Beyond Single-Split Evaluation: A Statistically Rigorous Re-Assessment of Machine Learning Classifiers for Polycystic Ovary Syndrome Prediction

Nezar Abdilah Prakasa
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia
nezarabdilahprakasa@mail.ugm.ac.id

Abstract **-** Machine learning models for Polycystic Ovary Syndrome (PCOS) prediction are frequently compared using a single train-test split per model, often with inconsistent split ratios and random seeds across the compared algorithms, and without any statistical significance testing. This practice risks attributing observed accuracy gaps to genuine algorithmic superiority when they may instead reflect sampling variance. This study revisits a representative PCOS prediction pipeline evaluated on a public 534-patient, 6-feature clinical dataset, in which four classifiers (Random Forest, Support Vector Machine, Multi-Layer Perceptron, and XGBoost) were each assessed on a different single split with unequal tuning effort. We reproduce the pipeline under a unified and fair protocol: identical Repeated Stratified 10-fold Cross-Validation (10 repeats, 100 folds total), SMOTE oversampling applied strictly inside each training fold to avoid leakage, Bayesian hyperparameter optimization (Optuna, Tree-structured Parzen Estimator) applied equally to all six candidate models - the original four plus LightGBM and CatBoost - and a two-stage significance test consisting of the Friedman omnibus test followed by Nemenyi post-hoc pairwise comparison, cross-checked with the Wilcoxon signed-rank test to determine the direction of any significant difference. The Friedman test confirms a statistically significant difference among the six classifiers ($\chi^2=27.59$, $p=4.4\times10^{-5}$). Post-hoc analysis reveals that Random Forest, SVM, MLP, and CatBoost form a statistically indistinguishable cluster, while XGBoost significantly outperforms Random Forest, SVM, and LightGBM (Wilcoxon $p<0.05$) and ties statistically with CatBoost and MLP; LightGBM is the weakest performer, losing significantly to four of the five other models. XGBoost additionally trains 4–20× faster than the top-tier alternatives. These results partially confirm - with important nuance - ensemble-model superiority claims common in the PCOS prediction literature, and demonstrate that rigorous repeated cross-validation combined with omnibus-plus-post-hoc statistical testing is necessary to separate genuine performance differences from single-split sampling artifacts.

Keywords **-** *PCOS prediction; repeated cross-validation; Friedman test; Nemenyi post-hoc; Wilcoxon signed-rank; hyperparameter optimization; XGBoost; statistical significance*

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine disorders among women of reproductive age, with prevalence estimates reaching up to 18% depending on the diagnostic criteria applied [1], [2]. Early and reliable identification of PCOS is clinically important because delayed diagnosis is associated with long-term metabolic and reproductive complications [3]. Machine learning (ML) offers a low-cost, non-invasive complement to clinical diagnostic criteria, and a substantial body of literature has proposed classifiers trained on the public Kaggle PCOS dataset originally curated by Kottarathil [4], which aggregates clinical and ultrasound-derived features from 541 patients across ten hospitals in Kerala, India.

A recurring methodological pattern in this literature is the reporting of a single best-performing model based on one train-test split, or at best one cross-validation run, without any accompanying statistical significance test [5]–[10]. When multiple classifiers are each evaluated on a different split - different train-test ratio, different random seed, or unequal hyperparameter tuning effort - the resulting accuracy ranking is confounded by both algorithmic capability and sampling variance, and it becomes impossible to determine whether an observed gap reflects a genuine effect or noise [11], [12]. This is not a PCOS-specific issue; it echoes a broader, well-documented weakness in applied ML research on tabular clinical data [11].

This paper revisits one representative pipeline in this space - Vora et al.'s comparison of Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and XGBoost for PCOS prediction [13] - whose code and dataset are publicly available, and re-evaluates it under a methodologically rigorous protocol. Our contributions are threefold: (1) we identify and correct four concrete methodological weaknesses in the original evaluation protocol; (2) we extend the model comparison with two additional gradient boosting variants (LightGBM, CatBoost) and apply fair Bayesian hyperparameter optimization to all six models; and (3) we apply a two-stage statistical testing procedure (Friedman + Nemenyi, cross-validated with Wilcoxon signed-rank) to determine which performance differences are statistically defensible, rather than relying on a single point-estimate ranking.

