IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

Deep Learning for Automatic Recognition of Polycystic Ovary Syndrome in Ultrasound Images: A Review of Classification, Segmentation, and YOLO-based Detection Approaches

Amruta Vitthal Aware[1] and Dr. Rajshree Pankaj Wankhade[2]
[1] M.Tech. Student, Department of Electronics and Telecommunications Engineering
[2]Professor, Department of Electronics and Telecommunications Engineering
G. H. Raisoni University, Amravati, India
[1] amruta8990mp@gmail.com, [2] rajshree.wankhade@ghrua.edu.in

Abstract : *Polycystic ovary syndrome (PCOS) afflicts about one in ten women of reproductive age, and polycystic ovarian morphology (PCOM) detected through ultrasound is one of the three Rotterdam diagnostic criteria of this syndrome. The estimation of follicle count and ovarian volume in PCOS is currently performed manually. Both methods are rather time-consuming, operator-dependent, and highly inconsistent between different experts. Deep learning is one approach to automatically interpreting ovarian ultrasound results, and there has been an explosion of research on this topic since 2024. This systematic review analyzes deep learning applications to ultrasound images of the ovaries in PCOS patients, published from 2024 to 2026. Four methodological categories of research on this topic are identified: image-level classification using convolutional, ensemble, and transformer networks; segmentation of ovaries and follicles; object detection with You Only Look Once (YOLO) family detectors and YOLOv26 without non-maximum suppression; and multimodal models that fuse image features with other patient data. Comparisons are conducted using public datasets, evaluation measures, and validation strategies. Classification models typically yield accuracies in the range of 95-99.9%, but most of these studies rely on the same public dataset and often image-wise splits for training and testing. In an audit of the PCOSGen data used in the challenge, it was discovered that eliminating near-duplicate images reduces the number of images by over 95%. This means that some of the reported accuracies are likely inflated due to leakage. One study that reports external validation and comparison with sonographers is for a two-center YOLOv11 detector.*

Keywords: Deep learning, object detection, polycystic ovary syndrome, polycystic ovarian morphology, transvaginal ultrasound, vision transformer, YOLOv26.

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is the most frequent endocrinopathy among women in their reproductive age group. The repercussions of this disease extend beyond the ovaries. Women suffering from PCOS have a greater chance of being affected by insulin resistance, type 2 diabetes mellitus, dyslipidemia, cardiovascular disorders, endometrial hyperplasia, subfertility, anxiety, and depression [1]. A 2024 meta-analysis of 35 studies found the global prevalence of PCOS to be 9.2% [2]. Furthermore, a 2025 meta-analysis of 88 studies involving over 561,000 subjects found that the prevalence of this disease was highest when the Rotterdam criteria were used for diagnosis [3]. Even at the conservative level of estimation, PCOS affects more than a hundred million women, and a vast majority of them are diagnosed late or not at all.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

Diagnosis in adults is made in accordance with the Rotterdam consensus, which requires two of the following three characteristics: oligo- or anovulation, hyperandrogenism (clinical or biochemical), and PCOM (polycystic ovarian morphology). The international 2023 guideline on evidence-based diagnostics narrowed down the ultrasound criterion. If a vaginal transducer with 8 MHz frequency or more is used, then PCOM is diagnosed based on an FNPO of 20 follicles and/or an ovarian volume of 10 mL or more in either ovary without a corpus luteum, cyst, or a dominant follicle present [5]. The guideline also accepts AMH (Anti-Müllerian hormone) as a replacement for ultrasound diagnosis in adults. It excludes the use of ovarian morphology in adolescents, as multifollicular ovaries are physiological for this age [5]. A 2024 meta-analysis of adolescent populations revealed that the prevalence changes significantly if PCOM is excluded from the criterion [6]. Nevertheless, in adult females, ultrasound remains the most direct method for obtaining morphological data.

