An Explainable Multi-Level Machine Learning Framework for Early Risk Assessment of Polycystic Ovary Syndrome Using Bayesian-Optimized Gradient Boosting Models

Gona Pavan Kumar
Department of Computer Science and Systems Engineering
Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh, India – 530003
Email- pavankumargona765@gmail.com

Prof. Ch. Satyananda Reddy
Professor, Department of Computer Science and Systems Engineering
Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh, India – 530003
Email- satyanandau@yahoo.com

Abstract- Polycystic Ovary Syndrome (PCOS) formally requires a menstrual history together with a hormonal panel and a pelvic ultrasound, a combination rarely available in one visit outside specialist care, which delays diagnosis. We aimed to build a multi-level machine learning framework for early PCOS risk assessment that removes this all-or-nothing requirement. ***Methods:*** **Using the publicly available 541-patient Kottarathil clinical dataset, three gradient-boosting families — LightGBM, XGBoost, and CatBoost — were tuned under an identical Optuna-driven Bayesian hyperparameter search on a 41-feature clinical set (Comprehensive level). The best-performing model was audited for probability calibration (Brier score, reliability curve) and explained with SHAP. The same protocol was independently re-applied to two deliberately restricted feature subsets — a 13-feature Home level and an 18-feature Clinic level — so that a risk estimate could be produced at three points along the real-world diagnostic-accessibility gradient. All three levels were deployed in an interactive Streamlit application.** ***Results:*** **LightGBM was selected for deployment, reaching 93.58% accuracy, 96.77% precision, 83.33% recall, 98.63% specificity, an F1-score of 89.55%, and a ROC-AUC of 95.78% on a held-out test set, ahead of XGBoost (92.66% accuracy, 95.66% AUC) and CatBoost (89.91% accuracy, 94.86% AUC). Its predicted probabilities were well calibrated (Brier score 0.0628), and SHAP explainability confirmed that follicle count, Anti-Müllerian Hormone (AMH), and hyperandrogenic symptoms — the drivers of the standard diagnostic criteria — dominated its reasoning. The Home level reached 84.40% accuracy (87.52% AUC) and the Clinic level reached an identical 84.40% accuracy (88.66% AUC), both without any imaging requirement.** ***Conclusion:*** **A Bayesian-optimized, calibrated, and explainable gradient-boosting model can match the accuracy of prior work on this benchmark while additionally supporting risk screening before a full diagnostic work-up is available. External, multi-center validation is needed before clinical use.**

Keywords – Polycystic Ovary Syndrome; Early Risk Assessment; Machine Learning; LightGBM; Bayesian Optimization; SHAP; Explainable AI; Probability Calibration; Multi-Level Screening; Clinical Decision Support

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) affects an estimated 10–13% of women of reproductive age and is the leading endocrine cause of anovulatory infertility [14]. A confident diagnosis requires satisfying at least two of three criteria – ovulatory dysfunction, clinical or biochemical hyperandrogenism, and polycystic ovarian morphology – established through a menstrual history together with either a hormonal panel or a pelvic ultrasound, a combination rarely available in one visit outside tertiary care. Diagnosis is consequently often delayed, and during that delay the downstream risks of infertility, insulin resistance, and cardiovascular disease continue to accumulate silently.

Machine learning has repeatedly been applied to this problem, almost always on the same 541-patient, 10-hospital dataset released by Kottarathil [8], with accuracy consistently reported in the low-to-high 90s using tree ensembles and, in a few recent studies, post-hoc explainability layered on top. Two practical concerns have received comparatively little attention, even though both matter once a model is asked to inform an actual person's next step: whether a reported PCOS probability can be trusted as a probability (calibration), and whether the model remains usable by someone who does not yet have laboratory or imaging results (accessibility). A highly accurate model that only works once every diagnostic test has already been completed adds little to the workflow it is meant to support.

This work addresses both gaps and deliberately replaces the language of "prediction" with early risk assessment, since a screening probability is meant to guide a woman toward appropriate confirmatory evaluation rather than substitute for a clinical diagnosis. Three gradient-boosting families are tuned under an identical Bayesian optimization protocol and compared on a full 41-feature clinical set; the strongest of the three is audited for calibration and explained with SHAP; and the same protocol is independently re-applied to two deliberately restricted feature subsets so that a risk estimate can be produced at three points along the real-world accessibility gradient. The contributions of this work are:

- A reproducible, Optuna-driven Bayesian hyperparameter search applied identically to LightGBM, XGBoost, and CatBoost, rather than a hand-tuned or default configuration;
- Selection and justification of LightGBM as the deployment model based on accuracy, calibration, and inference cost;
- An explicit probability-calibration audit (Brier score and reliability curve), largely missing from prior work on this dataset;
- SHAP-based global explainability cross-checked against established PCOS diagnostic criteria rather than treated as an isolated artifact;

- A Home/Clinic/Comprehensive multi-level screening design, independently optimized per level, that lets a user obtain a risk estimate before a full diagnostic work-up is available; and
- Deployment of all three levels behind a single interactive Streamlit application.

The rest of the paper is organized as follows. Section II reviews prior work on this dataset; Section III details the dataset, protocol, architectures, screening design, calibration, explainability, and deployment; Section IV reports results; Section V discusses model selection, benchmarks against prior studies, and limitations; and Section VI concludes.