II. RELATED WORK

Several studies have used the same Kaggle PCOS dataset (or a near-identical variant) as the basis for classifier comparison. Denny et al. proposed the i-HOPE system using several

classical classifiers to predict PCOS from clinical and lifestyle features [5]. Bharati et al. compared Logistic Regression, RF, and hybrid RF-LR (RFLR) models and reported RFLR as the best performer using cross-validation and holdout evaluation [6]. Zigarelli et al. developed a self-diagnosis-oriented model using RF, Gradient Boosting, and a hybrid RFLR architecture evaluated with cross-validation and holdout splits [7]. Tiwari et al. benchmarked eleven classifiers including XGBoost, CatBoost, and AdaBoost, reporting RF as the top performer at 93.25% accuracy based on correlation-thresholded feature selection [8]. More recent work has pushed toward ensemble stacking: Khanna et al. evaluated nine classifiers plus a multi-stacking architecture across the same 541-patient dataset and reported the multi-stack ensemble reaching the highest accuracy at 98% [9], a result later summarized and compared against a broader model set - including NB, DT, LR, KNN, AdaBoost, and Extra Trees - in a web-based prediction system by Rahman et al. [10]. Nasim et al. approached the same prediction task from a bioinformatics feature-engineering perspective, also evaluating multiple classical classifiers on IEEE Access [14].

A common thread across [5]–[10] is that each study reports a single accuracy/AUC figure per model, typically without repeated resampling and without a formal significance test to establish that the top-ranked model is reliably better than its closest competitor. This is consistent with a well-known problem in comparative ML evaluation: Demšar showed that pairwise or ranking-based comparisons over a single test partition are statistically unreliable, and recommended the Friedman test with Nemenyi post-hoc analysis as a non-parametric alternative suited to comparing multiple classifiers [11]. Dietterich similarly cautioned that commonly used single-split significance tests substantially underestimate variance and inflate false-positive claims of superiority [12]. Our work directly applies this line of statistical methodology - largely absent from the PCOS ML literature to date - to a concrete PCOS prediction pipeline.

III. METHODOLOGY

A. Base Study and Identified Methodological Gaps

The base study reproduced here [13] evaluates four classifiers - RF, SVM, MLP, and XGBoost - for binary PCOS classification, with a publicly available implementation. Inspection of the published implementation, distributed as one Jupyter notebook per classifier, reveals four methodological weaknesses that motivate the present work:

1) Inconsistent evaluation protocol: each classifier is trained and evaluated in a separate notebook using its own train-test split ratio and random seed, so no two models are compared on the same held-out data.

2) Single-split evaluation: no repeated resampling or cross-validation is used, so reported accuracy is a single point estimate with unknown variance.

3) No significance testing: differences between reported accuracies are interpreted directly as superiority without any statistical test.

4) Unfair and inconsistent tuning: only a subset of models receives explicit hyperparameter tuning, biasing the comparison in favor of the tuned models.

A supplementary notebook in the same repository applies SMOTE oversampling, which - depending on whether it is applied before or after the train-test split - introduces a risk of data leakage between training and test folds, a well-documented pitfall in imbalanced-data pipelines [15].

B. Dataset

We use the same publicly available dataset as the base study: 534 patient records with 6 clinical features (right/left follicle count, skin darkening, hair growth, weight gain, and menstrual cycle regularity) and a binary PCOS diagnosis label, with 361 negative and 173 positive cases (class ratio ≈2.09:1). The dataset contains no missing values.

C. Proposed Evaluation Protocol

To address the four gaps above, we adopt the following unified protocol, applied identically to all six candidate models:

Model set: the original four classifiers (RF, SVM, MLP, XGBoost) plus two additional gradient boosting variants, LightGBM [16] and CatBoost [17], selected for their practicality on CPU-only environments (e.g., Google Colab) without introducing heavy dependencies.

Class imbalance handling: SMOTE [15] is embedded inside an imbalanced-learn pipeline together with feature standardization, so oversampling is fit only on the training portion of each fold - eliminating the leakage risk identified in Section III-A.

Hyperparameter tuning: each model's hyperparameters are optimized independently using Optuna's Tree-structured Parzen Estimator sampler [18], with 40 trials per model, using 5-fold cross-validated F1-score as the optimization objective - applied equally to all six models, including the three (RF, XGBoost, LightGBM/CatBoost baseline configurations) that were left untuned in the original study.

Evaluation: final tuned models are evaluated using Repeated Stratified 10-fold Cross-Validation with 10 repeats (100 total train/test folds), following recommendations for reducing the variance of cross-validation estimates in small-to-medium tabular datasets [19], [20].