The problem is with the interpretation of the image. Counting up to twenty follicles with a diameter of 2 to 9 mm in one slice is a very tedious process, and the number of follicles counted depends on the transducer frequency, preset of the machine, angle of the probe, and operator skill. A meta-analysis published in 2024 revealed significant heterogeneity in the threshold values, transducer frequencies, and accuracies reported across different studies for ultrasound diagnostic criteria [7]. Problems with speckle noise, poor contrast of follicle borders, and touching/overlapping follicles only add to the difficulty.

Deep learning is very much tailored for pattern recognition tasks like this. CNNs learn texture and shape representations from raw pixel data, vision transformers capture long-range dependencies across the entire image, and one-shot object detection identifies and classifies an object in a single pass. A previous IEEE review on the use of machine learning for PCOS, involving 34 publications from 2015 to 2022, found that most research was focused on table-based clinical information; imaging datasets were few, unbalanced, and not standardized; ultrasound noise was not considered at all; and object detection methods like YOLO were underutilized [8].

Much has been achieved in this field since then. From 2024 to 2026, various methods were presented, including transfer learning ensembles optimized by genetic algorithms [9], light CNN and transformer fusion models with less than three million parameters [10], networks for follicle and cyst segmentation [11], [12], YOLO models for follicle and ovary detection [13], [14], and multimodal models combining ultrasound with clinical and hormonal information [15], [16]. Current accuracy values routinely surpass 98%. However, those figures are difficult to compare. The studies employ different datasets and splits, measure results with different metrics, and rarely test on an independent dataset from another hospital. An audit of the PCOSGen challenge dataset integrity in 2026 detected substantial contamination in the form of exact and near-duplicates [17], thus raising questions about the robustness of the results achieved to clean patient-level testing. On the other hand, the YOLO family switched to end-to-end architectures. In particular, YOLOv26 dispensed with NMS and DFL, and adopted a target-aware label assignment [18], [19].

The papers referenced in this paper were identified using IEEE Xplore, Scopus journals, PubMed, SpringerLink, and arXiv, with combinations of the following keywords: polycystic ovary syndrome, PCOM, ovarian ultrasound, follicle, deep learning, CNN, transformer, segmentation, and YOLO. This systematic review favors the inclusion of papers published between 2024 and 2026. Papers are included in this systematic review if they have made use of deep learning on either ovarian ultrasound or ovarian ultrasound features to detect PCOS and/or PCOM with quantified outcomes. Papers making use of other modalities apart from ovarian ultrasound or using questionnaires alone are not considered. Clinical guidelines, meta-analyses, and descriptions of the YOLO detection algorithm have been included as needed to aid understanding of the imaging papers. Preprints that contribute information which is otherwise unavailable have been included, especially regarding the PCOSGen integrity audit, and are noted as such in the reference list.

None of the reviews that currently exist connect these advancements while considering the way in which the models have been validated. This paper aims to do this. The contributions of this paper are as follows. Firstly, this paper categorizes the work related to the use of deep learning on PCOS ultrasound in the period 2024-2026 into a taxonomy by classification, segmentation, detection, and multimodal methods. Secondly, it discusses the publicly available datasets used for PCOS ultrasound and the integrity issues they face. Thirdly, it compares the reported results side by

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

side and correlates each number with the validation method used to obtain it. Fourthly, it traces the development of YOLO detectors from YOLOv8 to YOLOv26 and explains why the latest one works for ovarian ultrasound. Finally, it establishes an agenda for research for clinically viable PCOS recognition. A generic data-processing pipeline used by most of the reviewed systems is shown in Fig. 1.

