Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Leakage-Aware PCOS Diagnosis Across Independent Ultrasound and Clinical–Biochemical Data Sources

Nisar Ali
Abasyn University, Peshawar, Pakistan Email: nisaricup@gmail.com

Muhammad Abbas
Natioanl University of Modern Language, Peshawar, Pakistan
Email: muhammad.abbas@numl.edu.pk

Muzamil Shah
Natioanl University of Modern Language, Peshawar, Pakistan
Email: mshah@numl.edu.pk

Umair Chand
The Julius Maximilians University of Würzburg, Berufserfahrung, Germany
Email: umairrehman009gmail.com

Ayesha Tariq
Abasyn University, Peshawar, Pakistan Email: ayeshashawal7@gmail.com

Hira Javed
Abasyn University, Peshawar, Pakistan Email: javedanfal2315@gmail.com

Syeda Sariya Shah
Abasyn University, Peshawar, Pakistan Email: sariyashh@gmail.com

ABSTRACT

Artificial-intelligence studies for polycystic ovary syndrome (PCOS) frequently report high discrimination, yet duplicated images, non-independent partitions, preprocessing leakage, and optimistic model selection can materially distort performance. This study evaluates two independent PCOS data sources through separate leakage-controlled branches: ovarian ultrasound images and clinical–biochemical records. The ultrasound audit began with 10,000 valid images. Exact and orientation-aware perceptual duplicate analysis identified 17 cross-label conflicting groups, removed 3,721 exact duplicate files after conflict exclusion, and retained 1,637 images in 718 analysis groups. A duplicate-group-safe 70/15/15 split was used for multi-seed transfer learning. ResNet50 was selected by mean validation PR-AUC, then Platt-calibrated and evaluated once at a locked threshold on 273 test images, achieving ROC-AUC 0.9926, PR-AUC 0.9975, sensitivity 0.9259, specificity 1.0000, F1 0.9615, and Brier score 0.0139. Cluster-bootstrap 95% confidence intervals were 0.9756–1.0000 for ROC-AUC and 0.8772–0.9976 for

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

F1. The independent clinical branch contained 541 observations and 41 leakage-screened features. Global nested model selection across five outer folds yielded ROC-AUC 0.9501, PR-AUC 0.9091, sensitivity 0.8644, specificity 0.8984, F1 0.8338, and Brier score 0.0798. Grad-CAM visualized image-model attribution; clinical SHAP analysis was excluded because the executed notebook recorded a preprocessing-fit error. The results show that strong PCOS prediction can coexist with substantial data-integrity risks, making leakage control, locked evaluation, calibration, and uncertainty estimation central to defensible reporting. The two cohorts were not patient-linked; no image–clinical fusion was performed.

Keywords: Polycystic Ovary Syndrome, Medical Ultrasound, Machine Learning, Deep Learning, Data Leakage, Model Calibration, Clinical Prediction

Introduction

Polycystic ovary syndrome (PCOS) is a heterogeneous endocrine and metabolic disorder that affects reproductive, metabolic, and psychological health across the life course. Contemporary evidence-based guidance recognizes diagnostic heterogeneity rather than a single defining marker, with adult diagnosis based on combinations of ovulatory dysfunction, clinical or biochemical hyperandrogenism, and ovarian morphology or, in specified adult settings, anti-Müllerian hormone (AMH) [1, 2]. Reviews also emphasize that prevalence estimates vary with criteria and population, and that delayed or inconsistent diagnosis remains common [3]. This heterogeneity creates a plausible role for computational methods that can model nonlinear relationships among clinical findings, biochemical markers, and ovarian imaging without reducing PCOS to a single proxy variable.

Machine learning (ML) and deep learning (DL) have consequently been applied to both structured clinical data and ovarian ultrasound. Systematic reviews describe a rapidly growing literature spanning classical classifiers, ensemble models, convolutional neural networks, and explainability methods [4, 5]. Ultrasound studies have reported very high image-classification accuracy with transfer learning or hybrid feature extrac-tion [6–8], while newer studies have moved toward patient-separated and external evaluation [9]. Clinical studies likewise report strong discrimination using anthropomet-ric, hormonal, metabolic, and ultrasonographic variables [10–14]. These results support the technical feasibility of AI-assisted PCOS assessment, but performance values are not directly comparable when the underlying cohorts, targets, preprocessing pipelines, and validation designs differ.

Reliability is particularly important for small or reused medical datasets. Data leakage can arise when duplicates or observations from the same subject enter dif-ferent partitions, when feature selection or preprocessing is fitted before resampling, when target-proxy variables are retained, or when thresholds and model configurations are optimized using test labels. Leakage has been identified as a recurring source of optimistic performance across scientific ML applications [15]. Current reporting guid-ance for prediction models and medical imaging therefore places substantial weight on transparent data partitioning, model-

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

development separation, calibration, uncertainty, and external validation [16, 17]. For ovarian ultrasound, exact duplicates, transformed copies, and visually near-identical images can be especially problematic when public datasets are assembled from heterogeneous or repeated sources. A random image-level split can then estimate recognition of repeated content rather than generalization to independent image groups.

The present study addresses this methodological concern by evaluating two indepen-dent PCOS data sources through separate leakage-controlled branches. The ultrasound branch implements exact and orientation-aware near-duplicate analysis, duplicate-group-safe partitioning, training-only augmentation, multi-seed transfer learning, validation-based model selection, probability calibration, threshold locking, cluster-bootstrap uncertainty, paired model comparison, attribution visualization, and runtime benchmarking. The clinical–biochemical branch performs identifier and target-proxy auditing, leakage-safe preprocessing and feature selection, and global nested model selection with outer-fold prediction. Because the ultrasound and clinical records do not contain verified cross-source patient identifiers, the study does not perform or claim patient-level multimodal fusion.

