Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Toward trustworthy ovarian ultrasound frame classification through leakage-aware and calibrated deep learning

Kalakotla Arunima[1]

arunima.steffy18167@gmail.com

SR University

J Bhavana[2]

SR University

Uday Chandra S[3]

SR University

Article

Keywords: ovarian ultrasound, dataset leakage, duplicate detection, deep learning, calibration, reproducibility, PCOSGen

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11266008/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Toward trustworthy ovarian ultrasound frame classification through leakage-aware and calibrated deep learning

Kalakotla Arunima[1*], J Bhavana[2] and Uday Chandra S[3]

[1,2] School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India

[3] School of Engineering, SR University, Warangal, Telangana, India

Correspondence: arunima.steffy18167@gmail.com[1] , j.bhavana@sru.edu.in[2] , chandra.98145@gmail.com[3]

Abstract

Duplicate and visually similar images can inflate performance estimates when they cross dataset partitions, while test-guided calibration and threshold selection can introduce additional optimism. We developed an integrity-controlled workflow for frame-level ovarian ultrasound classification using the public PCOSGen training and test archives. Exact duplicates were identified with SHA-256 hashing, and cross-partition near copies were screened using perceptual hashing followed by correlation verification. The audit identified 347 near-copy train-test pairs and led to the exclusion of 309 contaminated development frames. Development data were partitioned by perceptual-similarity proxy groups. ResNet-18, MobileNetV3-Small and a shared-encoder global-local gated model were trained, calibrated and evaluated without using test labels for model selection, temperature scaling, threshold selection or ensemble weighting. On 1,448 cleaned held-out frames, the validation-weighted ensemble achieved 0.9309 accuracy, 0.9026 balanced accuracy, 0.9682 sensitivity, 0.8370 specificity, 0.8272 Matthews correlation coefficient, 0.9555 area under the receiver-operating-characteristic curve and 0.9765 average precision. Its Brier score was 0.0734. The ensemble improved the validation-selected sensitivity-specificity balance but did not improve discrimination significantly relative to MobileNetV3-Small (paired area-under-the-curve difference -0.00113, 95% interval -0.01027 to 0.00813; bootstrap $p = 0.78$). A quality-only control achieved an area under the curve of 0.8260, indicating residual acquisition-related information. These findings support integrity auditing, locked-test evaluation and calibrated reporting for public ultrasound benchmarks, while limiting interpretation to frame-level classification rather than patient-level diagnosis.

Keywords: ovarian ultrasound; dataset leakage; duplicate detection; deep learning; calibration; reproducibility; PCOSGen

Introduction

Polycystic ovary syndrome is a heterogeneous endocrine, reproductive and metabolic disorder. Contemporary guidelines require an integrated clinical assessment after exclusion of alternative causes, and ultrasound evidence of polycystic ovarian morphology represents only one component of the diagnostic process [1]. An image-only model can therefore classify the labels attached to ovarian ultrasound frames, but it cannot establish a patient-level diagnosis without menstrual, clinical, biochemical and demographic information. The present study consequently addresses frame-level ovarian ultrasound classification and does not claim autonomous diagnosis of polycystic ovary syndrome.

Ultrasonography is non-ionizing, widely available and comparatively inexpensive, but its interpretation is affected by speckle noise, weak boundaries, operator technique, acquisition settings and variable ovarian appearance. Deep learning can provide consistent image representations, yet apparent performance depends as much on the evaluation design as on the network architecture. Repeated frames, copied images and visually similar acquisitions can allow a network to exploit anatomy, overlays, compression patterns or device signatures shared across partitions rather than generalizable pathological information.

Prior ovarian ultrasound studies have used pretrained convolutional networks, handcrafted preprocessing, segmentation, super-resolution and ensemble learning [2-5,10-12]. Their reported results are not directly rankable because the datasets, endpoints, class definitions, experimental units and partitioning procedures differ. More generally, image-level rather than subject-level splitting has been shown to inflate medical image-classification performance [6,13], and data handling and test-set isolation remain important sources of bias in radiology artificial intelligence [7,8].

Accuracy and area under the receiver-operating-characteristic curve are also insufficient descriptions of reliability. Class imbalance can obscure clinically relevant errors, good ranking discrimination does not guarantee calibrated probabilities, and a threshold chosen after inspecting the test set produces an optimistic operating point [9,19]. Visual explanations do not independently establish that predictions are based on ovarian anatomy rather than image borders, overlays or acquisition characteristics. A defensible assessment should therefore combine integrity auditing, isolated validation, discrimination, calibration, class-specific performance, uncertainty analysis, confounding controls, computational cost and qualitative model inspection.