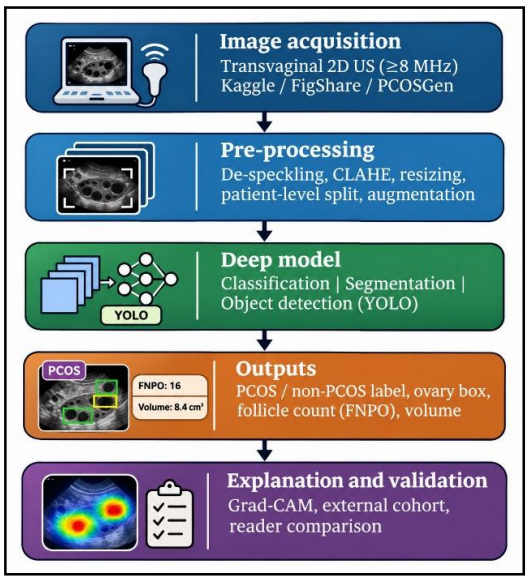

Fig. 1. Generic pipeline of deep learning-based PCOS recognition from ovarian ultrasound, from image acquisition to explanation and validation.

The rest of this paper is organized into five more sections. Section II describes the clinical definition of PCOM, the reasons behind variation in ovarian ultrasound data, and the datasets used. Section III provides an overview of the four categories of deep learning approaches, along with a discussion on YOLO detection in particular. Section IV covers the evaluation criteria and the results for each. Section V addresses open problems and directions for future work, while Section VI concludes this review.

II. CLINICAL BACKGROUND AND ULTRASOUND DATA RESOURCES

Any automated PCOS detection system needs to replicate the clinical criteria, operate on data that are both noisy and operator-dependent, and train on the relatively small amount of publicly available data. This chapter will consider each constraint one at a time, as they underlie several design considerations and limitations in the studies reviewed.

A. Diagnostic Criteria and Polycystic Ovarian Morphology

In the Rotterdam definition, PCOM is only one of three requirements, and PCOS can be present without it [4]. According to the 2023 guideline, PCOM is evaluated according to each ovary, not per image, and consists of the presence of more than 20 follicles of 2-9 mm in diameter and/or more than 10 mL ovarian volume, assessed using a transducer of at least 8 MHz without a corpus luteum, cyst, or dominant follicle [5]. Only follicle numbers per section and ovarian area can be used if it is impossible to conduct a complete ovarian examination [5]. The results of the meta-analysis by Pea et al. on the diagnostic value of ultrasound parameters have shown considerable heterogeneity among studies due to different equipment and counting methods [7].

Two implications can be drawn concerning model design. If the model assigns a label to an individual frame as either PCOS or normal, then it is performing a different task, since the decision under the guideline is based on the entire ovary and, for diagnosis, also on the patient's endocrine profile. Moreover, if the labels were obtained using a mixture of adolescents and adults, or transabdominal and transvaginal scans, then these labels do not match any of the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

definitions provided in the guidelines. Hence, a study such as the one done by Zhao et al. [14], in which labels are obtained from patients using the Rotterdam criteria and using only intracavitary scans, is much more clinically relevant.

B. Sources of Variability in Ovarian Ultrasound

Ovaries in ultrasound images exhibit multiplicative noise, acoustic shadowing, and poor boundaries between follicles and stroma. Chen et al. observed that the speckle noise blurs the boundary of follicles so much that the usual segmentation network under-segments the follicles, and they introduced an edge detection branch into the loss function to compensate for this issue [20]. The image appearance is further influenced by the transducer frequency, gain, depth, and dynamic range settings used on each device. The PCOSGen dataset, for example, was created using YouTube videos, ultrasoundcases.info, and Kaggle, whose origins are unknown [21]. Although documented heterogeneity aids generalization, undocumented heterogeneity does not indicate whether the model has learned the morphological structure of the ovaries or its own signature.