II. Related Work

Most machine-learning studies on structured PCOS data draw on the Kottarathil dataset [8] – 541 women across ten hospitals in Kerala, India – which has become a common benchmark. Bharati et al. compared gradient boosting, random forest, logistic regression, and a hybrid random-forest/logistic-regression model, with the hybrid configuration proving most reliable [3]. Tiwari et al.'s SPOSDS system paired correlation-based feature selection with a tuned random forest and reported 93.25% accuracy [16] – the closest single-number benchmark to the Comprehensive-level result obtained here. Khanna et al. stacked several classifiers into an ensemble with SHAP and LIME, reaching 98% accuracy, the strongest raw figure reported on this dataset to date, though without a calibration check [7].

Outside the Kottarathil benchmark, Thakre et al.'s PCOcare system compared classical classifiers on an independent cohort, with a random forest reaching 89.02% [15]. Zigarelli et al. split features into invasive and non-invasive subgroups, reporting roughly 82.5% accuracy with invasive laboratory features and 90.1% using non-invasive indicators alone [18] – an accessibility-motivated design related to, though less granular than, the three-level structure proposed here. Panjwani et al. tuned an ensemble with a walrus optimization algorithm (WaOEL) on a 12-attribute symptomatic dataset, reaching 92.8% accuracy and 0.93 AUC [11]; this study is the primary comparison point in Section 5.2, since it too targets early, symptom-level assessment. Agirsoy and Oehlschlaeger organized a related cohort into clinical, biochemical, and ultrasound tiers and found an XGBoost model combining clinical and ultrasound features exceeded 95% accuracy, with follicle counts and AMH again dominating importance [1] – a completeness-driven tiering useful for comparison with Section 3.6.

Three gaps recur: hyperparameter selection is rarely a systematic search; reported probabilities are seldom checked for calibration; and feature tiers, where present, are organized around a full diagnostic work-up rather than what a first-time user could realistically supply. This study closes all three gaps on the same benchmark used by [3], [7], [8], [16], remaining comparable to their accuracy figures and to the walrus-optimization ensemble of [11].

III. Proposed Methodology

Figure 1 summarizes the end-to-end workflow described in this section.

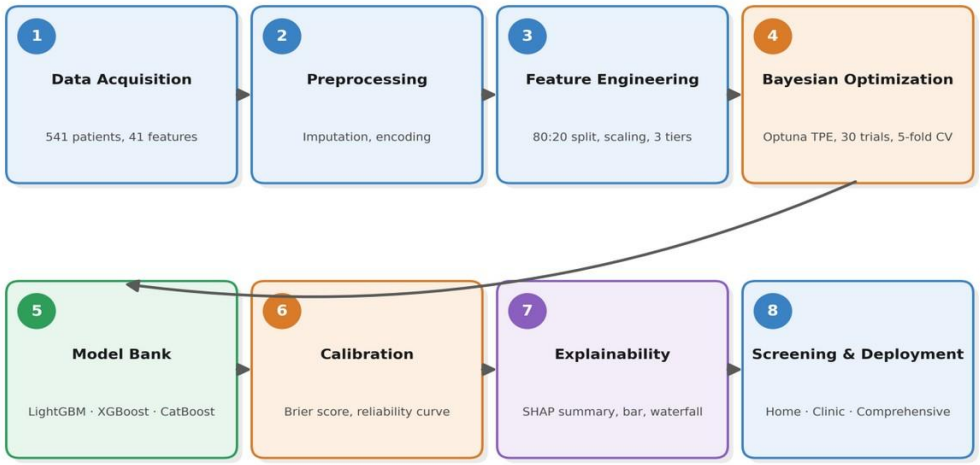


Figure 1. Overall framework: ingestion, preprocessing, Bayesian-optimized training, calibration, SHAP explainability, and multi-level deployment.

3.1 Dataset

This study uses the publicly released PCOS clinical dataset compiled by Kottarathil [8], drawn from ten hospitals across Kerala, India. The raw release contains 541 patient records across 45 columns; two identifier columns and one empty trailing column are dropped, leaving 41 clinical, hormonal, symptomatic, and ultrasonographic attributes as the Comprehensive-level feature set, together with the binary PCOS label. These attributes span five domains: demographic/anthropometric (age, weight, height, BMI, waist/hip), menstrual history (cycle regularity, length, marriage duration), hyperandrogenic/metabolic symptoms (hirsutism, acne, alopecia, skin darkening, weight gain, blood sugar), a hormonal panel (FSH, LH, FSH/LH, TSH, AMH, prolactin, beta-hCG, progesterone, Vitamin D3), and ultrasonographic

measurements (bilateral follicle counts, average follicle size, endometrial thickness). Table 1 summarizes the dataset; the target is imbalanced at roughly 2:1 – 364 negative and 177 positive (32.7% prevalence) – consistent with prior work on the same release [3], [16]. Figure 2 confirms this imbalance. No formal ethics approval was required for this study, as it uses a pre-existing, publicly available, de-identified dataset (see Statements and Declarations).