The principal contributions are fourfold. First, the ultrasound analysis quantifies and controls duplicate-related leakage before model development rather than assuming folder-level independence. Second, image-model selection, calibration, and threshold choice are separated from the locked test set, with uncertainty estimated by resampling analysis groups. Third, the clinical branch treats outer-fold predictions from global nested selection as the primary performance estimate instead of reporting a full-data refit as an unbiased test result. Fourth, the two branches are reported as complementary but independent PCOS evaluations, preserving the evidential boundary imposed by the absence of patient-level linkage.

Related Work

Ultrasound-based PCOS artificial intelligence

Ovarian ultrasound provides morphological information that can be exploited by image classifiers, but its use in ML is sensitive to acquisition heterogeneity, labeling quality, and sample independence. Suha and Islam [6] combined VGG16-derived representations with an ensemble classifier on a relatively small ultrasound collection and reported near-perfect classification. Kermanshahchi et al. [8] similarly evaluated ML-based classification on pelvic ultrasound. These studies demonstrate the separability that can exist in curated image collections, yet the strength of their clinical inference depends on whether image partitions correspond to independent subjects or genuinely independent examinations.

Moral et al. [7] are methodologically informative because they explicitly identified substantial overlap between the originally supplied training and test sets of a public ultrasound dataset, discarded the overlapping test set, and created a new split before evaluating CystNet. Their observation illustrates why nominal train/test folders can-not be treated as proof of independence. More recent image pipelines have extended PCOS analysis through super-resolution and segmentation [18] or transfer learning with complementary biomarker analysis [19]. Zhao et al. [9] provide a stronger valida-tion example: their

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

ultrasound recognition model used patient-level separation in a development center and an independent-center external cohort. Such designs are more informative for generalization than high accuracy from an image-level split alone.

Clinical and biochemical machine learning

Structured PCOS prediction commonly uses menstrual history, anthropometry, man-ifestations of hyperandrogenism, hormonal measurements, and ultrasound-derived counts or dimensions. Silva et al. [20] used ML to characterize clinical and laboratory patterns associated with PCOS phenotypes. Elmannai et al. [10] and Khanna et al. [11] evaluated explainable or stacked ML systems on tabular PCOS datasets, illustrating how feature selection and ensemble learning can achieve high discrimination in rela-tively small cohorts. Tong et al. [12] focused on AMH and steroid/hormonal variables, showing that clinically targeted predictor sets can also support strong classification.

Agirsoy and Oehlschlaeger [13] evaluated clinical, biochemical, and ultrasonographic feature combinations in 541 women and reported high XGBoost performance. Their additional dataset produced perfect test metrics, which the authors themselves treated cautiously because overfitting, leakage, or strong similarity between training and test data could not be excluded. In contrast, Yao et al. [14] developed an explainable XGBoost model using a larger multicenter cohort and reported similar internal and external AUCs, providing a more direct assessment of transportability. Evans et al.

[21] showed that a small set of clinically selected features could retain high predictive discrimination in a 151-participant multicenter retrospective cohort, although validation remained internal.

Explainability and reliable probability estimates

Explainability is frequently used to make model behavior inspectable rather than to establish biological causality. Grad-CAM highlights image regions that contribute to a convolutional network output [22]; SHAP decomposes a prediction into feature-associated contributions under a specified background distribution [23]. PCOS studies have used SHAP to rank clinical predictors and examine individual predictions [10, 13, 14]. These tools can reveal whether a model attends to plausible regions or variables, but they do not prove that the highlighted structures or features cause PCOS.

Calibration is a separate requirement. A model can discriminate well while assigning probabilities that are systematically too high or too low. This distinction matters when probabilities might inform triage or decision support. Reporting guidelines such as TRIPOD+AI and CLAIM now encourage clear separation between discrimination, calibration, threshold-dependent classification, and clinical utility [16, 17]. The present study therefore reports Brier score and expected calibration error alongside ROC-AUC, PR-AUC, sensitivity, specificity, and F1.

Methodological reliability and comparison of recent studies

Table 1 summarizes five experimental studies selected for methodological relevance to the current dual-source design. The studies span ultrasound

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

classification, explicit overlap handling, patient-separated external imaging validation, combined clini-cal/ultrasound tabular prediction, and compact clinical prediction. The table is intended

Table 1 Five methodologically relevant experimental PCOS-AI studies. Performance is study-specific and should not be ranked across non-equivalent cohorts.

Study	Year	Modality	Cohort	Method / validation	Main finding	Key limitation or strength
Suha and Islam [6]	2022	US	594 images	VGG16 features + ensemble ML; image train/test split	Near-perfect reported classification	Small set; subject independence not established in the reported protocol.
Moral et al. [7]	2024	US	1,924 retained public images	Deep learning; patient-separated development + independent-center external validation	96.54% dense and 97.75% RF accuracy	Strong external discrimination
Zhao et al. [9]	2025	US	933 patients, two centers; 1,722 images	CystNet; new split + 5-fold CV after discarding an overlapping supplied test set	Explicit overlap audit; patient or duplicate-group independence of the new split not documented.	Methodologically stronger general-ization design; transportability beyond the study setting remains to be tested.
Agirsoy and Oehlschlaeger [13]	2025	Clinical + US variables	541 women; additional 320-instance dataset	Multiple ML models, XGBoost, SHAP; CV/held-out and separate-dataset evaluation	Clinical + USG + AMH AUC 0.9947; perfect additional-dataset metrics	

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Authors explicitly caution about overfitting, leakage, or train/test similarity.to compare study design rather than rank numerical performance across non-equivalent datasets.