This study evaluates the reliability of performance estimates obtained from the public PCOSGen archives. The workflow audits exact duplicates, verifies cross-partition near copies, constructs perceptual-similarity proxy groups, trains calibrated transfer-learning models and locks all development decisions before final test evaluation. A shared-encoder global-local gated model is included as a candidate representation, but the primary contribution is the integrity-controlled evaluation protocol rather than a claim of universal architectural superiority.

Results

Dataset integrity audit

The PCOSGen training archive contained 3,200 labelled frames and the test archive contained 1,468 frames. Within the training archive, SHA-256 hashing identified 80 exact-duplicate groups. Sixty-four images from label-conflicting groups were removed, and 48 redundant same-label images were collapsed, leaving 3,088 candidate development frames. The test audit identified 19 exact-duplicate groups; removal of two label-conflicting images and 18 redundant same-label images produced a cleaned test set of 1,448 frames.

Cross-partition screening identified 347 verified near-copy pairs involving 309 unique development frames. Those 309 frames were excluded from model development, leaving 2,779 development frames. Because patient, examination, video, device and source identifiers were unavailable, perceptual-similarity groups were used only as conservative proxies for dependence and were not interpreted as patient groups.

Table 1. Final analytical partitions after integrity control

Partition	Purpose	Total	Negative	Positive
Training	Parameter estimation	2,225	626	1,599
Validation	Early stopping, calibration, threshold and ensemble selection	554	189	365
Test	Final locked evaluation	1,448	411	1,037

Validation-based model selection

All neural models completed training without numerical failure. MobileNetV3-Small and ResNet-18 stopped after nine epochs, whereas the gated model stopped after five epochs. After deterministic test-time augmentation and temperature scaling, the validation areas under the curve were 0.8348 for MobileNetV3-

Small, 0.8286 for ResNet-18 and 0.8238 for gated fusion. MobileNetV3-Small was therefore selected as the non-proposed ensemble member before test labels were accessed. Validation-derived ensemble weights were 0.5084 for MobileNetV3-Small and 0.4916 for gated fusion.

The lower and more variable validation discrimination of the gated model indicated that additional fusion capacity did not provide stable standalone improvement on the 554-frame validation partition. Validation areas under the curve were also markedly lower than test-set values, suggesting that the public test archive was easier or structurally different from the proxy-group validation split. The test result should not be interpreted as evidence of unusually strong generalization.

Held-out test performance

The validation-weighted ensemble achieved the highest threshold-dependent performance, with 0.9309 accuracy, 0.9026 balanced accuracy, 0.9682 sensitivity, 0.8370 specificity, 0.9526 F1 score and 0.8272 Matthews correlation coefficient. Its confusion matrix contained 344 true negatives, 67 false positives, 33 false negatives and 1,004 true positives. Relative to MobileNetV3-Small, the ensemble improved balanced accuracy by 3.36 percentage points and specificity by 6.81 points while introducing one additional false negative.

Table 2. Frame-level test performance at validation-selected thresholds

Model	Accuracy	Balanced accuracy	Sensitivity	Specificity	F1	MCC
Prevalence control	0.7162	0.5000	1.0000*	0.0000	0.8346	0.0000
Quality-only control	0.7783	0.7094	0.8689	0.5499	0.8488	0.4360
ResNet-18	0.8854	0.8370	0.9489	0.7251	0.9222	0.7091
MobileNetV3-Small	0.9123	0.8690	0.9691	0.7689	0.9406	0.7787
Gated fusion	0.8957	0.8405	0.9682	0.7129	0.9301	0.7352
Validation-weighted ensemble	0.9309	0.9026	0.9682	0.8370	0.9526	0.8272

*The prevalence control assigned every frame to the positive class; its sensitivity therefore has no discriminatory value.

The gated model did not outperform the lightweight baseline as a standalone classifier. Its usefulness in the ensemble therefore reflects complementary threshold-level errors rather than superior discrimination. Because the experiment used one training seed and did not include global-only, local-only and ungated-fusion ablations, the ensemble improvement cannot be attributed specifically to the gating mechanism.

Discrimination calibration and uncertainty

ResNet-18 produced the highest area under the curve (0.9574) and average precision (0.9838), while the ensemble produced the lowest Brier score (0.0734). MobileNetV3-Small achieved the lowest ten-bin expected calibration error among the neural models (0.0613). No model therefore dominated discrimination, calibration, operating-point performance and computational cost simultaneously.

