

Early prediction of endometriosis risk through explainable artificial intelligence

Magui Anca-Elena
Babeş-Bolyai University

WeADL 2026 Workshop

The workshop is organized under the umbrella of WinDMiL, project funded by CCCDI-UEFISCDI, project number [PN-IV-P7-7.1-PED-2024-0121](#), within PNCDI IV



- ➊ Introduction
- ➋ Problem Statement
- ➌ Related Work
- ➍ Proposed Methodology
- ➎ Experimental Results
- ➏ Comparison to Related Work
- ➐ Conclusion & Future Work

What is Endometriosis?

- A chronic gynecological pathology affecting **10% of women**.
- Severe pain, dysmenorrhea, and infertility, reducing quality of life.
- A disease where formal diagnosis currently relies on **invasive laparoscopy**.

Problem Statement & Objective

The Problem

- Diagnostic delay of **7–11 years**.
- Existing ML models rely on invasive biomarkers or imaging.
- Lack of early, accessible screening tools.

Our Objective

- Non-invasive screening via self-reported symptoms.
- Explainable ML early-warning system.
- A trustworthy tool for patients and clinicians.

Research Questions

- RQ1:** What is the optimal combination of feature selection methods and classification algorithms for endometriosis risk prediction based on self-reported symptoms?
- RQ2:** Can the selected model's predictions be explained in clinically interpretable terms, and do the identified risk factors align with established medical knowledge?
- RQ3:** How does the proposed model compare to existing machine learning approaches in the literature, both in terms of statistical performance and clinical utility?

- **Goldstein et al. [1]**: Baseline study (58 $\rightarrow$ 24 features). Used model-dependent wrapper methods; reported $F_1 > 0.93$.
- **Chapron et al. [2]**: Reduced 34 features to 5 using Logistic Regression. Lacks non-linear machine learning capabilities.
- **Nezhat et al. [3]**: AI-based EndoRA app. High sensitivity (93.1%) but extremely low specificity (5.9%) due to selection bias.

Proposed Methodology

- Consensus-based symptom reduction (from 58 to 35) while ensuring strict data isolation.
- Nested Cross-Validation and Bayesian optimization (Optuna) to prevent model overfitting.
- TreeSHAP explainability to guarantee transparent, dual-level risk predictions.
- Decision Curve Analysis to evaluate clinical utility and real-world net benefit.

Feature Selection

- **Dataset:** Goldstein et al. [1] (886 patients, 58 symptoms).
- Exclusive use of filter methods to prevent overfitting and algorithm bias.
- **Final Result:** 35 highly predictive symptoms (**39.7% reduction**).

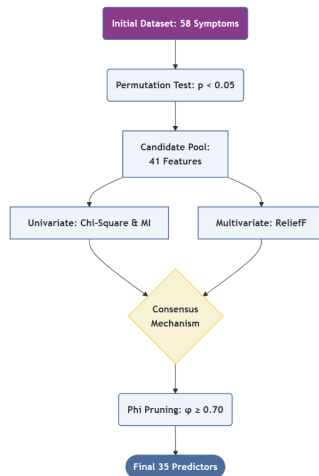


Figure: Proposed consensus-based feature selection pipeline.

Model Training & Optimization

- Comparative evaluation of 4 classifiers: SVM (RBF), Random Forest, XGBoost, and CatBoost.
- Generalization performance estimated using 10-fold **Nested Cross-Validation**.
- Bayesian hyperparameter optimization using the **Optuna** framework (100 trials).
- Optimal decision thresholds calculated via the **Youden Index** ($J = \text{Sensitivity} + \text{Specificity} - 1$) to maximize screening safety.

Explainability Framework

- Provides a unified framework for both **global** and **local** interpretability.
- Employs **TreeSHAP** for tree-based ensembles, with **KernelSHAP** as a model-agnostic fallback.
- Grounded in cooperative game theory: assigns a fair **Shapley value** reflecting each symptom's diagnostic contribution.
- Includes a global **surrogate decision tree** to approximate and visually summarize the model's overall logic.

Decision Curve Analysis (DCA)

- Quantifies real-world **Net Benefit** across clinical thresholds p_t :

$$NB(p_t) = \frac{TP}{N} - \frac{FP}{N} \cdot \frac{p_t}{1 - p_t}$$

- Compares the model against default “treat all” / “treat none” strategies

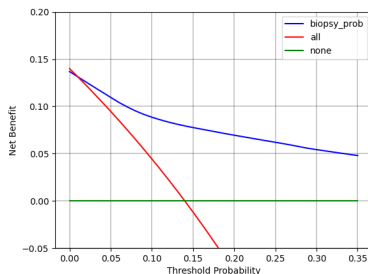


Figure: Illustrative DCA plot comparing a predictive model against default clinical strategies.