Taken together, these studies show that both imaging and structured clinical variables can support highly discriminative PCOS models. They also illustrate why raw accuracy is an incomplete measure of evidence quality. A model evaluated after subject-separated external validation answers a different question from a model evaluated after random image splitting. Likewise, perfect performance on a second public dataset is not automatically equivalent to prospective external validation. The present work is positioned around leakage control and evaluation discipline uplicate groups are isolated before ultrasound splitting, the test set is locked after validation-based selection and calibration, and the clinical branch uses nested outer-fold predictions as its principal estimate. It does not claim direct superiority over the studies in Table 1.

Materials and Methods

Study design and evidential separation

The executed analysis comprised two independent data sources: an ovarian-ultrasound image collection and a clinical–biochemical spreadsheet. No verified patient identifier linked an ultrasound image to a clinical record across the two sources. Each branch was therefore developed and evaluated independently, and results were combined only at the level of scientific interpretation. Figure 1 summarizes this design

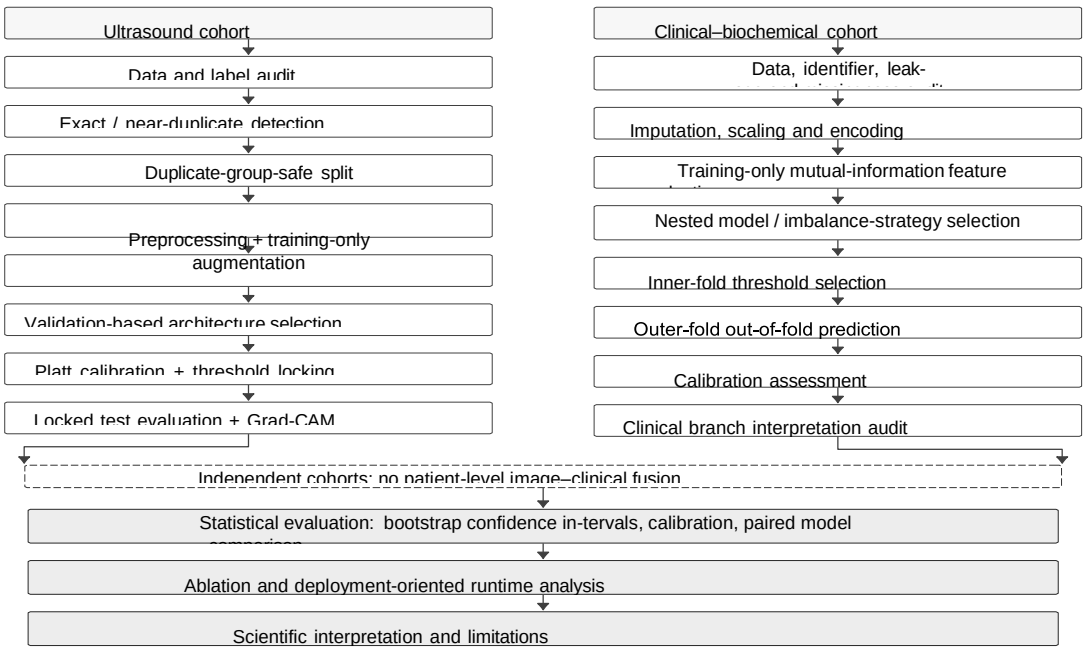


Fig. 1 Leakage-aware study workflow. The ultrasound and clinical–biochemical branches were independent. They were not fused at patient level

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

because no verified cross-source linkage was available.

Ultrasound data audit and duplicate control

The image manifest contained 10,000 valid files, with 5,000 labeled PCOS and 5,000 labeled non-PCOS before duplicate control. No corrupt files were reported. True patient identifiers were unavailable, so independence could not be enforced at patient level. Instead, the notebook constructed analysis groups from content similarity.

Exact duplicates were identified by MD5 hashing. Near-duplicate analysis used perceptual hashes computed under eight orientation transforms, with candidate search based on a BK-tree and Hamming-distance threshold of 8. Candidate relationships were consolidated with union–find to create duplicate/orientation groups. This audit yielded 116,178 near/orientation candidate pairs. Seventeen groups contained cross-label conflicts and were excluded from downstream modeling. After conflict exclusion, 3,721 exact duplicate files were removed. The final image cohort comprised 1,637 images in 718 analysis groups; 225 groups contained more than one retained near/orientation image. The largest group contained 62 images, with mean and median group sizes of 2.28 and 1, respectively.

The retained images were partitioned by analysis group, not by individual file, using a 70/15/15 train/validation/test target. Five hundred candidate split assignments were searched, and split seed 369 was selected by the notebook's balance criterion. The final partitions contained 1,125 training images (789 PCOS, 336 non-PCOS), 239 validation images (167 PCOS, 72 non-PCOS), and 273 test images (189 PCOS, 84 non-PCOS). Assertions confirmed no path, MD5, or duplicate-group overlap across partitions. Because patient IDs were not available, this procedure is described as duplicate-group-safe splitting rather than patient-level splitting.

Ultrasound preprocessing, transfer learning, and locked evaluation

Images were resized to 224 × 224 RGB. Augmentation was applied only to the training partition and consisted of random horizontal flipping, rotation factor 0.06, zoom 0.10, and contrast 0.10. Validation and test transformations were deterministic. Contrast-limited adaptive histogram equalization was disabled in the executed configuration, and no CLAHE ablation result is reported.