Table 3. Discrimination calibration operating threshold and CPU inference cost

Model	AUROC (95% interval)	Average precision	Brier score	ECE	Threshold	ms per frame
ResNet-18	0.9574 (0.9472-0.9661)	0.9838	0.0830	0.0837	0.4977	184.1
MobileNetV3-Small	0.9567 (0.9445-0.9676)	0.9780	0.0799	0.0613	0.6112	35.6

Model	AUROC (95% interval)	Average precision	Brier score	ECE	Threshold	ms per frame
Gated fusion	0.9453 (0.9298-0.9591)	0.9691	0.0947	0.1340	0.3700	279.3
Validation-weighted ensemble	0.9555 (0.9416-0.9676)	0.9765	0.0734	0.1061	0.5366	314.9*

*Approximate sequential inference time for the two ensemble members; probability combination adds negligible cost.

For the ensemble, proxy-group bootstrap intervals were 0.9177-0.9433 for accuracy, 0.8831-0.9209 for balanced accuracy, 0.9564-0.9788 for sensitivity, 0.8004-0.8734 for specificity, 0.9428-0.9611 for the F1 score and 0.7920-0.8593 for the Matthews correlation coefficient. The area-under-the-curve interval was 0.9416-0.9676 and the average-precision interval was 0.9668-0.9849. These intervals quantify proxy-group sampling uncertainty and should not be interpreted as patient-level confidence intervals.

The prespecified paired comparison used MobileNetV3-Small because it was the validation-best baseline. The ensemble-minus-baseline area-under-the-curve difference was -0.00113, with a 95% paired-bootstrap interval from -0.01027 to 0.00813 and a two-sided bootstrap p value of 0.78. The study therefore did not demonstrate superior discrimination. The defensible comparative claim is limited to improved sensitivity-specificity balance at thresholds selected using validation data.

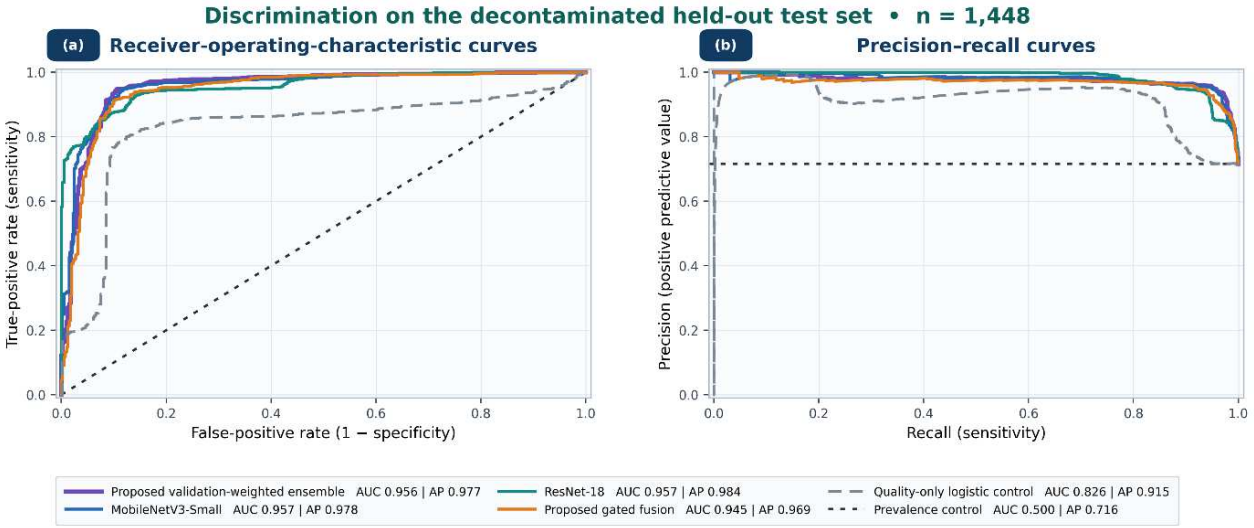

Figure 1. Receiver-operating-characteristic and precision-recall curves on the decontaminated held-out test set. The quality-only control demonstrates residual acquisition-related predictability, while differences among the neural models are small.

Temperature scaling produced values of 0.830 for MobileNetV3-Small, 0.901 for ResNet-18 and 1.620 for gated fusion. Although the ensemble improved the Brier score, its expected calibration error was worse than those of MobileNetV3-Small and ResNet-18. Calibration was therefore mixed rather than uniformly improved.

Decision analysis confounding controls and explanations

Decision-curve analysis showed positive net benefit for the neural models across a broad range of threshold probabilities relative to the treat-none strategy and, at higher thresholds, the treat-all strategy. Because the endpoint was a frame label and no clinical action was prespecified, these curves are exploratory and do not demonstrate patient benefit [20].

The quality-only logistic-regression control achieved 0.8260 area under the curve and 0.7094 balanced accuracy using file size, dimensions and grayscale summary features. Its performance was below that of the neural models but high enough to demonstrate residual source or acquisition-related information. High neural performance therefore cannot be attributed exclusively to ovarian morphology.