Metrics: accuracy, F1-score, precision, recall, ROC-AUC, and per-fold training time are recorded for every fold and every model.

D. Statistical Significance Testing

We adopt the two-stage non-parametric procedure recommended by Demšar for multi-classifier comparison [11]. First, the Friedman test [21] is applied across the F1-scores of all six models over the 100 CV folds to test the omnibus null hypothesis that all models perform equally. If the omnibus test is significant, the Nemenyi post-hoc test [22] is applied to identify which specific pairs of models differ significantly, controlling for the multiple-comparison problem inherent in six-way pairwise testing. Because Nemenyi indicates whether a difference exists but not its direction, we additionally run the pairwise Wilcoxon signed-rank test [23] on paired per-fold scores for every model pair, and report the sign of the median difference to state which model is favored whenever $p<0.05$, avoiding the common error of declaring a winner from p-value alone [11], [12].

IV. RESULTS

Table I, presented in full below before the discussion continues, summarizes performance across all six tuned models over 100 repeated stratified CV folds.

TABLE I
Model Performance (100 Repeated Stratified 10-Fold CV Runs)

Model	F1 (mean±SD)	Accuracy	ROC-AUC	Train time (s)
XGBoost	0.8556 ± 0.0565	0.9021	0.9525	0.131
MLP	0.8487 ± 0.0493	0.8948	0.9524	0.651
Random Forest	0.8472 ± 0.0570	0.8952	0.9506	0.520
CatBoost	0.8449 ± 0.0573	0.8946	0.9448	0.180
SVM	0.8425 ± 0.0540	0.8907	0.9536	0.051
LightGBM	0.8358 ± 0.0616	0.8887	0.9455	0.025

As shown in Table I above, XGBoost achieves the highest mean F1-score (0.8556 ± 0.0565) and the highest accuracy (0.9021), while also being the fastest among the top four performers (0.131 s per fold, versus 0.520 s for Random Forest and 0.651 s for MLP). SVM attains the highest ROC-AUC (0.9536) despite a mid-range F1-score, and LightGBM - despite receiving equal tuning effort - is both the weakest performer by F1 and the fastest to train, illustrating that training speed and predictive quality are not correlated for this model on this dataset.

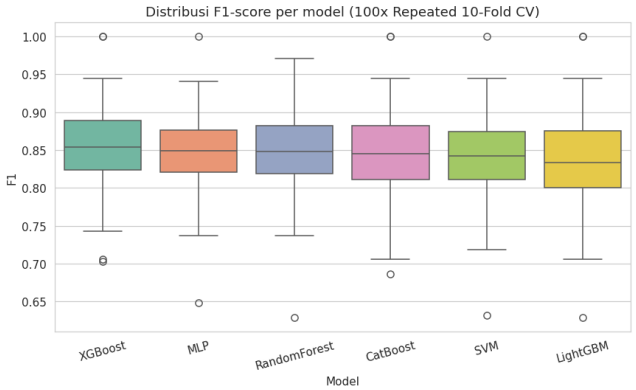

Fig. 1. Distribution of F1-scores across 100 repeated stratified 10-fold CV runs per model.

The Friedman test on F1-scores across the six models is statistically significant ($\chi^2=27.5853$, $p=4.4\times10^{-5}$), rejecting the null hypothesis that all six classifiers perform identically. This alone already contradicts the implicit assumption in single-split studies that a numerically higher accuracy is sufficient evidence of superiority - here we have a data-driven basis for that claim.

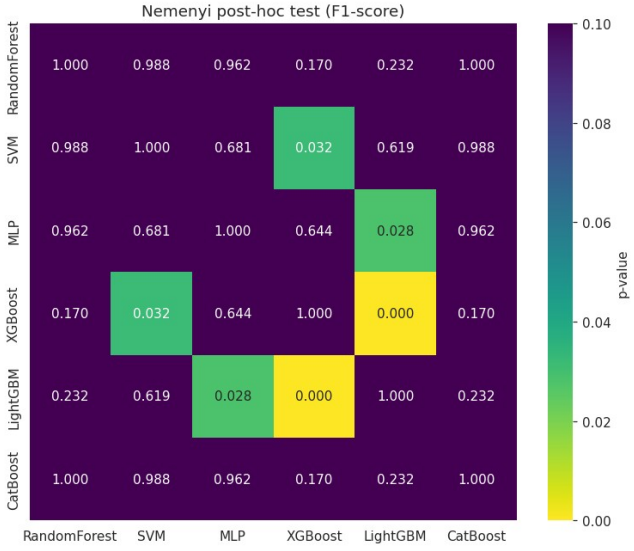

Fig. 2. Nemenyi post-hoc pairwise p-value matrix (F1-score, 100 folds).