Four ImageNet-initialized architectures were evaluated: MobileNetV2 [24], ResNet50 [25], EfficientNetV2B0 [26], and EfficientNetV2B0 augmented with squeeze-and-excitation attention [27]. Each architecture was trained with seeds 42, 52, and 62. The base network was initially frozen; the classification head used dropout 0.40, a 128-unit ReLU layer with $L_2 = 10^{-4}$, and dropout 0.30. Training used AdamW with weight decay 10^{-5}, learning rate 3×10^{-4} for the head stage and 10^{-5} for fine-tuning, and binary cross-entropy. Up to 15 head-training and 15 fine-tuning epochs were allowed, with early stopping patience 5 and learning-rate reduction factor 0.3 after two stagnant epochs. The last 40 eligible layers were fine-tuned while batch-normalization layers remained frozen. Class weights were 1.6741 for the non-PCOS class and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

0.7129 for the PCOS class.

Architecture selection was based on mean validation PR-AUC across the three seeds. After selecting the architecture without test inspection, validation probabilities were used for Platt calibration and threshold selection. The threshold maximized Youden's index,

$$J = \text{sensitivity} + \text{specificity} - 1, \quad (1)$$

and was then locked before test inference. For binary targets $y_i \in \{0, 1\}$ and predicted probabilities p_i, probability calibration was summarized by the Brier score

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

The principal test report includes ROC-AUC, PR-AUC, sensitivity, specificity, precision, F1, balanced accuracy, Matthews correlation coefficient (MCC), Cohen's kappa, and Brier score.

Ultrasound uncertainty, paired tests, attribution, and runtime

Uncertainty for the selected locked image model was estimated with 2,000 bootstrap replicates clustered by analysis-group identifier. This preserves dependence among retained near-duplicate/orientation images within a group. Pairwise thresholded predic-tions from all four image architectures were compared using McNemar tests, followed by Holm multiplicity correction. For $b + c > 0$, a continuity-corrected McNemar statistic can be written as

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}, \quad (3)$$

where b and c are discordant pair counts. When $b = c = 0$, there are no discordant cases to test; this was the executed outcome for every architecture pair. Grad-CAM [22] was generated for representative true-positive, true-negative, and false-negative examples from a constituent model. No false-positive example existed in the locked test predictions. These maps are interpreted solely as model attributions. They do not establish pathology, causal mechanisms, or agreement with a clinician.

Deployment-oriented timing was measured in the executed CPU environment. The selected ResNet50 constituent model was benchmarked through Keras and after TensorFlow Lite conversion at batch size one. The final clinical random forest was timed for single-row predict proba. These timings are not treated as edge-device or clinical deployment validation because no dedicated edge hardware or energy measurement was used.

Clinical–biochemical data audit and preprocessing

The clinical spreadsheet worksheet Full new contained 541 rows and 45

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

columns, with 177 PCOS and 364 non-PCOS outcomes in the executed target encoding. Identifier-like variables Sl. No, Patient File No., and Unnamed: 44 were excluded from modeling, leaving 41 safe features. No exact duplicate rows were removed. The notebook found no feature with an extreme univariate AUC at its screening threshold of 0.995, reducing evidence for a simple target proxy but not proving the absence of all possible leakage. No repeated-patient grouping variable was available; clinical resampling was therefore row-level.

Missingness was sparse apart from Unnamed: 44, which was already excluded: II beta-HCG, AMH, marriage status, and fast-food variables each had one missing value. Candidate feature sets were predefined as clinical-history/anthropometric (22 features), biochemical/hormonal (11), ultrasound measurements (5), clinical plus ultrasound (27), and all safe features (41). Blood group, haemoglobin, and number of abortions were retained only in the full-safe set.

Within training resamples, numerical features were median-imputed and standardized, while categorical features were most-frequent-imputed and one-hot encoded.

Mutual-information SelectKBest was applied inside the modeling pipeline, with can-didate feature counts of 5, 10, 15, or all. Candidate imbalance handling included no resampling and RandomOverSampler. Model families were logistic regression, support vector machine (SVM) [28], random forest [29], and XGBoost [30].

Global nested clinical model selection

The clinical branch used five outer folds and three inner folds. Within each outer-training set, the inner search jointly considered predefined feature groups, model families, imbalance strategy, feature-selection setting, and model hyperparameters using ROC-AUC. After the best inner candidate was selected, inner out-of-fold probabilities were used to select a fold-specific Youden threshold without access to the outer assessment fold. The selected pipeline was then fitted on the full outer-training fold and evaluated on the outer fold. Concatenated outer-fold probabilities constitute the principal clinical estimate.

This distinction is important because the notebook also refitted a final full-data candidate (full-safe features with random forest, threshold 0.44) for prospective use. That refit is not reported as unbiased efficacy. Clinical calibration was evaluated on nested outer-fold probabilities using Brier score and expected calibration error (ECE). The notebook attempted SHAP analysis for the refitted clinical model, but the executed output recorded a NotFittedError from the one-hot encoder. Accordingly, no SHAP plots or feature-attribution claims are included in this manuscript.

Experimental environment

The executed environment used Python 3.12.13, TensorFlow 2.20.0, NumPy 2.0.2, pandas 2.2.2, OpenCV 4.13.0, XGBoost 3.3.0, and SHAP 0.52.0. The runtime was Linux on x86 64 with 12.67 GB RAM and no detected GPU. The global random seed was 42, with additional image-training seeds as described above. Dataset provenance and licensing metadata were not documented in the supplied notebook and are therefore not inferred here.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Results

Data-integrity audit and cohort construction

Table 2 reports the key data-audit outcomes. The most consequential ultrasound finding was the scale of repeated content. Although the raw folders were balanced at 5,000 images per class, duplicate and conflict control reduced the modelable collection to 1,637 images. The final partitions were PCOS-predominant because splitting occurred after group-level exclusion and grouping constraints. No group overlap was detected across train, validation, and test partitions.