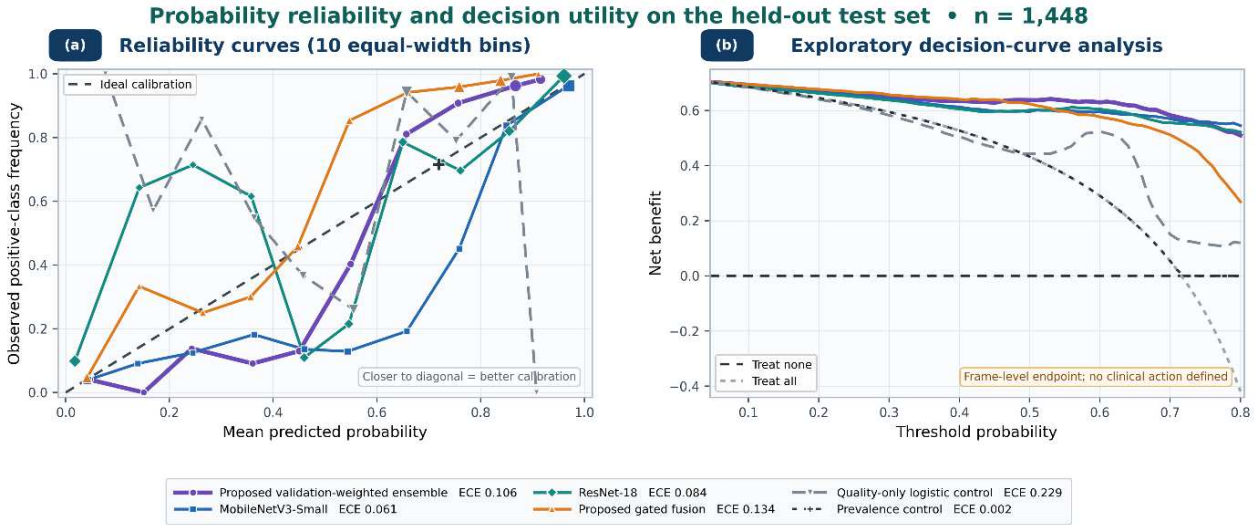

Figure 2. Probability reliability and exploratory decision analysis on the held-out test set. Calibration curves use ten equal-width bins; decision curves cover threshold probabilities from 0.05 to 0.80.

Grad-CAM maps sometimes highlighted echogenic or cystic-appearing image regions, but other maps emphasized edges, corners, measurement overlays and peripheral structures. The visualizations support qualitative model auditing but do not provide anatomical or causal validation. The montage used an exploratory cutoff of 0.50, which differs from the gated model's validation-derived threshold of 0.3700 and cannot reproduce the operating-point errors in Table 2. Expert ovarian-region annotation and quantitative overlap analysis would be required to evaluate anatomical relevance [21].