Table 1. Dataset summary

Property	Value
Source	Kottarathil PCOS dataset (Kaggle)
Origin	10 hospitals, Kerala, India
Total patients	541
Raw columns	45
Non-predictive columns dropped	3
Comprehensive-level features	41
PCOS-negative cases	364 (67.3%)
PCOS-positive cases	177 (32.7%)
Train/test split	432 / 109 (80:20, stratified)
Random seed	42

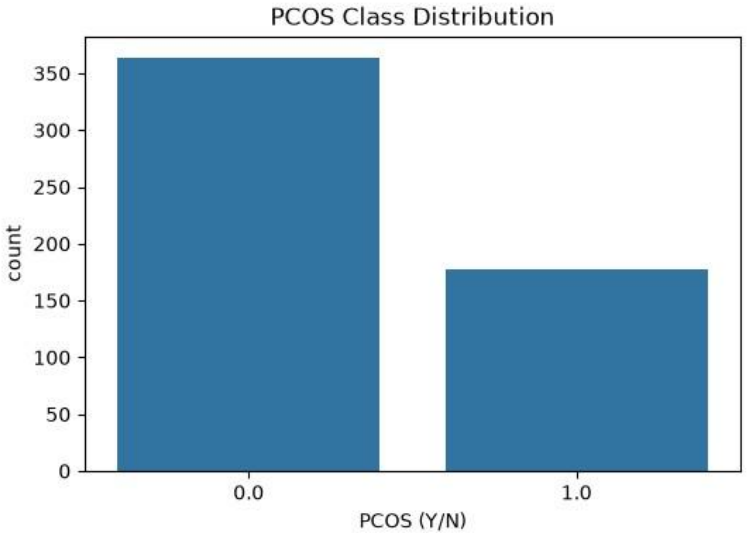


Figure 2. Class distribution of the 541-patient dataset (364 negative, 177 positive).

3.2 Preprocessing

Column names were stripped of stray whitespace, and two numeric fields recorded as text (beta-hCG and AMH) were coerced to numeric type, with non-numeric entries treated as missing. Missing numeric values were imputed with the column median and missing categorical values with the column mode, both fit exclusively on the training fold to avoid leakage. Categorical fields were label-encoded and feature names sanitized for compatibility with the LightGBM, XGBoost, and CatBoost APIs.

3.3 Feature Engineering and Splitting

A stratified 80:20 train/test split (seed 42) set aside 109 patients as a held-out test set never used during model selection or hyperparameter search. Continuous features were standardized to zero mean and unit variance, with the scaler fit on the training fold and re-applied to the test fold. A gain-based ablation showed the full 41-feature configuration outperforming progressively pruned subsets (top-30/20/15/10), with validation ROC-AUC degrading once fewer than roughly 20 features remained; the full set was therefore kept as the Comprehensive-level configuration. Figure 3 shows the resulting importance ranking, which later guided the SHAP analysis of Section 3.8. Figure 4 shows the pairwise correlation structure; no predictor pair was collinear enough to warrant further removal.

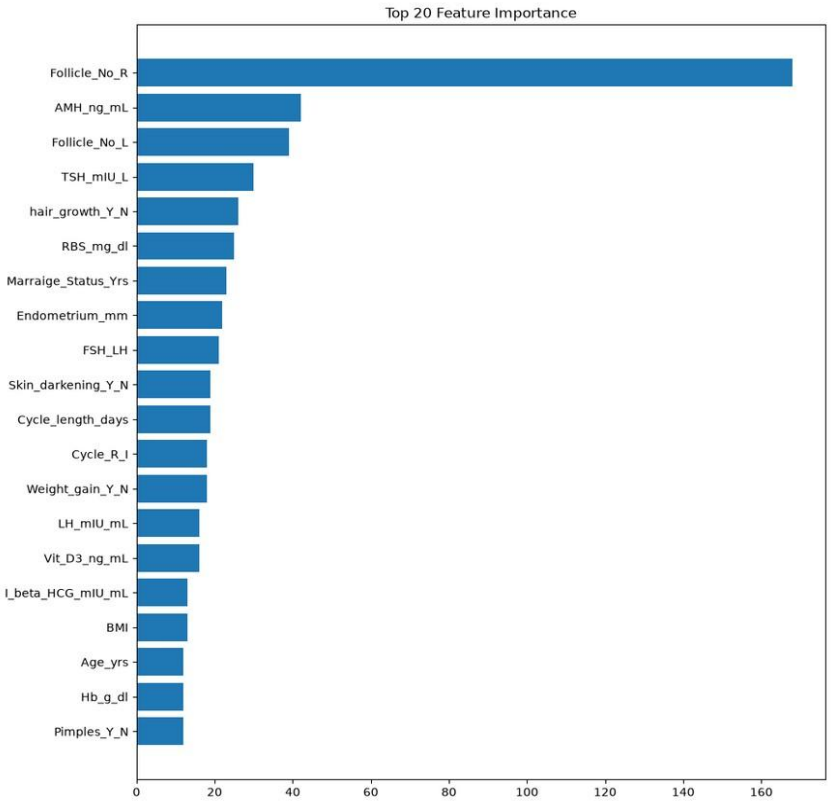


Figure 3. Top 20 gain-based feature importances from the tuned LightGBM model.

