Performance Metrics for Decision Trees

2.1.3 Random Forest

Random forest, an ensemble learning method, combines the strength of multiple decision trees to enhance predictive accuracy and robustness. In the context of endometriosis detection, random forest models leverage a multitude of trees trained on different subsets of data and features. This ensemble approach allows the model to capture a broad spectrum of symptom interactions, contributing to improved overall performance.

2.2 Data collection

The dataset included 56 endometriosis symptoms that were compiled based on an extensive review of relevant literature. The dataset used in this project consists of 800 women examples, each containing the symptoms related to endometriosis. Of these, 474 had a diagnosis of endometriosis and 412 had no diagnosis, that is, did not undergo a diagnostic procedure. It is not a continuous dataset : each entry is labeled with a binary response (0 or 1) indicating the presence or absence of the respective symptom. We note that it is possible that some proportion of the undiagnosed women suffer from endometriosis but have not yet been diagnosed. Such respondents may introduce bias into our model and cause false negatives. Nevertheless, as the percentage of endometriosis is estimated between 5 and 10%, we expect such bias to be relatively small.

Metric	Mean	Std
Recall	0.8946	0.0426
Specificity	0.9350	0.0366
Precision	0.9390	0.0319
F1-score	0.9154	0.0258
Accuracy	0.9138	0.0256

TABLE 3 – Performance Metrics for Random Forest

2.3 Symptom importance analysis

To enhance the model's performance and avoid redundancy, we applied the Jaccard Index to the dataset. The Jaccard Index is a measure of similarity between two sets. In our context, it helps identify and eliminate redundant symptoms, ensuring a more concise and informative set of features. It is calculated as the size of the intersection of two sets divided by the size of their union. In the context of our dataset :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Here, A represents the set of symptoms in one entry, and B represents the set of symptoms in another entry. We used the Jaccard Index to iteratively evaluate different subsets of symptoms and determine the optimal number of informative features for our model.

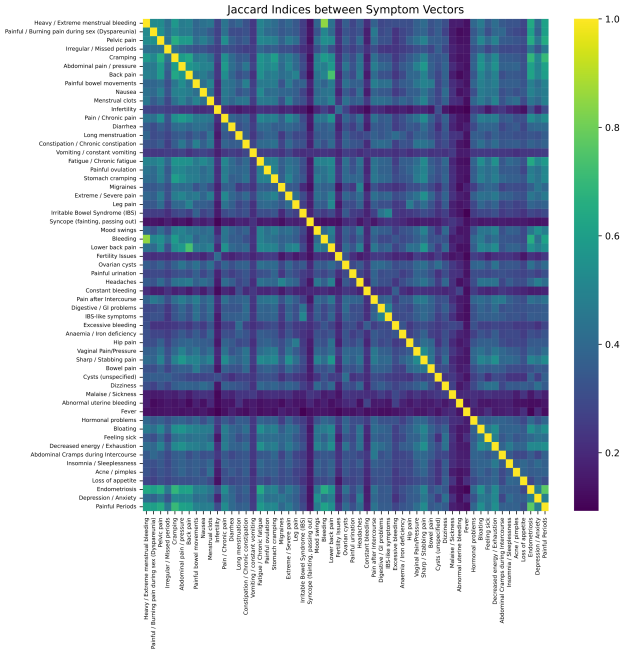

FIGURE 1 – Jaccard Index for Different Symptom Subsets. In this representation, the lighter the color, the higher the similarity between 2 symptoms.

3 Results

Based on the acquired outcomes, the Random Forest model has marginally superior performance. Our objective was to enhance the model's interpretability and potentially boost predictive accuracy by evaluating the importance of features. To achieve this, we focused on the most relevant variables through the assessment of feature importance. Additionally, we visualized the ROC Curve and the Confu-

sion Matrix and computed the Area Under the Curve (AUC) across different configurations of feature sets.

The Matthews Correlation Coefficient (MCC) is a metric commonly used to evaluate the performance of binary classification models. It takes into account true positive (T_P), true negative (T_N), false positive (F_P), and false negative (F_N) values, providing a balanced measure even in the presence of class imbalance. The MCC formula is given by :

$$MCC = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_P) \times (T_P + F_N) \times (T_N + F_P) \times (T_N + F_N)}}$$

The MCC value ranges from -1 to 1, where 1 indicates perfect classification, 0 indicates no better than random, and -1 indicates total disagreement between the prediction and the actual labels.

Model	MCC
Logistic Regression	0.75
Decision Trees	0.82
Random Forest	0.88

TABLE 4 – Matthews Correlation Coefficient (MCC) for Different Models

Additionally, in the context of the Random Forest model, we observe that beyond incorporating 15 features, the model's performance remains nearly constant, as indicated by the F1 score. Regarding the Area Under the Curve (AUC), a very slight increase is noted, albeit not substantial enough to be deemed statistically significant. This observation suggests that the inclusion of additional features beyond the 15th does not significantly contribute to the model's predictive capacity, at least in terms of the considered performance metrics.