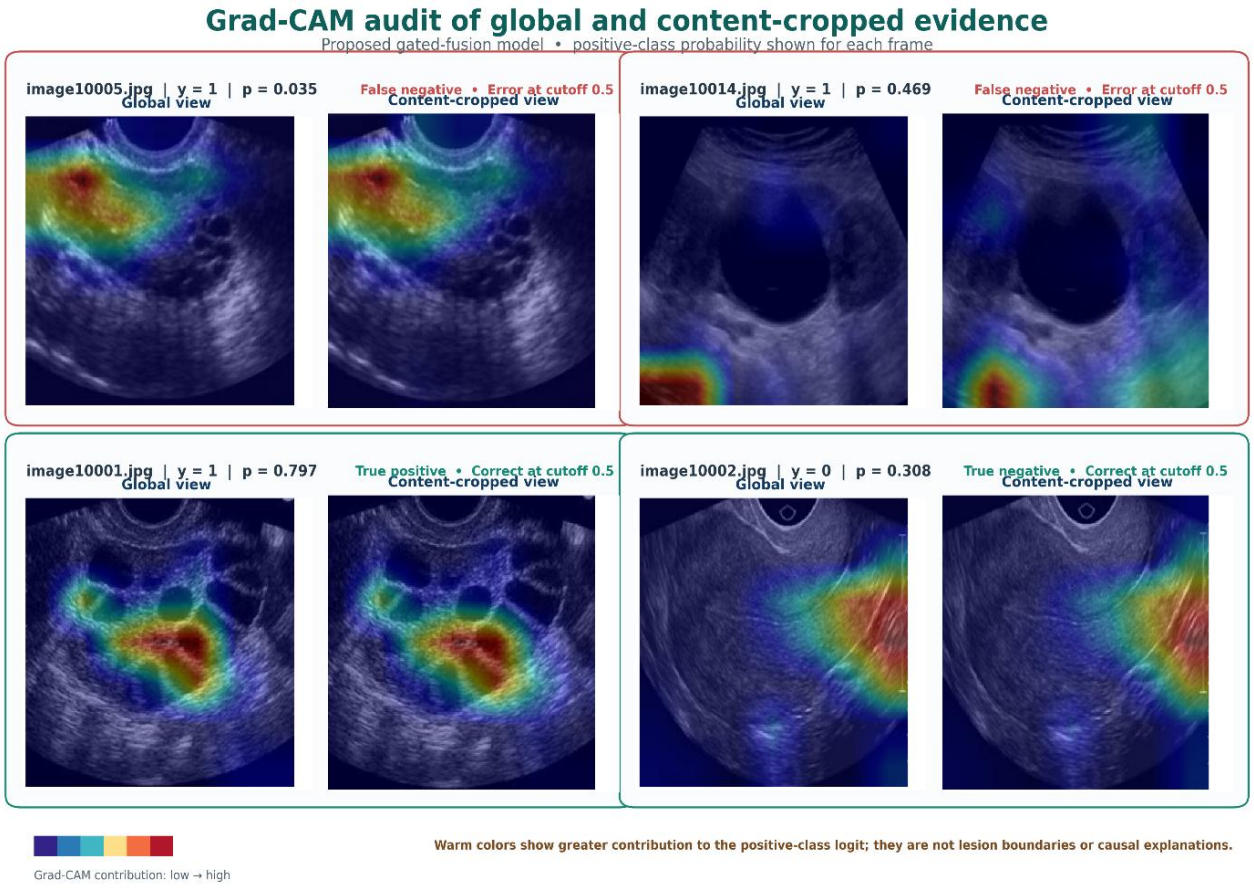

Figure 3. Representative Grad-CAM audits for global and content-cropped inputs. Red areas indicate greater contribution to the positive-class logit and should not be interpreted as causal lesions. Examples were selected using an exploratory cutoff of 0.50.

Computational efficiency

MobileNetV3-Small trained in 756.6 s (12.6 min) on the observed CPU environment, compared with 3,745.1 s (62.4 min) for ResNet-18 and 3,958.7 s (66.0 min) for gated fusion. Test inference required 35.6 ms per frame for MobileNetV3-Small, 184.1 ms for ResNet-18 and 279.3 ms for gated fusion. Sequential ensemble inference required approximately 314.9 ms per frame. The complete workflow, including auditing, training, prediction, bootstrap analysis, visualization and export, required 9,842.3 s (164 min). MobileNetV3-Small offered the best accuracy-runtime compromise, whereas the ensemble improved threshold-dependent performance at approximately nine times the per-frame CPU inference cost.

Discussion

The principal finding is that dataset integrity materially affects the interpretation of public ovarian ultrasound benchmarks. The audit identified conflicting exact duplicates and 347 cross-partition near-copy pairs that could otherwise permit shared content to influence both development and evaluation. After decontamination, all neural models retained areas under the curve above 0.94, demonstrating that high frame-level discrimination was not solely a consequence of the detected duplicate contamination.

The second finding is that performance was multidimensional. The ensemble achieved the best balanced accuracy, specificity, F1 score, Matthews correlation coefficient and Brier score, but ResNet-18 achieved the highest discrimination and MobileNetV3-Small achieved the best efficiency and expected calibration error. The lack of a significant area-under-the-curve difference prevents a claim that the ensemble is a superior discriminator. Its advantage is confined to the validation-selected operating point.

The quality-only control and the validation-test performance gap reveal continuing dataset limitations. An area under the curve above 0.8 from non-anatomical summary variables suggests that acquisition source, image dimensions, compression or other superficial properties remain associated with labels. The substantially higher test than validation performance further indicates distributional differences between the public archives. These findings argue against interpreting the reported scores as evidence of patient-level or cross-centre generalization.

The study also illustrates why post-development safeguards should be treated as part of the method. All calibration, threshold selection, ensemble-member selection and ensemble weighting were completed using validation data. Paired uncertainty analysis separated a positive threshold-dependent result from a non-significant discrimination difference. This evaluation discipline is transferable to other public medical-image collections in which repeated content and incomplete provenance are plausible.

Several limitations remain. PCOSGen contains internet-sourced frames without patient, examination, video, scanner, site, demographic, hormonal or clinical identifiers. Perceptual-hash groups reduce obvious visual dependence but do not establish independent patient sampling. The experiment used one training seed and did not include global-only, local-only, ungated-fusion or alternative-ensemble ablations. The Grad-CAM analysis was qualitative and did not use expert ovarian-region annotations. No independent clinical cohort was available. Finally, the healthy/unhealthy frame labels do not constitute Rotterdam-criteria diagnosis.