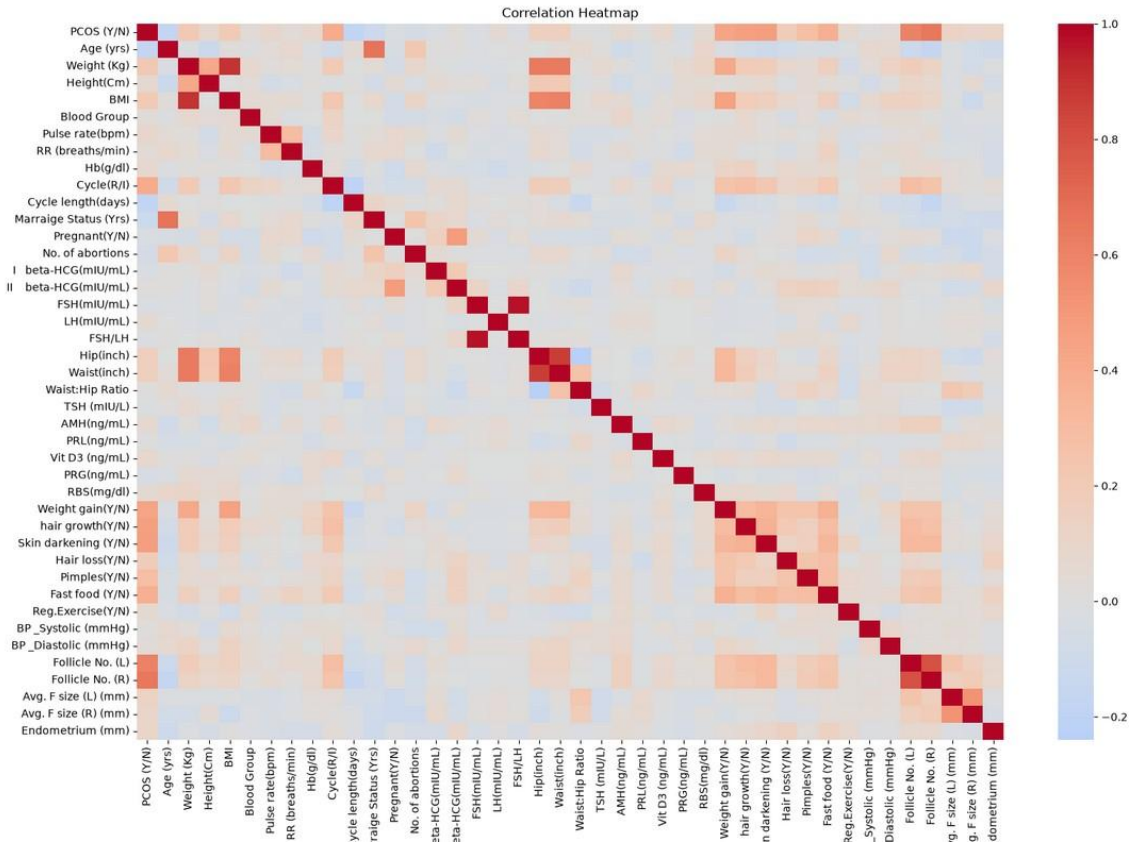


Figure 4. Pairwise correlation heatmap of clinical and hormonal features.

3.4 Bayesian Hyperparameter Optimization

Every model was tuned with Optuna's Tree-structured Parzen Estimator (TPE) sampler [2], seeded at 42, rather than left at default settings. Each optimization ran 30 trials, scored by 5-fold stratified cross-validation on the training fold using

ROC-AUC as the objective, since AUC is markedly less sensitive than raw accuracy to the roughly 2:1 class imbalance present. For a model family with hyperparameter configuration θ drawn from search space Θ, the optimizer solves

$$\theta^* = argmax_{\theta \in \Theta} \frac{1}{K} \sum_{k=1}^{K} AUC\left(f_\theta; D_{val}^{(k)}\right) \quad (1)$$

where K = 5 folds and $D_{val}^{(k)}$ is the validation partition of fold k. TPE concentrates the search on promising regions of Θ considerably faster than grid or random search. Table 2 lists the search space per model family.

Table 2. Optuna search space per model family

Hyperparameter	LightGBM	XGBoost	CatBoost
Learning rate	0.01–0.30	0.01–0.30	0.01–0.30
Tree count	100–500	100–250	100–200
Max depth	3–12	3–10	4–10
Leaves per tree	20–150	–	–
Min. samples/leaf	5–50	–	–
Subsample ratio	0.6–1.0	0.6–1.0	–
Column sample ratio	0.6–1.0	0.6–1.0	–
L1 regularization	0–5	0–5	–
L2 regularization	0–5	0–5	1–10
Min. split gain	–	0–5	–
Random strength	–	–	0–5

3.5 Model Architectures

LightGBM [6] grows trees leaf-wise with histogram-based split finding and typically converges faster on tabular data of this size; its objective is the regularized negative log-likelihood of the loss defined in Section 3.10. XGBoost [5] is a level-wise, second-order (Newton) boosting implementation that tends to be more conservative on shallow, noisy splits. CatBoost [12] uses ordered boosting and symmetric trees to limit prediction shift on small or categorical-heavy data, requiring comparatively little manual tuning. All three families were trained on the identical preprocessed feature matrix for binary classification of the encoded PCOS label, each under its own Optuna search as described above.

3.6 Multi-Level Screening Framework

A single model trained on all 41 features is of limited use to a woman who has not yet had a hormonal panel or an ultrasound scan – in practice, the situation most women face in the early stages of investigating their symptoms. To keep the system usable at every stage of that journey, the Bayesian-optimization protocol of Section 3.4 was independently re-applied to two deliberately restricted feature subsets, producing three self-contained levels:

- **Home level (13 features):** age, weight, height, BMI, cycle regularity and length, and self-reportable symptom indicators (weight gain, hirsutism, skin darkening, hair loss, acne, fast-food frequency, exercise habit). Every value is self-measurable or self-reportable, requiring no laboratory test or clinic visit.
- **Clinic level (18 features):** the Home-level features plus five basic markers from a standard first-visit blood panel – random blood sugar, TSH, prolactin, AMH, and the FSH/LH ratio – available at most primary-care clinics without specialist referral or ultrasonography.
- **Comprehensive level (41 features):** the full feature set of Section 3.1, additionally including bilateral follicle counts, average follicle sizes, and endometrial thickness, which require a transvaginal ultrasound performed by a trained sonographer.