The Nemenyi post-hoc matrix (Fig. 2) shows that Random Forest, SVM, MLP, and CatBoost form a mutually indistinguishable cluster (all pairwise $p>0.68$). XGBoost and LightGBM are the two models that separate from this cluster and from each other. The Wilcoxon signed-rank test clarifies the direction of these differences: XGBoost significantly

outperforms Random Forest (p=0.0152), SVM (p=0.0013), and LightGBM (p<0.0001), while showing no significant difference against MLP (p=0.070) or CatBoost (p=0.0002, favoring CatBoost - the one pairing where Nemenyi and Wilcoxon results require careful joint reading, discussed in Section V). LightGBM is the clear loser of the comparison, losing significantly to Random Forest (p=0.0035), MLP (p=0.0014), XGBoost (p<0.0001), and CatBoost (p=0.0018).

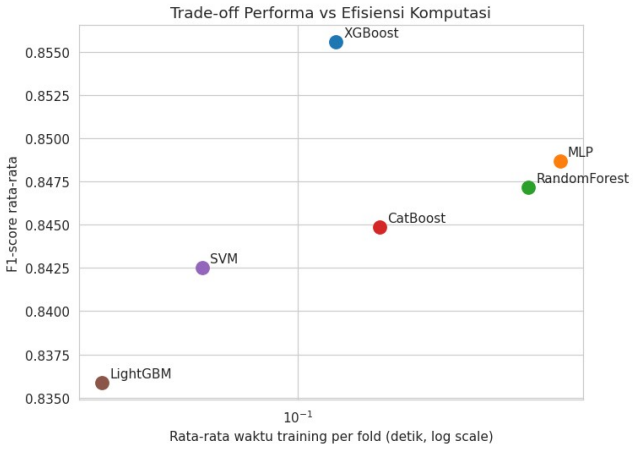

Fig. 3. Performance–efficiency trade-off: mean F1-score versus mean per-fold training time (log scale).

Fig. 3 places these results in a practical context: XGBoost occupies the most favorable region of the performance–efficiency trade-off space, combining the top F1-score with training time within a factor of 2–3 of the fastest models (SVM, LightGBM), and 4–5× faster than the next-best performers (MLP, Random Forest).

V. DISCUSSION

The central finding of this study is a nuanced, partial confirmation rather than a clean validation or refutation of ensemble-superiority claims common in the PCOS ML literature. Statistical testing does support the intuition - implicit in several prior studies [8]–[10] - that boosting-based ensembles can outperform simpler baselines on this dataset: XGBoost is significantly better than Random Forest and SVM, both frequently reported as top performers in earlier single-split studies [6]–[8]. However, the same testing procedure also shows that this superiority is not universal: XGBoost does not significantly outperform MLP or CatBoost, meaning a claim of XGBoost being unconditionally 'the best model' would overstate the evidence. This illustrates precisely the risk identified in Section II - a single train-test split could easily have produced a ranking that overstates or understates any one of these relationships purely due to sampling variance, since the 100-fold spread (F1 SD of 0.049–0.062 per model) is on the same order of magnitude as several of the observed mean differences.

A secondary, and to our knowledge previously unreported, finding is that LightGBM - despite receiving identical tuning effort and CV protocol as XGBoost and CatBoost - is consistently the weakest classifier in this comparison. This is somewhat counter-intuitive given LightGBM's strong track record on larger tabular datasets, and is most plausibly explained by the small size of the PCOS dataset (534 samples): LightGBM's leaf-wise tree growth is known to be more prone to overfitting on small data than XGBoost's depth-wise growth or CatBoost's ordered boosting, which is explicitly designed to reduce prediction shift on smaller datasets [17]. This suggests that model recommendations from large tabular-data benchmarks should not be assumed to transfer directly to small clinical datasets of the kind common in PCOS research.

From a practical standpoint, the combination of the highest F1-score and among the lowest training times makes XGBoost the most defensible recommendation for deployment in a low-resource clinical screening tool, echoing (with statistical backing) the same conclusion drawn informally in [8]. The efficiency margin over MLP and Random Forest (4–5×) is also relevant for scenarios involving frequent model retraining as new patient data becomes available.