Experimental Results: UMAP

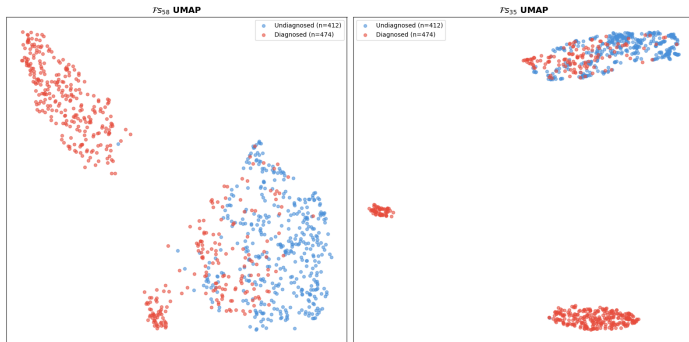


Figure: UMAP projection of $\mathcal{F}_{508}$ (left) vs. $\mathcal{F}_{35}$ (right).

- Three patient clusters suggesting distinct clinical phenotypes [5].
- **Sharper boundaries in $\mathcal{F}_{35}$** — noise elimination confirmed.

Experimental Results: Classification – RQ1

Table: Best configuration per model (35 features, hold-out test set)

Model	Config	AUC	AUPRC	F1	Sens.	Spec.	Prec.	Acc.
SVM (RBF)	Default	0.976	0.981	0.920	0.905	0.928	0.935	0.916
RF	Default + Youden	0.976	0.981	0.936	0.916	0.952	0.956	0.933
XGBoost	Default + Youden	0.977	0.982	0.918	0.884	0.952	0.955	0.916
CatBoost	Tuned + Youden	0.977	0.981	0.920	0.905	0.928	0.935	0.916

- **Selected Model:** RF ($t_{opt} = 0.538$) — default settings generalized better than Optuna-tuned ones.
- **Noise Removal:** Removing 23 features yielded $< 1\%$ AUC drop ($0.982 \rightarrow 0.976$).
- **Bootstrap** ($n = 1000$): $F_1 \in [0.897, 0.968]$, $AUC \in [0.956, 0.991]$ (95% CI).

Explainability: Global SHAP Analysis – RQ2

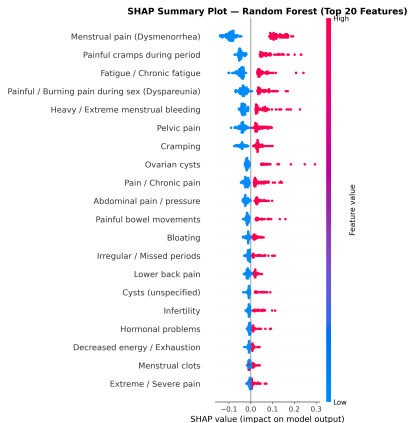


Figure: SHAP beeswarm plot for the selected model.

- **Top predictors:**
Dysmenorrhea, painful cramps, chronic fatigue, dyspareunia, heavy bleeding
- Findings align with established clinical guidelines [4]

Explainability: Surrogate Decision Tree

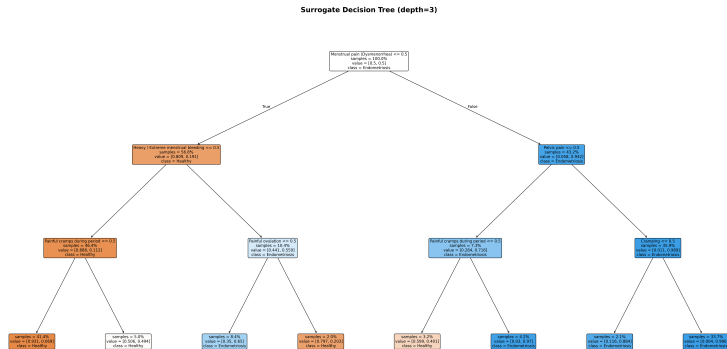


Figure: Depth-3 surrogate tree mimics the full Random Forest with 93.3% fidelity.