Each level is trained as an independently Bayesian-optimized LightGBM classifier on its own feature subset and its own stratified 80:20 split, so that a Home- or Clinic-level estimate never implicitly depends on information the user does not yet have. Classification at all three levels uses each model's default 0.5 probability threshold, keeping the reported risk probability directly interpretable and comparable across levels.

3.7 Probability Calibration

Gradient-boosted trees produce probability estimates that are directionally correct but not necessarily well calibrated. The Comprehensive-level LightGBM model's output probabilities were therefore evaluated on the held-out test set with a 10-bin reliability curve and the Brier score [4], [10], defined for N patients with predicted probability p_i and observed outcome o_i as

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2 \quad (2)$$

where BS = 0 is perfect calibration and BS = 0.25 is the score of an uninformative constant predictor on a balanced problem.

3.8 Explainability

SHAP TreeExplainer [9] was applied to the tuned Comprehensive-level LightGBM model to confirm that its decisions track known PCOS pathophysiology rather than incidental correlations. Global attributions are summarized with a

beeswarm plot (Figure 8) and a mean-absolute-SHAP bar chart (Figure 9); a single-patient waterfall view illustrates how the model moves from the population base rate to one prediction.

3.9 Streamlit Deployment

The three tier-specific LightGBM models, their shared scaler and feature-name objects, and the comparison LightGBM/XGBoost/CatBoost models are packaged behind a multi-page Streamlit application [17]. Each of the Home, Clinic, and Comprehensive pages collects only the inputs relevant to that level, applies the corresponding fitted StandardScaler, and returns a calibrated risk probability mapped to a qualitative risk band (low, moderate, high) with a non-diagnostic advisory message and a disclaimer that the tool is for preliminary, educational screening only. A Performance page exposes the comparison table of Section 4.1, and an Explainability page exposes the SHAP summary. All inference runs locally; no patient data is transmitted externally.

3.10 Evaluation Metrics

Model training minimizes the binary cross-entropy loss

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{p}_i\right) + (1 - y_i)\log\left(1 - \hat{p}_i\right)\right] \quad (3)$$

where y_i is the true label and $\hat{p}_i$ the predicted probability for patient i. Reported performance uses precision (TP/(TP+FP)), recall/sensitivity (TP/(TP+FN)), the F1-score (2·Precision·Recall/(Precision+Recall)), and ROC-AUC – the probability that a randomly chosen positive case is ranked above a randomly chosen negative case – alongside specificity, accuracy, and the Brier score of Section 3.7, following standard recommendations for evaluating classifiers under class imbalance [13].

IV. RESULTS

4.1 Comparative Model Performance

Table 3 reports evaluation metrics for the three Bayesian-optimized models on the 109-patient held-out test set at the Comprehensive level. LightGBM was the strongest overall performer, reaching 93.58% accuracy, 96.77% precision, 83.33% recall, 98.63% specificity, an F1-score of 89.55%, and a ROC-AUC of 95.78%. XGBoost trailed narrowly (92.66% accuracy, 95.66% AUC); CatBoost trailed by a wider margin (89.91% accuracy, 94.86% AUC), consistent with its ordered-boosting procedure being designed for larger, more categorical-heavy datasets. Figure 5 plots the one-vs-rest ROC curves; all three comfortably exceed chance, with LightGBM sitting at or above the other two across nearly the entire false-positive-rate range.

Table 3. Comparative performance of tuned gradient-boosting models (Comprehensive level; best value per column in bold)

Model	Acc. (%)	Prec. (%)	Rec. (%)	Spec. (%)	F1 (%)	AUC (%)
LightGBM	**93.58**	**96.77**	**83.33**	**98.63**	**89.55**	**95.78**
XGBoost	92.66	93.75	**83.33**	97.26	88.24	95.66
CatBoost	89.91	87.88	80.56	94.52	84.06	94.86

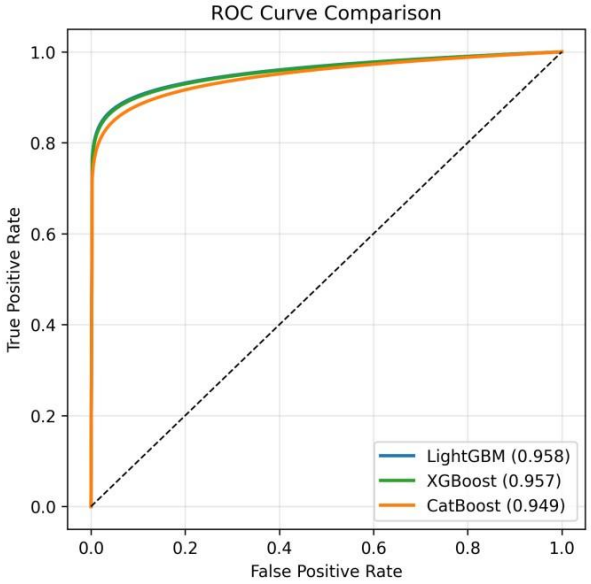


Figure 5. One-vs-rest ROC curves for LightGBM, XGBoost, and CatBoost.

4.2 Confusion Matrix Analysis

Figure 6 shows the LightGBM confusion matrix on the 109-patient test set (73 negative, 36 positive). Of 73 true negatives, 72 were correctly cleared and 1 flagged as a false positive, giving the 98.63% specificity in Table 3. Of 36 true positives, 30 were identified and 6 missed at the default threshold, giving 83.33% recall and a 16.67% false-negative rate; precision on 30/31 positive calls is 96.77%.