C. Public Datasets

The most commonly used databases in the studies reviewed are listed in Table I. The Kaggle database, comprising 3,856 images, is the most popular [22]. This database is used, for instance, in the EffiDenseGenOp ensemble [9], the VGG16 network designed by Shanmugavadivel et al. [23], and the PCOSFusionNet hybrid architecture [24]. FigShare database of 11,784 to 12,680 images was compiled from publicly available sources [25]. The PCOSGen database was published in the context of the Auto-PCOS challenge and consists of a training set with 3,200 images and a separate test set; it is labeled by a gynecologist [21], [26]. The database by Borna et al. includes 3 classes – normal, polycystic, and dominant-follicle ovaries [27], while DC-UNet was trained on approximately 42,000 two-dimensional slices of the USOVA3D follicle volumes [11]. The only prospectively acquired, guideline-labeled cohort of the reviewed studies that is not publicly available was created by Zhao et al. [14].

TABLE I: PUBLIC AND REFERENCE ULTRASOUND DATASETS USED IN PCOS STUDIES

Dataset	Size/classes	Remarks
Kaggle PCOS [22]	3,856 images; PCOS, normal	Most used benchmark; no patient IDs [9], [24]
FigShare [25]	11,784–12,680 images; binary	Aggregated from earlier public sets [10], [24]
PCOSGen [21]	3,200 train + test; healthy, unhealthy	Frames from web videos; heavy duplication [17]
Borna et al. [27]	Normal, PCO, dominant follicle	Three-class; best accuracy 76.2%
USOVA3D slices [11]	≈42,000 2D slices	Follicle masks; segmentation only
Zhao et al. [14]	933 patients, 1,722 images	Prospective, two centers; not public

The public does not create patient IDs, the transducer frequency, or the guidelines used in labeling. This is the reason why patient splitting, the typical protection against leakage in medical imaging, is almost impossible with public PCOS data.

III. DEEP LEARNING APPROACHES FOR PCOS ULTRASOUND ANALYSIS

The reviewed studies are classified into four groups based on their output, as presented in Fig. 2. The classification networks output one label for each image. The segmentation networks output masks of ovaries, follicles, or cysts, which can then be counted and measured by area. The detection networks output the bounding boxes labeled with their class labels, which enable the real-time localization of the ovary and its follicles. Multimodal models include image- or ultrasonography-based features alongside other clinical and laboratory variables.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

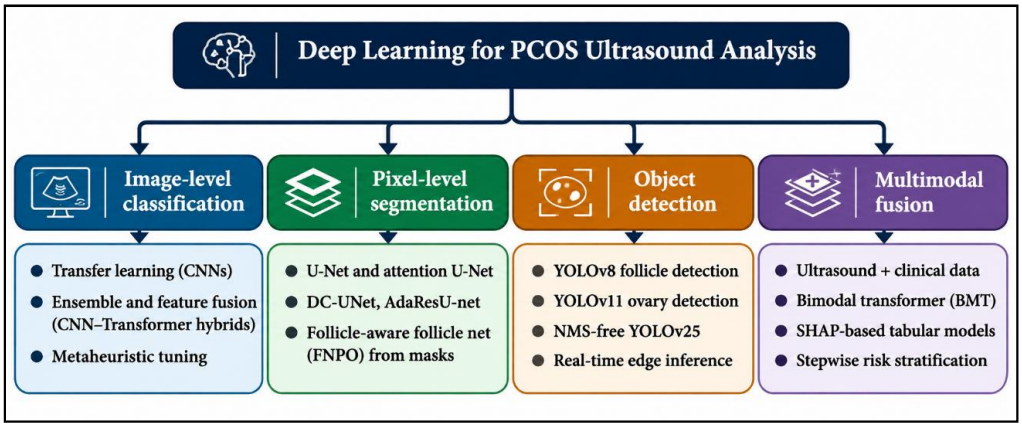

Fig. 2. Taxonomy of deep learning approaches for PCOS ultrasound analysis reviewed in this paper, with representative techniques in each family.