The clinical dataset contained 541 unique rows at the available grouping level and 41 modeling features after identifier exclusion. The target distribution was 177 PCOS and 364 non-PCOS. Because no repeated-patient identifier was available, clinical nested validation cannot be interpreted as patient-grouped evaluation beyond the row identifiers present in the file.

Table 2 Dataset and leakage-audit summary extracted from the executed notebook.

Item	Executed result
Raw ultrasound files	10,000 valid; 5,000 PCOS and 5,000 non-PCOS; 0 corrupt
Near/orientation candidate pairs	116,178
Cross-label conflict groups excluded	17
Exact duplicate files removed	3,721
Retained ultrasound cohort	1,637 images in 718 analysis groups
Retained groups with > 1 image	225; largest group 62 images
Ultrasound split	Train 1,125 (789/336), validation 239 (167/72), test 273 (189/84), PCOS/non-PCOS
Leakage assertions	No path, MD5, or duplicate-group overlap across splits
Clinical cohort	541 rows: 177 PCOS, 364 non-PCOS
Clinical predictors	41 safe features after excluding 3 identifier-like columns
Clinical duplicates / grouping	0 exact duplicate rows removed; no repeated-patient group variable available
External validation	Not performed in either study branch

Table 3 Ultrasound architecture comparison. Validation values are means across three seeds; test metrics use each architecture's executed threshold. ResNet50 was selected before test inference by validation PR-AUC.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Architecture	Val ROC-AUC	Val PR-AUC	Val Brier	Test ROC-AUC	Test PR-AUC	Test F1	Test Brier	FP / FN
MobileNetV2	0.9943	0.9982	0.00071	0.9948	0.9980	0.9615	0.0502	0 / 14
ResNet50 (selected)	**0.9981**	**0.9993**	**0.0043**	0.9926	0.9975	0.9615	**0.0139**	0 / 14
EfficientNetV2B0	0.9948	0.9984	0.00050	0.9938	0.9978	0.9615	0.0406	0 / 14
EfficientNetV2B0+SE	0.9949	0.9984	0.00044	0.9923	0.9975	0.9615	0.0333	0 / 14

Ultrasound architecture selection and locked test performance

Across three seeds, ResNet50 achieved the highest mean validation PR-AUC (0.9993) and ROC-AUC (0.9981), and the lowest reported validation Brier score (0.0043), so it was locked as the primary architecture before test inference. Platt calibration was applied and the locked Youden threshold was 0.9666.

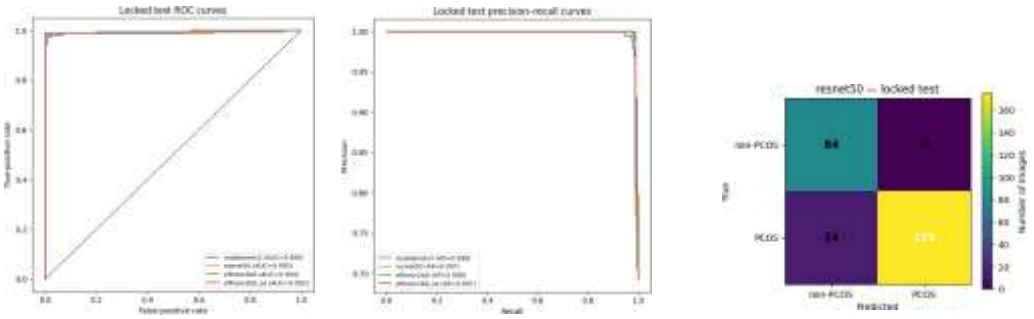
On the 273-image test set, the selected ResNet50 produced 175 true positives, 84 true negatives, 14 false negatives, and no false positives. This corresponded to accuracy 0.9487, sensitivity 0.9259, specificity 1.0000, precision 1.0000, F1 0.9615,

balanced accuracy 0.9630, MCC 0.8909, Cohen's kappa 0.8850, ROC-AUC 0.9926, PR-AUC 0.9975, and Brier score 0.0139. Figure 2 shows the actual notebook ROC, precision–recall, and confusion-matrix outputs.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about



(a) ROC curves (b) Precision–recall curves (c) Confusion matrix **Fig. 2** Actual notebook outputs for the locked ultrasound evaluation. Curves show all evaluated architectures; the confusion matrix corresponds to the selected calibrated ResNet50 at its locked threshold.

The four architectures had different probability rankings and calibration errors but identical thresholded test decisions: every model produced 84 true negatives, zero false positives, 14 false negatives, and 175 true positives. Consequently, all six pairwise McNemar comparisons had $b = c = 0$, raw $p = 1.0$, and Holm-adjusted $p = 1.0$. These results do not support a claim that any architecture had statistically superior thresholded test classification.

Cluster bootstrap reflected uncertainty more conservatively than point estimates alone. For ResNet50, the 95% interval was 0.8559–0.9967 for accuracy, 0.8772–0.9976 for F1, 0.9756–1.0000 for ROC-AUC, 0.9891–1.0000 for PR-AUC, and 0.8906–0.9976
for balanced accuracy. The relatively broad lower bounds for threshold-dependent metrics are relevant when interpreting a test set containing 718 source groups overall but only 273 retained images in the final test partition.

Clinical nested-validation performance

Global nested selection did not identify the same pipeline in every outer fold. The full-safe feature group with random forest was selected in three folds, full-safe with SVM in one fold, and clinical-plus-ultrasound with SVM in one fold. Inner best ROC-AUC values ranged from 0.953 to 0.962. This selection variability indicates that several candidate pipelines were competitive within the limited cohort.