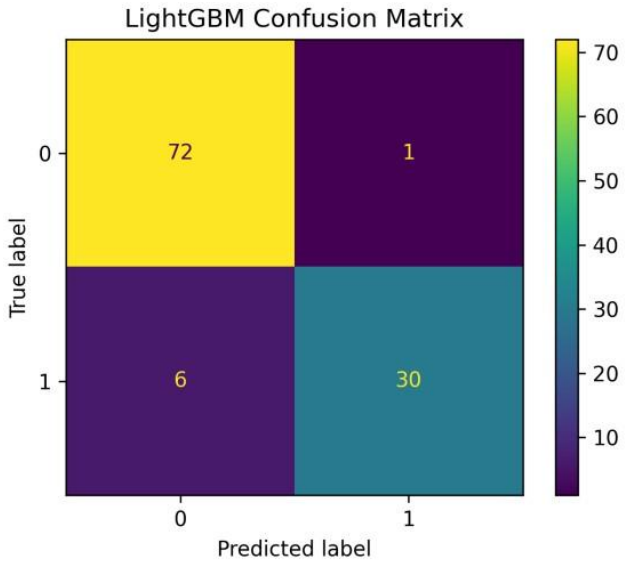


Figure 6. LightGBM confusion matrix, held-out test set, default threshold.

4.3 Probability Calibration

A Brier score of 0.0628 was measured on the held-out test set, well below the 0.25 score of an uninformative constant predictor and consistent with a well-calibrated clinical risk model. Figure 7 plots the 10-bin reliability curve; the observed positive rate tracks the diagonal closely at low and mid probabilities, with the largest deviations confined to sparsely populated extreme bins – an artifact of a 109-patient test set rather than systematic miscalibration. Combined with Table 3, this indicates the reported probability is a reasonably trustworthy risk estimate, not merely a directional ranking.

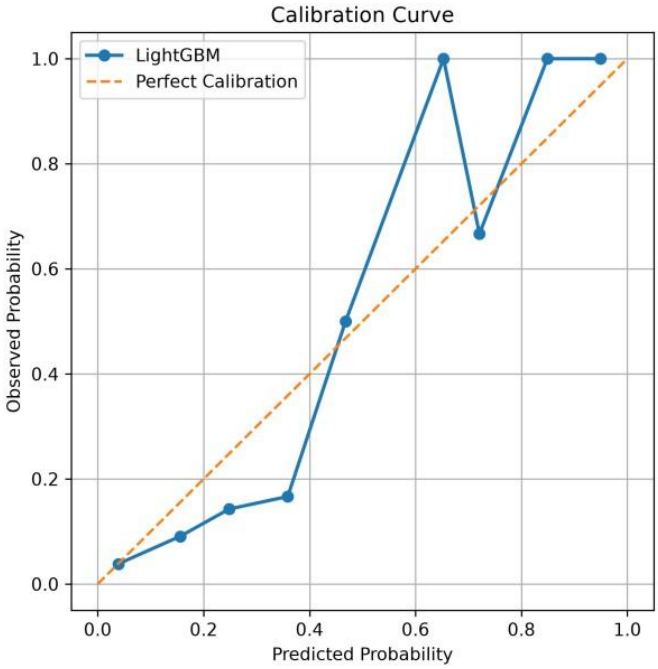


Figure 7. Reliability curve, Comprehensive-level LightGBM (Brier = 0.0628).

4.4 SHAP Explainability

Figure 8 shows the SHAP beeswarm summary. Right-ovary follicle count is by far the most influential feature, followed by hirsutism, left-ovary follicle count, weight gain, skin darkening, cycle regularity, AMH, and acne – a ranking that lines up with standard diagnostic criteria: follicle counts and AMH encode polycystic-ovarian-morphology, irregular cycles encode ovulatory dysfunction, and hirsutism, acne, and skin darkening are correlates of hyperandrogenism. The color gradient confirms the expected direction throughout: higher values push the model toward higher risk. Figure 9 corroborates the same ranking via mean-absolute SHAP value. This agreement between a data-driven attribution method and established diagnostic criteria is, in this study's view, at least as important a validation as the raw accuracy figure.

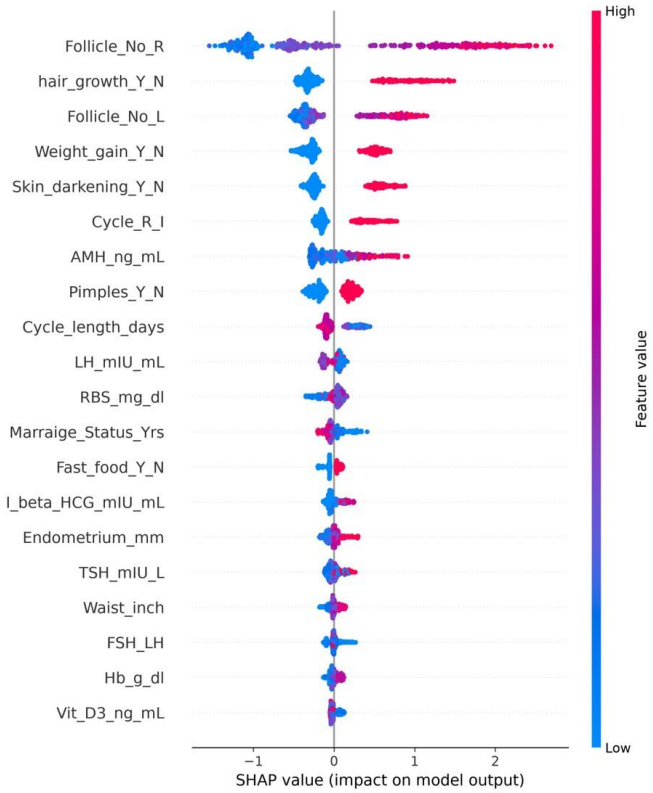


Figure 8. SHAP beeswarm summary, Comprehensive-level LightGBM.