Future work should freeze the current pipeline and evaluate repeated random seeds, explicit fusion ablations and independent source-aware cohorts. Patient-aligned ultrasound, clinical and biochemical variables would permit clinically meaningful modelling and true patient-level partitioning. Device and demographic subgroup analysis, expert-annotated ovarian regions, prospective validation and calibration monitoring are necessary before any decision-support or deployment claim.

Methods

Study design and prediction target

This retrospective computational study evaluated binary frame-level ovarian ultrasound classification. The positive class denoted a frame labelled unhealthy or apparently abnormal, and the negative class denoted a frame labelled healthy or apparently normal. The target was selected because it was the common binary endpoint available in the supplied archives. The output was not designed to diagnose polycystic ovary syndrome, assess disease severity or replace integrated clinical evaluation.

Data sources label harmonization and integrity control

The PCOSGen training and test archives were obtained from their public Zenodo releases [14,15]. The training archive provided 3,200 frames (903 negative and 2,297 positive), and the test archive provided 1,468 frames (416 negative and 1,052 positive). Files were grayscale-decoded, matched to their supplied label tables and described by image dimensions, file size, grayscale mean and standard deviation, entropy and fractions of extreme black and white pixels. Labels were harmonized to 0 for apparently normal and 1 for apparently abnormal frames.

SHA-256 hashes identified byte-identical files. Label-conflicting duplicate groups were excluded, and same-label copies were collapsed. Cross-partition near-copy screening used a 64-bit perceptual hash. A training-test pair was considered a verified near copy when its Hamming distance was below four and the correlation between corresponding 96 x 96 grayscale images was at least 0.98. The screening used test-image pixels and file identifiers but not test labels. Development images involved in verified cross-partition near-copy relationships were excluded.

The remaining 2,779 development frames were divided using five-fold stratified group partitioning, with fold zero selected by the prespecified split seed 20260811. Exact perceptual-hash values served as proxy-group identifiers, preventing the same proxy group from appearing in training and validation. These groups are conservative similarity controls rather than patient identifiers.

Image preprocessing and global local views

Images were represented with three identical channels. The global view retained the complete frame, resized with aspect-ratio-preserving letterboxing to 224 x 224 pixels. For the local view, a grayscale foreground mask retained pixels above the higher of the 12th intensity percentile or intensity value 8. The detected content box was expanded by 4% on each side, cropped and letterboxed to 224 x 224 pixels. If fewer than 100 foreground pixels were detected, the full frame was retained.

Training augmentation included horizontal flipping with probability 0.5, rotation within 8 degrees, horizontal and vertical translation up to 4%, scaling from 0.94 to 1.06, brightness jitter of 0.12 and contrast jitter of 0.15. Validation and test preprocessing were deterministic. ImageNet channel means and standard deviations were used for normalization.

Models

MobileNetV3-Small [16] and ResNet-18 [17] served as global-view transfer-learning baselines. ImageNet pretrained weights were used, the classification heads were replaced with two-output linear layers, and the first 65% of backbone parameter tensors were frozen.

The candidate global-local model used a shared ImageNet-pretrained EfficientNet-B0 encoder [18]. Global and local batches were concatenated for one shared encoder call and then separated before adaptive average pooling, producing 1,280-dimensional representations for each view. The first 60% of encoder parameter tensors were frozen. A feature-wise sigmoid gate, generated from the concatenated representations by a 320-unit hidden layer with rectified-linear activation and dropout of 0.15, controlled the relative global and local contribution.

The classifier input concatenated three 1,280-dimensional components: the gated representation, the absolute difference between global and local features and their element-wise interaction. The resulting 3,840-dimensional vector was passed through layer normalization, a 384-unit fully connected layer with Gaussian error linear unit activation, dropout of 0.30 and a two-output classification layer.