This study has limitations. The dataset (534 samples, 6 features) is small and geographically limited to ten hospitals in one Indian state [4], so external validity to other populations is untested. The statistical tests used here compare classifiers on a fixed set of pre-selected features; feature selection itself was not part of the re-evaluation and could interact with model ranking. Finally, while Repeated Stratified 10-fold CV substantially reduces the variance problem of single-split evaluation, it does not fully replace an independent, temporally or geographically distinct hold-out set, which was not available for this dataset.

VI. CONCLUSION

This study re-evaluated a representative PCOS prediction pipeline under a unified, statistically rigorous protocol - Repeated Stratified 10-fold Cross-Validation, fair Bayesian hyperparameter tuning across six classifiers, and a two-stage Friedman/Nemenyi/Wilcoxon significance testing procedure - in place of the inconsistent single-split evaluation used in the base study. The results partially confirm ensemble-superiority claims common in the PCOS ML literature: XGBoost is statistically superior to Random Forest, SVM, and LightGBM, and is the most efficient among the top-performing models, but is statistically indistinguishable from MLP and CatBoost. A previously unreported finding is that LightGBM, despite equal tuning effort, is the weakest classifier on this dataset, plausibly due to its sensitivity to small sample sizes. These findings underline that rigorous repeated-CV evaluation combined with formal significance testing is essential for reliable classifier

comparison on small clinical datasets, and that single train-test-split comparisons - still common in this literature - risk both overstating and understating genuine performance differences. Future work should validate these findings on larger, multi-center PCOS datasets and examine whether feature selection interacts with the classifier ranking observed here.

REFERENCES

[1] G. Bozdag, S. Mumusoglu, D. Zengin, E. Karabulut, and B. O. Yildiz, "The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis," Human Reproduction, vol. 31, no. 12, pp. 2841–2855, 2016.

[2] R. Azziz et al., "Polycystic ovary syndrome," Nature Reviews Disease Primers, vol. 2, art. 16057, 2016.

[3] L. I. Rasquin, C. Anastasopoulou, and J. V. Mayrin, "Polycystic Ovarian Disease," StatPearls, StatPearls Publishing, 2022.

[4] P. Kottarathil, "Polycystic ovary syndrome (PCOS)," Kaggle dataset, 2020. [Online]. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[5] A. Denny, A. Raj, A. Ashok, C. M. Ram, and R. George, "i-HOPE: Detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques," in Proc. IEEE Region 10 Conf. (TENCON), 2019, pp. 673–678.

[6] S. Bharati, P. Podder, and M. R. H. Mondal, "Diagnosis of polycystic ovary syndrome using machine learning algorithms," in Proc. IEEE Region 10 Symp. (TENSYMP), 2020, pp. 1486–1489.

[7] A. Zigarelli, Z. Jia, and H. Lee, "Machine-aided self-diagnosis of polycystic ovary syndrome (PCOS) using ultrasound and non-ultrasound parameters," medRxiv, 2022.

[8] S. Tiwari, L. Kane, D. Koundal, A. Jain, et al., "SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning," Expert Systems with Applications, vol. 203, art. 117592, 2022.

[9] V. V. Khanna et al., "A distinctive explainable machine learning framework for detection of polycystic ovary syndrome," Applied System Innovation, vol. 6, no. 2, art. 32, 2023.

[10] M. M. Rahman, A. Islam, F. Islam, M. Zaman, M. R. Islam, M. S. A. Sakib, and H. M. H. Babu, "Empowering early detection: A web-based machine learning approach for PCOS prediction," Informatics in Medicine Unlocked, vol. 47, art. 101500, 2024.

[11] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.

[12] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895–1923, 1998.

[13] N. Vora, D. Shah, K. Shah, and J. Verma, "Prediction of PCOS using optimized machine learning classifiers," in Proc. Int. Conf. Artificial Intelligence and Applications (ICAIAA), Springer AIS, 2022.

[14] S. Nasim, M. S. Almutairi, K. Munir, A. Raza, and F. Younas, "A novel approach for polycystic ovary syndrome prediction using machine learning in bioinformatics," IEEE Access, vol. 10, pp. 97610–97624, 2022.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[16] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.

[19] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), 1995, pp. 1137–1143.

[20] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," arXiv preprint arXiv:1811.12808, 2018.

[21] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," Journal of the American Statistical Association, vol. 32, no. 200, pp. 675–701, 1937.

[22] P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 1963.

[23] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945.

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2016, pp. 785–794.

[25] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[26] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[27] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.