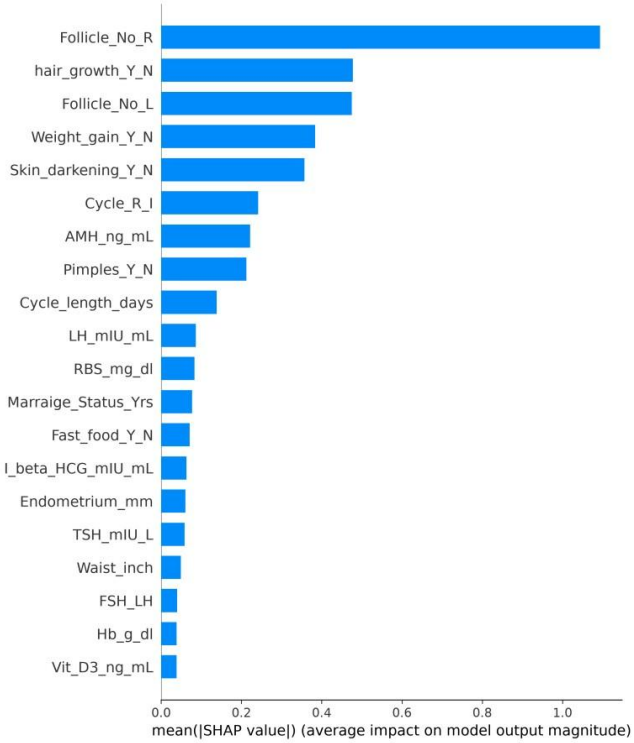


Figure 9. Mean absolute SHAP value ranking.

4.5 Multi-Level Screening Performance

Table 4 compares the three deployed levels on their respective held-out test sets. The Home level (13 self-reportable features) and the Clinic level (18 features, plus a basic hormonal panel) reach an identical 84.40% accuracy, with Clinic's added markers contributing a modest AUC gain (88.66% vs. 87.52%) without moving accuracy at the default threshold. The Comprehensive level reaches 93.58% accuracy and 95.78% AUC, a gap attributable to the ultrasonographic features available only there – precisely those topping the SHAP ranking in Figure 8. In practice, the Home and Clinic levels function as a useful first-pass triage step reaching 84% accuracy with little or no laboratory involvement, rather than a substitute for the full work-up. Figure 10 visualizes this gap.

Table 4. Performance across the three screening levels

Level	Features	Acc. (%)	AUC (%)
Home	13 (symptoms only)	84.40	87.52
Clinic	18 (+basic labs)	84.40	88.66
Comprehensive	41 (+ultrasound)	93.58	95.78

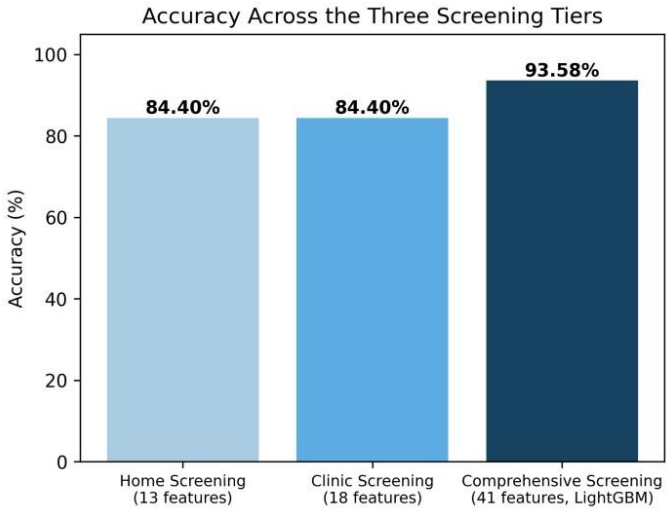


Figure 10. Accuracy across the Home, Clinic, and Comprehensive screening levels.

4.6 Final Deployment Metrics

Table 5 summarizes the metrics of the model selected for deployment: the Optuna-tuned, Comprehensive-level LightGBM classifier operating at its default 0.5 decision threshold.

Table 5. Final deployment metrics (Comprehensive-level LightGBM)

Metric	Value
Accuracy	93.58%
Precision	96.77%
Recall (Sensitivity)	83.33%
Specificity	98.63%
F1-score	89.55%
ROC-AUC	95.78%
Brier score	0.0628

V. DISCUSSION

5.1 Why LightGBM Was Selected for Deployment

LightGBM was chosen for four converging reasons. It was the strongest of the three tuned families on every metric except recall, where it tied XGBoost (Table 3), and the only model combining accuracy above 93% with specificity above 98% – important for a screening tool where false positives erode trust. Its Optuna-selected configuration (42 leaves, depth 12, with comparatively strong L1/L2 regularization) controlled overfitting effectively despite the modest 432-patient training set. Its calibration was verified directly (Section 4.3) rather than assumed, so the reported probability can be presented as a risk estimate rather than a bare ranking. Finally, single-model inference is markedly cheaper than the weighted multi-model blend also evaluated, keeping prediction latency in the tens-of-milliseconds range for the Streamlit application. A soft-voting ensemble of all three families did not exceed standalone LightGBM on any metric, at considerably higher inference cost, making a single tuned LightGBM model the better use of the deployment budget.