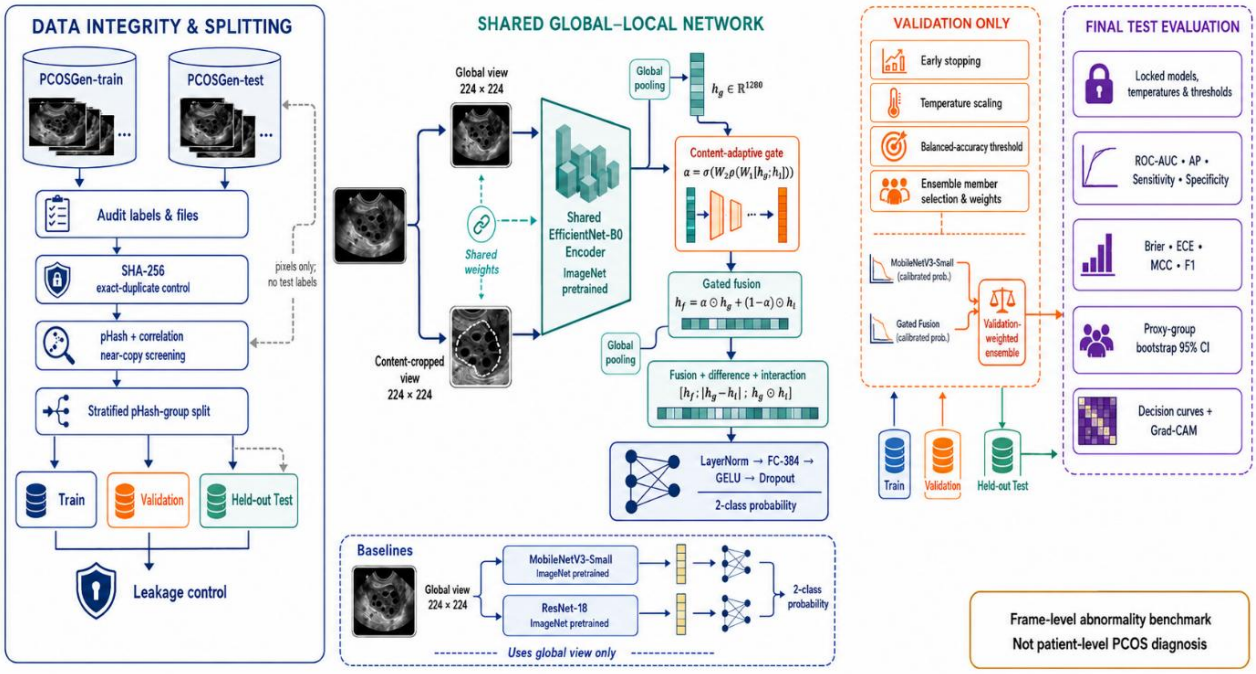

Figure 4. Integrity-controlled workflow and shared-encoder global-local gated architecture. Dashed boundaries separate validation-only operations from final locked-test evaluation.

Training calibration and ensemble construction

Models were optimized with class-weighted cross-entropy, label smoothing of 0.04, AdamW, an initial learning rate of 3 x 10^-4, weight decay of 10^-4, batch size 32 and gradient-norm clipping at 5. A cosine-annealing schedule was applied over a maximum of 12 epochs. Training stopped after four epochs without a validation area-under-the-curve improvement greater than 10^-4. The checkpoint with the highest validation area under the curve was retained. The low-runtime experiment used one prespecified model seed, 42, and was executed on a CPU. Logits from original and horizontally flipped inputs were averaged during validation and testing.

A non-image quality control used standardized file size, dimensions, grayscale mean, grayscale standard deviation, entropy and extreme-pixel fractions in logistic regression. A constant training-prevalence control assessed performance attributable to class imbalance.

Each neural model was temperature-scaled using validation logits by minimizing validation cross-entropy [19]. For each model, the operating threshold that maximized validation balanced accuracy was selected within the range 0.05-0.95. The proposed ensemble combined the gated model with the non-proposed candidate having the highest validation area under the curve. Each member's weight was proportional to its validation area under the curve minus 0.5, with a lower bound of 10^-6, and weights were normalized to sum to one. The ensemble threshold was selected from the combined validation probabilities. Test labels did not affect model fitting, early stopping, calibration, member selection, weighting or threshold selection.

Evaluation and statistical analysis

Evaluation included accuracy, balanced accuracy, sensitivity, specificity, positive and negative predictive values, F1 score, Matthews correlation coefficient, area under the receiver-operating-characteristic curve, average precision, Brier score and ten-bin expected calibration error. Receiver-operating-characteristic, precision-recall, reliability and exploratory decision curves were generated. Decision analysis covered threshold probabilities from 0.05 to 0.80 [20].

Ninety-five percent percentile intervals were estimated by bootstrap resampling whole perceptual-hash proxy groups rather than individual frames. The ensemble was compared with the validation-best baseline using the paired proxy-group bootstrap difference in area under the curve. Superiority was defined a priori as a paired 95% interval that excluded zero. These intervals quantify uncertainty under the available proxy grouping and are not patient-level intervals.

Grad-CAM was calculated from the final EfficientNet block for selected misclassified and correctly classified frames [21]. Separate maps were generated for global and local views. The maps were interpreted as qualitative attention audits rather than anatomical, causal or clinical explanations.

Reproducibility and ethical considerations

Input-archive hashes, preprocessing rules, hyperparameters, model seed, fitted temperatures, classification thresholds, bootstrap outputs, per-frame predictions, training histories, runtimes and checkpoints were recorded by the executable workflow. The analysis used previously released, de-identified image files and involved no prospective recruitment, intervention or access to direct personal identifiers. The absence of patient-level provenance limits clinical interpretation and does not permit verification of independent patient sampling.