Explainability: Multi-Model Agreement

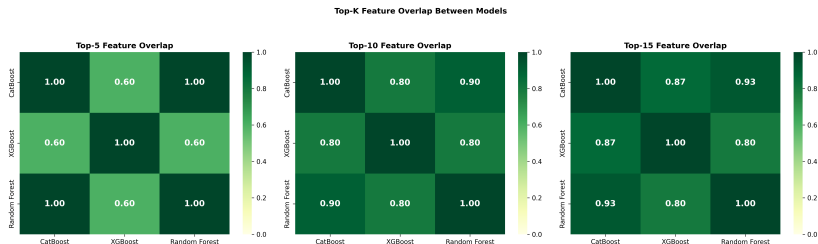


Figure: Top-K feature overlap between model pairs.

- CatBoost and RF: **100% overlap** in Top 5 predictors
- 80%–93% agreement across Top 10–15 features
- Confirms **true biological signals**, not model-specific artifacts

Clinical Validation: Decision Curve Analysis

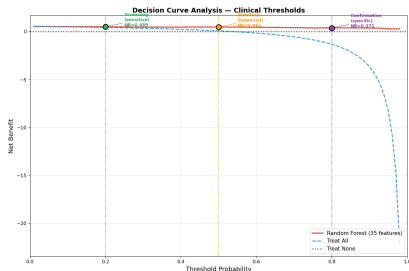
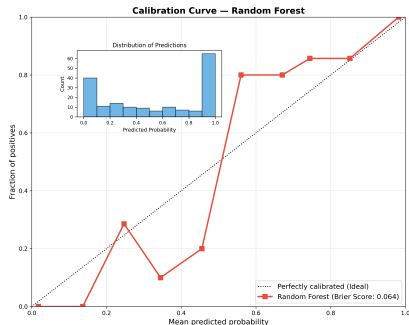


Figure: Net Benefit of the Random Forest model vs. treat-all and treat-none strategies.

- **Screening (20%):** NB = 0.489, +0.072 over treat all
- **Standard (50%):** NB = 0.461, equivalent to **39 extra true positives** per 100 patients
- **Confirmation (80%):** NB = 0.371, safely guiding surgical decisions

Clinical Validation: Calibration Curve



- Brier Score: **0.064**
- Predictions concentrated near 0 and 1
- No need for post-calibration

Comparison to Related Work – RQ3

Controlled re-evaluation under identical pipeline

Feature Set	Feat.	F1	Spec.	Acc.
Goldstein et al. [1]	24	0.905	0.892	0.899
This Work	35	0.936	0.952	0.933

- Re-evaluated Goldstein's features under our *identical, stricter protocol* for a fair comparison
- Our consensus pipeline: **+3.1% F1, +3.4% Accuracy**
- DCA at 50%: our model detects **~5 additional true positives** per 100 patients vs. Goldstein (39.3 vs. 34.8)

Response to Research Questions

- RQ1** ✓ Consensus filter selection + Youden-adjusted Random Forest yields optimal screening ($F1 = 0.936$, $Acc = 0.933$)
- RQ2** ✓ TreeSHAP provides clinically aligned explanations — top predictors (dysmenorrhea, fatigue, dyspareunia) match established guidelines [4]
- RQ3** ✓ +3.1% F1 over baseline under identical conditions; clinical value confirmed via DCA ($NB = 0.461$)

Conclusion

- Non-invasive, symptom-based screening to reduce diagnostic delays.
- Transparent, clinically verifiable logic to ensure Trustworthy AI.
- Ready for integration into patient-facing mobile applications.

- Prospective clinical trials across diverse populations.
- Multi-cycle temporal modeling (LSTM / Transformers).
- Wearable sensor integration (temperature, HRV).

Thank you for your attention!

References



Goldstein, A., et al. (2023).

Self-report symptom-based endometriosis prediction using machine learning
Scientific Reports, vol. 13, no. 1, p.5499.



Chapron, C., et al. (2022).

A new validated screening method for endometriosis diagnosis based on patient questionnaires
EClinicalMedicine, vol. 44, p. 101263.



Nezhat, C., et al. (2023).

Use of the Free Endometriosis Risk Advisor App as a Non-Invasive Screening Test for Endometriosis in Patients with Chronic Pelvic Pain and/or Unexplained Infertility.
Journal of Clinical Medicine, vol. 12, no. 16, p. 5234.



K. T. Zondervan, et.al. (2020).

Endometriosis.
New England Journal of Medicine,382(13):1244–1256.



Duncan, Jordann-Mishaël et.al. (2024),

Management of endometriosis: a call to multidisciplinary approach
Journal of osteopathic medicine,125(6), 305–313