Concatenated outer-fold predictions yielded ROC-AUC 0.9501 and PR-AUC 0.9091. At the fold-specific inner-selected thresholds, accuracy was 0.8872, sensitivity 0.8644, specificity 0.8984, precision 0.8053, F1 0.8338, balanced accuracy 0.8814, MCC 0.7497, Cohen's kappa 0.7486, and Brier score 0.0798. The confusion matrix contained 327 true negatives, 37 false positives, 24 false negatives, and 153 true positives. Bootstrap 95%
intervals were 0.9300–0.9674 for ROC-AUC, 0.8643–0.9452 for PR-AUC, 0.7913–0.8735
for F1, and 0.8505–0.9111 for balanced accuracy.

Calibration, ablation, attribution, and runtime

The selected image model had Brier score 0.0139 and ECE 0.0202 on the

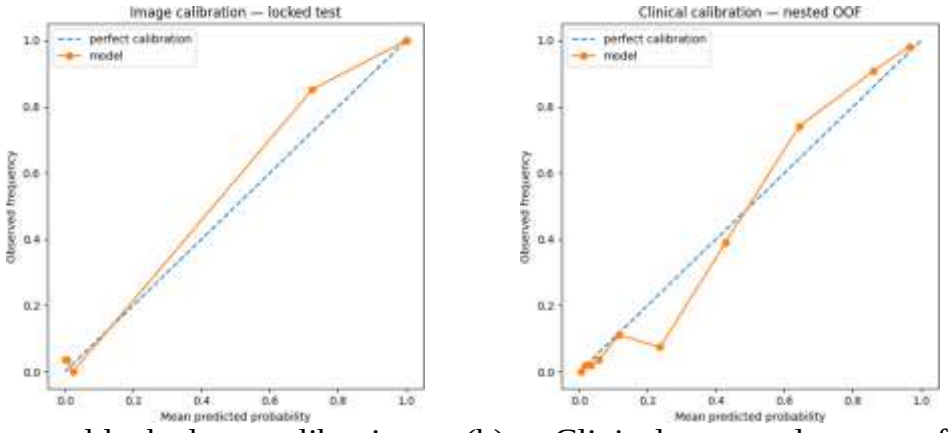
Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

locked test set. The clinical nested outer-fold probabilities had Brier score 0.0798 and ECE

Table 4 Primary performance estimates with 2,000-replicate bootstrap confidence intervals. Image bootstrap was clustered by duplicate-analysis group; clinical estimates use nested outer-fold predictions.

Branch / metric	Estimate	95% CI low	95% CI high
Ultrasound accuracy	0.9487	0.8559	0.9967
Ultrasound F1	0.9615	0.8772	0.9976
Ultrasound ROC-AUC	0.9926	0.9756	1.0000
Ultrasound PR-AUC	0.9975	0.9891	1.0000
Ultrasound balanced accuracy	0.9630	0.8906	0.9976
Clinical accuracy	0.8872	0.8595	0.9131
Clinical F1	0.8338	0.7913	0.8735
Clinical ROC-AUC	0.9501	0.9300	0.9674
Clinical PR-AUC	0.9091	0.8643	0.9452
Clinical balanced accuracy	0.8814	0.8505	0.9111



(a) Ultrasound locked-test calibration (b) Clinical nested outer-fold calibration

Fig. 3 Calibration plots generated by the executed notebook for the two

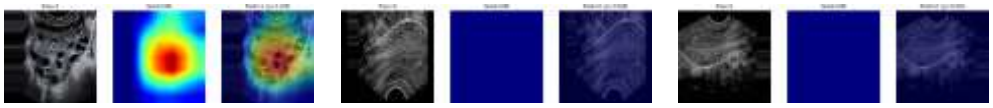
Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

independent branches.

0.0477. Figure 3 reproduces the actual notebook calibration plots. These values indicate stronger probability calibration in the ultrasound test evaluation, but the branches involve different datasets and validation structures and should not be compared as if they were a common cohort.

The squeeze-and-excitation ablation did not improve the EfficientNetV2B0 test result: without added SE, ROC-AUC was 0.9938, PR-AUC 0.9978, and F1 0.9615; with SE, the corresponding values were 0.9923, 0.9975, and 0.9615. The notebook explicitly recorded that a CLAHE ablation was not executed, so no CLAHE effect is claimed.

Grad-CAM examples are shown in Figure 4. The maps demonstrate that the selected image model can be interrogated spatially for individual predictions. They are not used as evidence of lesion localization, follicle detection, or diagnostic agreement with an



(a) True positive (b) True negative (c) False negative

Fig. 4 Representative Grad-CAM outputs generated by the notebook for the ultrasound branch. Heatmaps are model-attribution visualizations and are not clinical annotations.

Table 5 Deployment-oriented runtime results from the executed notebook.

Benchmark	Size (MB)	Median (ms)	P95 (ms)
ResNet50 Keras CPU, batch 1	224.65	364.13	522.53
ResNet50 TFLite CPU, batch 1	24.18	92.94	158.89
Clinical RF, single-row probability	3.12	18.39	23.23

expert. The absence of false positives in the locked test set meant that a false-positive Grad-CAM example was unavailable. Clinical SHAP results are not shown because the executed notebook reported that the one-hot encoder was not fitted in the SHAP transformation path; the manuscript therefore makes no SHAP-based claims about the present clinical model.

Runtime measurements are summarized in Table 5. TensorFlow Lite conversion reduced the selected image model file from 224.65 MB to 24.18 MB and reduced measured median CPU latency from 364.13 ms to 92.94 ms in the executed environment. The clinical random forest file was 3.12 MB with median single-row prediction latency of

18.39 ms. No power measurement or dedicated edge-device benchmark was

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

performed.