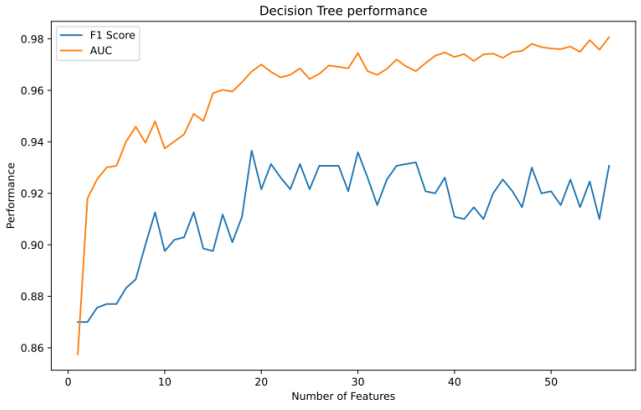

FIGURE 2 – F1 Score and AUC for the Random Forest Model

The Area Under the Curve (AUC) for the ROC curve provides a measure of the model's ability to distinguish between positive and negative instances. It represents the area

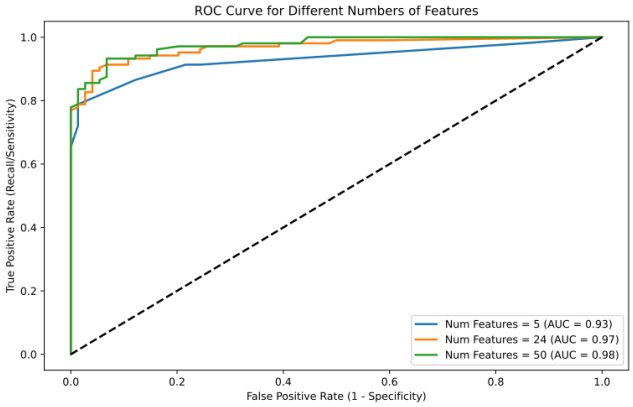

FIGURE 3 – ROC Curve for the Random Forest Model

under the ROC (Receiver Operating Characteristic) curve of the classification model. The ROC curve shows the relationship between the model's true positive rate T_P and false positive rate F_P at different thresholds. Notably, we observed on Figure 3 that increasing the number of features led to improved results, particularly for 50 features, achieving an AUC of 0.98. This observation contrasts with the common intuition and previous results that adding more features may decrease model precision. In our case, augmenting the feature set enhanced the model's discriminatory power.

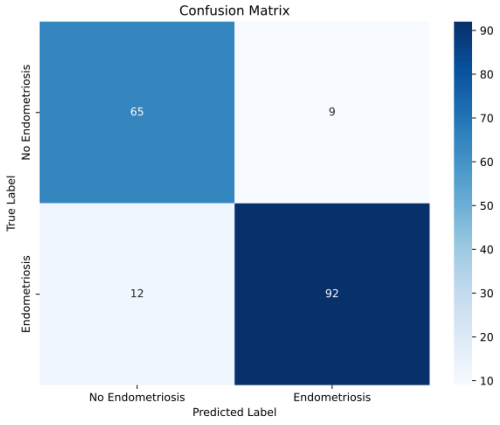

FIGURE 4 – Confusion Matrix for the Random Forest Model

Following the generation of the confusion matrix, which reveals 65 true positives, 92 true negatives, 12 false negatives, and 9 false positives, it is imperative to delve into a comprehensive analysis. Particularly noteworthy is the higher count of false negatives compared to false positives. In the domain of medical diagnostics, this outcome holds considerable significance.The elevated number of false negatives implies instances where the model fails to identify actual positive cases, potentially leading to undiagnosed individuals who are indeed afflicted by endometriosis. This phenomenon raises concerns about the model's sensitivity

and its ability to capture all positive cases within the dataset.

Conversely, the lower count of false positives is reassuring, as it signifies a reduced likelihood of misclassifying healthy instances as positive. While minimizing false positives is a positive aspect, the priority in medical applications often leans toward minimizing false negatives to ensure that individuals in need of attention are not overlooked.

4 Discussion and Future Work

In the quest for refining the diagnostic capabilities of our model, future investigations could benefit from exploring advanced machine learning algorithms, such as XGBoost. Renowned for its superior performance in medical domains, XGBoost offers robust predictive capabilities that may contribute to further enhancing our model's accuracy.

Références

[1] Zhang H, Zhang H, Yang H, Shuid AN, Sandai D, Chen X. Machine learning-based integrated identification of predictive combined diagnostic biomarkers for endometriosis. Front Genet. 2023 Nov 27 ;14 :1290036. doi : 10.3389/fgene.2023.1290036. PMID : 38098472 ; PMCID : PMC10720908.

[2] Urteaga, I., McKillop, M. & Elhadad, N. Learning endometriosis phenotypes from patient-generated data. npj Digit. Med. 3, 88 (2020). https ://doi.org/10.1038/s41746-020-0292-9

[3] Sivajohan, B., Elgendi, M., Menon, C. et al. Clinical use of artificial intelligence in endometriosis : a scoping review. npj Digit. Med. 5, 109 (2022). https ://doi.org/10.1038/s41746-022-00638-1

[4] Bendifallah, S., Puchar, A., Suisse, S. et al. Machine learning algorithms as new screening approach for patients with endometriosis. Sci Rep 12, 639 (2022). https ://doi.org/10.1038/s41598-021-04637-2