Data availability

The source PCOSGen training and test datasets are publicly available from Zenodo at https://doi.org/10.5281/zenodo.10430727 and https://doi.org/10.5281/zenodo.10960327, respectively. The derived audit tables, exclusion manifests, split assignments and per-frame predictions are available from the corresponding author on reasonable request.

Code availability

The executable analysis code used for integrity auditing, model development, calibration and evaluation is available from the corresponding author on reasonable request.

Author contributions

K.A. led study conception, implementation, analysis, visualization and manuscript drafting. J.B. contributed to methodological review, interpretation and manuscript revision. U.C.S. supervised the study and contributed to interpretation and manuscript revision. All authors reviewed and approved the manuscript.

Funding

This research received no specific grant or financial support from any funding agency or authority in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

References

[1] Teede, H. J. et al. Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. J. Clin. Endocrinol. Metab. 108, 2447-2469 (2023). https://doi.org/10.1210/clinem/dgad463.

[2] Suha, S. A. & Islam, M. N. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Sci. Rep. 12, 21724 (2022). https://doi.org/10.1038/s41598-022-21724-0.

[3] Moral, P., Mustafi, D., Mustafi, A. & Sahana, S. K. CystNet: an AI-driven model for PCOS detection using multilevel thresholding of ultrasound images. Sci. Rep. 14 (2024). https://doi.org/10.1038/s41598-024-75964-3.

[4] Reka, S., Prabha, T. S., Prasanna, M., Reddy, V. N. N. & Amirtharajan, R. Automated high precision PCOS detection through a segment anything model on super resolution ultrasound ovary images. Sci. Rep. 15 (2025). https://doi.org/10.1038/s41598-025-01744-2.

[5] Borna, M. R., Saadat, H., Sepehri, M. M., Torkashvand, H., Torkashvand, L. & Pilehvari, S. AI-powered diagnosis of ovarian conditions: insights from a newly introduced ultrasound dataset. Front. Physiol. 16, 1520898 (2025). https://doi.org/10.3389/fphys.2025.1520898.

[6] Tampu, I. E., Eklund, A. & Haj-Hosseini, N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. Sci. Data 9 (2022). https://doi.org/10.1038/s41597-022-01618-6.

[7] Rouzrokh, P. et al. Mitigating bias in radiology machine learning: 1. Data handling. Radiol. Artif. Intell. 4, e210290 (2022). https://doi.org/10.1148/ryai.210290.

[8] Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4, 100804 (2023). https://doi.org/10.1016/j.patter.2023.100804.

[9] Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 17, 230 (2019). https://doi.org/10.1186/s12916-019-1466-7.

[10] Barrera, F. J. et al. Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. Front. Endocrinol. 14, 1106625 (2023). https://doi.org/10.3389/fendo.2023.1106625.

[11] Hosain, A. K. M. S., Mehedi, M. H. K. & Kabir, I. E. PCONet: a convolutional neural network architecture to detect polycystic ovary syndrome from ovarian ultrasound images. In 8th International Conference on Engineering and Emerging Technologies (ICEET) (IEEE, 2022). https://doi.org/10.1109/ICEET56468.2022.10007353.

[12] Divekar, A. & Sonawane, A. Leveraging AI for automatic classification of PCOS using ultrasound imaging. arXiv:2501.01984 (2025). https://doi.org/10.48550/arXiv.2501.01984.

[13] Yagis, E. et al. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. Sci. Rep. 11 (2021). https://doi.org/10.1038/s41598-021-01681-w.

[14] Handa, P., Saini, A., Dutta, S., Pathak, H., Choudhary, N., Goel, N. & Dhanao, J. K. PCOSGen-train dataset. Zenodo (2023). https://doi.org/10.5281/zenodo.10430727.

[15] Handa, P., Saini, A., Dutta, S., Pathak, H., Choudhary, N., Goel, N. & Dhanao, J. K. PCOSGen-test dataset. Zenodo (2024). https://doi.org/10.5281/zenodo.10960327.

[16] Howard, A. et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 1314-1324 (2019). https://doi.org/10.1109/ICCV.2019.00140.

[17] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778 (2016). https://doi.org/10.1109/CVPR.2016.90.

[18] Tan, M. & Le, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, 6105-6114 (PMLR, 2019).

[19] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, 1321-1330 (PMLR, 2017).

[20] Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. Med. Decis. Making 26, 565-574 (2006). https://doi.org/10.1177/0272989X06295361.

[21] Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 618-626 (2017). https://doi.org/10.1109/ICCV.2017.74.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PCOSAIMRevisionColab.ipynb