Discussion

The central finding is that high PCOS classification performance remained after a sub-stantially stricter data-integrity process, but the audit also exposed why raw benchmark numbers require caution. The ultrasound folders initially contained 10,000 appar-ently valid and balanced images. After cross-label conflict exclusion, exact duplicate removal, and near/orientation grouping, only 1,637 images remained for modeling. This reduction is scientifically more important than a cosmetic preprocessing step: without grouping, highly similar content could occur on both sides of an image-level split, allow-ing a network to exploit repeated visual information. The large number of candidate near/orientation pairs and the presence of 62 images in the largest retained group show that file count alone materially overstated the number of independent visual units.

This observation aligns with the broader leakage literature [15] and with PCOS-specific evidence. Moral et al. [7] found enough overlap between the supplied public training and test folders to discard the original test set. Their report demonstrates that data-packaging boundaries in public imaging repositories are not necessarily statistical boundaries. The present study extends this principle by using exact hashes, transformed

perceptual hashes, a BK-tree search, union–find grouping, cross-label conflict exclusion, and group-safe splitting before training. Because genuine patient identifiers were absent, this remains weaker than patient-level separation; it is nevertheless a more defensible estimate of independence than random file-level splitting.

The selected ResNet50 achieved excellent discrimination on the locked test set, with ROC-AUC 0.9926 and PR-AUC 0.9975. However, three details temper any claim of architectural superiority. First, ResNet50 was chosen by validation PR-AUC, not by its test ROC-AUC; MobileNetV2 and EfficientNetV2B0 actually had slightly higher test ROC-AUC values. Second, all four architectures produced exactly the same binary decisions at their executed thresholds. McNemar tests therefore contained no discordant pairs and cannot distinguish their thresholded classification. Third, the cluster-bootstrap interval for test accuracy extended down to 0.8559. The most defensible interpretation is that several transfer-learning backbones performed similarly on this cleaned dataset, with ResNet50 providing the best validation-based selection criterion and the best Brier score among the reported test models.

Calibration and threshold locking are important because many PCOS-AI reports emphasize accuracy or AUC but do not separate probability quality from discrimination. The selected image model's Brier score of 0.0139 and ECE of 0.0202 indicate close agreement between predicted probabilities and observed outcomes in the locked test set. The threshold of 0.9666 was selected using validation data and then frozen, avoiding test-driven operating-point optimization. This design is closer to the separation recommended in modern clinical prediction reporting guidance [16, 17]. Nonetheless, calibration established on a single public or retrospective distribution can drift when acquisition devices, sonographers, prevalence, or inclusion criteria change.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

The clinical branch provides a complementary result. Nested outer-fold ROC-AUC was 0.9501 with PR-AUC 0.9091, and the bootstrap interval for ROC-AUC was 0.9300–0.9674. These values are lower than several very high results reported on similar-sized public PCOS tables [10, 11, 13], but direct numerical ranking would be misleading. The present estimate comes from global nested selection, where feature group, model family, resampling, feature selection, hyperparameters, and threshold choice were re-selected within each outer-training fold. This better separates development decisions from assessment than fitting a final tuned pipeline on all data and reporting its apparent or same-split performance.

Selection variability across the outer folds is itself informative. Random forest with the full-safe feature set was selected three times, while SVM variants were selected twice. The result suggests that the cohort does not identify a single overwhelmingly dominant pipeline. This is a plausible consequence of limited sample size and correlated PCOS predictors. It also argues against overstating the final full-data random-forest refit as the uniquely optimal clinical model. The refit may be useful as a frozen candidate for future validation, but its threshold 0.44 does not create a new unbiased performance estimate.

Recent clinical studies show the value of stronger external designs. Yao et al. [14] reported closely matched internal and external AUCs from a multicenter laboratory cohort, while Zhao et al. [9] performed independent-center external testing for ultra-sound. These studies address a limitation that the current work cannot resolve: neither

branch here has a genuinely independent external cohort. Conversely, the current analysis makes the internal validity problem unusually explicit. The key contribution is therefore not a claim that its AUC exceeds prior work, but evidence that strong discrimination persists after duplicate control, development/test separation, nested selection, and uncertainty estimation.

Explainability also requires restraint. The Grad-CAM examples indicate which image regions influenced representative predictions, not whether those regions cor-respond to confirmed follicles, stromal changes, or causal pathology. Expert review of the highlighted regions was not documented. On the clinical side, the appropriate integrity decision was to omit SHAP results after the executed pipeline failed at the preprocessing transformation step. Reconstructing SHAP values after the fact without rerunning and validating the notebook would violate the study's source-of-truth rule. This failure is reported rather than concealed because reliable explainability depends on using the exact fitted transformation and model objects.

Runtime results suggest that model representation affects computational cost: TensorFlow Lite reduced file size and median CPU inference time substantially relative to the Keras model in the measured environment. These timings may be useful for planning future deployment studies, but they are not evidence of embedded feasibility, power efficiency, hospital integration, or real-time performance on an ultrasound console. Likewise, the 18.39 ms clinical random-forest timing describes one CPU software environment only

Limitations

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Several limitations define the boundary of the present evidence. The two datasets were independent and lacked verified cross-source patient linkage, so patient-level image–clinical fusion was impossible. The ultrasound collection lacked true patient identifiers; duplicate-group splitting reduced content leakage but cannot guarantee that different analysis groups came from different people. Dataset provenance, licensing information, and qualified manual review of image labels were not documented in the supplied notebook. Seventeen cross-label duplicate/orientation groups were detected and excluded, which resolves their direct use in modeling but also demonstrates uncertainty in the raw labels or source organization.