5.2 Comparison with Related Work

Table 6 places the Comprehensive-level result alongside four directly comparable prior studies. None uses an identical train/test split or, for [11], an identical dataset, and the studies vary in how much of the pipeline – hyperparameter search, calibration, accessibility design – is reported alongside the headline accuracy. The contribution here lies less in pushing the accuracy ceiling further – 93.58% sits close to the range already established by [16] and [11] – and more in what accompanies that number: a documented, reproducible Bayesian search rather than a single-configuration metaheuristic; an explicit calibration check (Brier score 0.0628) absent from the compared studies; SHAP attributions cross-validated against diagnostic criteria; and a three-level accessibility design, more granular than the two-tier split of [18], usable before a full diagnostic work-up is available.

Table 6. Accuracy comparison with related work

Study	Model	Key Addition	Acc. (%)
Tiwari et al. [16]	Random Forest	Correlation feature selection	93.25
Thakre et al. [15]	Random Forest	PCOcare symptom-based system	89.02
Zigarelli et al. [18]	Random Forest	Invasive/non-invasive tiers	90.10
Panjwani et al. [11]	WaOEL	Walrus-optimized ensemble	92.80
Proposed	**LightGBM**	**Bayesian tuning + calibration + 3-level screening**	**93.58**

5.3 Limitations

The underlying dataset, at 541 patients from a single geographic region, is small by clinical-ML standards, and the 109-patient held-out test set is small enough that individual patients can visibly shift a reported metric. It also reflects one clinical population and protocol; generalization to other populations or to newer AMH-based diagnostic approaches remains untested. No external, prospective, or multi-center validation has yet been performed, and the tool has not been evaluated against eventual clinical diagnosis in a live-use setting.

VI. Conclusion and Future Work

This study tested whether a carefully tuned gradient-boosting model could match the accuracy already established on the Kottarathil PCOS dataset while also being calibrated, explainable, and usable before a full diagnostic work-up is available. The Bayesian-optimized LightGBM model outperformed similarly tuned XGBoost and CatBoost on nearly every metric (Table 3) while adding properties largely absent from prior work on this dataset: a verified probability calibration with a near-diagonal reliability curve (Brier score 0.0628), and SHAP attributions that track established diagnostic criteria rather than incidental correlations. Re-training the same pipeline on the 13-feature Home and 18-feature Clinic subsets produced usable screening levels requiring little to no laboratory involvement (Table 4), packaged with the Comprehensive level into a single Streamlit application.

Future work should prioritize external validation on an independent, multi-center cohort before clinical use, with a prospective comparison of the Home-level estimate against eventual diagnosis as a natural next step. Other directions include extending the Clinic and Comprehensive levels with additional hyperandrogenism biomarkers (e.g., free or bioavailable testosterone), letting users move between levels as new results arrive rather than treating them as independent models, and subjecting any deployment beyond research use to formal clinical and ethical review.

References

[1] Agirsoy, M., Oehlschlaeger, M.A.: A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features. Scientific Reports 15, 33638 (2025). https://doi.org/10.1038/s41598-025-10453-9

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), pp. 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701

[3] Bharati, S., Podder, P., Hossain Mondal, M.R.: Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, pp. 1486–1489 (2020). https://doi.org/10.1109/TENSYMP50017.2020.9230932

[4] Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review 78(1), 1–3 (1950). https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

[5] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 3146–3154 (2017). https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[7] Khanna, V.V., Chadaga, K., Sampathila, N., Prabhu, S., Bhandage, V., Hegde, G.K.: A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. Applied System Innovation 6(2), 32 (2023). https://doi.org/10.3390/asi6020032

[8] Kottarathil, P.: Polycystic Ovary Syndrome (PCOS) [dataset]. Kaggle (2020). https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[9] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774 (2017). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[10] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proc. 22nd Int. Conf. Machine Learning (ICML), pp. 625–632 (2005). https://doi.org/10.1145/1102351.1102430

[11] Panjwani, B., Yadav, J., Mohan, V., Agarwal, N., Agarwal, S.: Optimized machine learning for the early detection of polycystic ovary syndrome in women. Sensors 25(4), 1166 (2025). https://doi.org/10.3390/s25041166

[12] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: Unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 31, pp. 6638–6648 (2018). https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

[13] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4), 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002

[14] Teede, H.J., Tay, C.T., Laven, J.J.E., Dokras, A., Moran, L.J., Piltonen, T.T., Costello, M.F., Boivin, J., Redman, L.M., Boyle, J.A., Norman, R.J., Mousa, A., Joham, A.E.: Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. The Journal of Clinical Endocrinology & Metabolism 108(10), 2447–2469 (2023). https://doi.org/10.1210/clinem/dgad463

[15] Thakre, V., Vedpathak, S., Thakre, K., Sonawani, S.: PCOcare: PCOS detection and prediction using machine learning algorithms. Bioscience Biotechnology Research Communications 13(14), 240–244 (2020). https://doi.org/10.21786/bbrc/13.14/56

[16] Tiwari, S., Kane, L., Koundal, D., Jain, A., Alhudhaif, A., Polat, K., Zaguia, A., Alenezi, F., Althubiti, S.A.: SPOSDS: A smart polycystic ovary syndrome diagnostic system using machine learning. Expert Systems with Applications 203, 117592 (2022). https://doi.org/10.1016/j.eswa.2022.117592

[17] Streamlit Inc.: Streamlit documentation (2024). https://docs.streamlit.io

[18] Zigarelli, A., Jia, Z., Lee, H.: Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: Observational study. JMIR Formative Research 6(3), e29967 (2022). https://doi.org/10.2196/29967