The final retained image cohort was modest after duplicate control, and the cluster-bootstrap intervals remain wide for some threshold-dependent metrics. The clinical cohort was also small, with no repeated-patient grouping variable and no indepen-dent external validation. Both branches are secondary-data benchmarks rather than prospective diagnostic studies. Distribution shift across populations, scanners, imaging protocols, laboratories, and diagnostic criteria is therefore unresolved.

The image explainability analysis was not independently reviewed by sonographers or gynecologists, so Grad-CAM cannot be interpreted as validated morphological localization. Clinical SHAP analysis could not be completed in the executed notebook because the fitted preprocessing object was unavailable in the attempted transformation path; no clinical feature-attribution result is claimed. Runtime measurements were CPU software benchmarks, not edge-device or hospital-workflow validation, and energy consumption was not measured. These limitations preclude describing the models as clinically validated diagnostic devices.

Conclusions

Two independent PCOS data sources were evaluated through separate leakage-controlled ultrasound and clinical–biochemical branches. The ultrasound audit showed that repeated and conflicting image content can materially reduce the effective dataset size and should be resolved before partitioning. After duplicate-group-safe splitting, validation-based architecture selection, probability calibration, and threshold locking, the selected ResNet50 retained high locked-test discrimination with cluster-bootstrap uncertainty quantified around the point estimates. The clinical branch achieved strong nested outer-fold performance under leakage-safe preprocessing and global nested model selection rather than relying on a full-data refit as an efficacy estimate.

The study supports leakage-aware validation as a central requirement for PCOS AI, particularly when public datasets contain repeated images or small structured cohorts. Its practical contribution is methodological rather than a claim of clinical readiness or multimodal fusion. The appropriate next step is independent, patient-identified external validation with documented dataset provenance, expert-reviewed labels and attributions, and prospectively defined operating thresholds.

Author Contributions

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

[EDITABLE PLACEHOLDER: specify verified author contribution roles for Hafiz Ullah, Aqib Hanif, Sunbal Rani, Ayesha Tariq, Hira Javed, and Syeda Sariya Shah using the target journal's required taxonomy.]

Funding
[EDITABLE PLACEHOLDER: provide the verified funding statement, including grant identifiers if applicable.]

Institutional Review Board Statement
[EDITABLE PLACEHOLDER: provide the verified ethics/IRB statement and approval identifier, or a justified exemption statement if applicable.]

Informed Consent Statement
[EDITABLE PLACEHOLDER: provide the verified informed-consent statement appropriate to the underlying datasets and study design.]

Data Availability Statement

The executed notebook records local/Google Drive paths for the two source datasets but does not document a persistent public repository identifier, complete provenance,
or redistribution/licensing terms.

[EDITABLE PLACEHOLDER: provide the verified dataset source/repository, access URL or accession identifier, and licensing/availability restrictions before submission.]

Conflicts of Interest
[EDITABLE PLACEHOLDER: provide the verified conflicts-of-interest declaration for all authors.]

References

Teede, H.J., Tay, C.T., Laven, J., 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., Network, I.P.: Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Fertility and Sterility **120**(4), 767–793 (2023) https://doi.org/10.1016/j.fertnstert.2023.07.025

Stener-Victorin, E., Teede, H., Norman, R.J., Legro, R., Goodarzi, M.O., Dokras, A., Laven, J., Hoeger, K., Piltonen, T.T.: Polycystic ovary syndrome. Nature Reviews Disease Primers **10**, 27 (2024) https://doi.org/10.1038/s41572-024-00511-3

Joham, A.E., Norman, R.J., Stener-Victorin, E., Legro, R.S., Franks, S., Moran, L.J., Boyle, J.A., Teede, H.J.: Polycystic ovary syndrome. The Lancet Diabetes & Endocrinology **10**(9), 668–680 (2022) https://doi.org/10.1016/S2213-8587(22)00163-2

Barrera, F.J., Brown, E.D.L., Rojo, A., Obeso, J., Plata, H.,

Online ISSN 3007-3197
Print ISSN 3007-3189

Lincango, E.P., Terry, N., Rodr´ıguez-Guti´errez, R., Hall, J.E., Shekhar, S.: Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. Frontiers in Endocrinology **14**, 1106625 (2023) https://doi.org/10. 3389/fendo.2023.1106625

Verma, P., Maan, P., Gautam, R., Arora, T.: Unveiling the role of artificial intelligence (ai) in polycystic ovary syndrome (pcos) diagnosis: A comprehensive review. Reproductive Sciences **31**(10), 2901–2915 (2024) https://doi.org/10.1007/s43032-024-01615-7

Suha, S.A., Islam, M.N.: An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Scientific Reports **12**, 17123 (2022) https://doi.org/10.1038/s41598-022-21724-0

Moral, P., Mustafi, D., Mustafi, A., Sahana, S.K.: Cystnet: An ai driven model for pcos detection using multilevel thresholding of ultrasound images. Scientific Reports **14**, 25012 (2024) https://doi.org/10.1038/s41598-024-75964-3

Kermanshahchi, J., Reddy, A.J., Xu, J., Mehrok, G.K., Nausheen, F.: Development of a machine learning-based model for accurate detection and classification of polycystic ovary syndrome on pelvic ultrasound. Cureus **16**(7), 65134 (2024) https://doi.org/10. 7759/cureus.65134

Zhao, B., Wen, L., Huang, Y., Fu, Y., Zhou, S., Liu, J., Liu, M., Li, Y.: A deep learning-based automatic recognition model for polycystic ovary ultrasound images. Balkan Medical Journal **42**(5), 419–428 (2025) https://doi.org/10.4274/balkanmedj.galenos.2025.

2025-5-114

Elmannai, H., El-Rashidy, N., Mashal, I., Alohali, M.A., Farag, S., El-Sappagh, S., Saleh, H.: Polycystic ovary syndrome detection machine learning model based on optimized