A. Image-Level Classification Networks

Transfer Learning using ImageNet pre-trained models is widely adopted. Shanmugavadivel et al. fine-tuned a VGG16 model on a dataset of 3,856 images from Kaggle. They achieved 98.29% accuracy by training another SVM model on 541 medical records, achieving 94.44% in the same study [23]. Sundari et al. designed an attention-based CNN model for TV ultrasound images, using Bayesian Hyperparameter optimization, SMOTE sampling, and Grad-CAM visualization, achieving 94.8% accuracy, 93.2% sensitivity, and 95.5% specificity [28]. Abrar et al. used a combination of DenseNet201 and ResNet50 on Kaggle images, along with SHAP, Grad-CAM, and LIME interpretations [29], based on gradient-weighted class activation mapping [30].

But results on more difficult data are less encouraging. On their novel three-class dataset, Borna et al. discovered that the best-performing network was ResNet18, achieving 76.2% accuracy and 78.2% macro-F1 score [27]. The distance between these numbers and the ones mentioned above – 98% to 99% achieved on the binary classification Kaggle task is indicative more of the differences in the datasets than in the neural networks. In another study, Kermanshahchi et al. reported 100% accuracy and 82.6% precision on 1,932 pelvic images [31], which is impossible to achieve on the same test set.

B. Hybrid, Ensemble and Transformer Models

The next batch of studies involves stacking or fusing multiple backbone networks. EffiDenseGenOp first preprocesses images using a fuzzy inference system and then stacks EfficientNetB7 and DenseNet201 via genetic optimization, achieving 99.58% accuracy and 98.97% AUC in K-fold cross-validation [9]. The CoCo-HoloNet model implements holographic convolutional dense blocks, regularized by tangent runner optimization, achieving more than 99% accuracy [32]. PCOSFusionNet fuses HOG descriptors and VGG19 features using CLAHE and watershed pre-processing, achieving 98.49% and 98.30% on the 3,856- and 12,680-image datasets, respectively [24]. MMFEDL-DPCOS fuses five CNN architectures with the help of stacked autoencoders and reaches 98.68% on 9,000 images from the FigShare dataset [33]. Tasnim et al. fused features from the ResNet-18 and GoogLeNet models using an AdaBoost meta-learner and achieved 97.36% accuracy, whereas a custom lightweight head achieved 99.19% accuracy [34]. CystNet pre-processes images using watershed and multilevel thresholding, achieving 96.54%-97.75% accuracy on the Kaggle dataset and 94.39% on PCOSGen [35].

Transformers emerged in 2025 and 2026. WomenSafe has two branches: MobileNetV2 and vision transformers, both of which include squeeze-and-excitation blocks, and they are merged via global average pooling. Using FigShare images (11,784 of them), it achieved 95.31% accuracy with MCC 89.08% using 2.89 million parameters and 0.28 GFLOPs

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

[10]. Shah and Yadav used features from the Swin Transformer and applied random forest and gradient boosting classifiers, achieving 99.87%, whereas the ResNet50 ensemble achieved only 99.2% [36]. Hoque et al. have used DenseNet121, Swin Transformer, and ConvNeXt to create DenConST and achieved 85.69%, then refined it into DenConREST [37]. What stands out about all of these models is that increased capacity yields only a tiny improvement in performance on a saturated benchmark while increasing computational cost.

C. Segmentation of Ovaries and Follicles

Transformers emerged in 2025 and 2026. WomenSafe has two branches: MobileNetV2 and vision transformers, both of which include squeeze-and-excitation blocks, and they are merged via global average pooling. Using FigShare images (11,784 of them), it achieved 95.31% accuracy with MCC 89.08% using 2.89 million parameters and 0.28 GFLOPs [10]. Shah and Yadav used features from the Swin Transformer and applied random forest and gradient boosting classifiers, achieving 99.87%, whereas the ResNet50 ensemble achieved only 99.2% [36]. Hoque et al. have used DenseNet121, Swin Transformer, and ConvNeXt to create DenConST and achieved 85.69%, then refined it into DenConREST [37]. What stands out about all of these models is that increased capacity yields only a tiny improvement in performance on a saturated benchmark while increasing computational cost.

D. Object Detection with YOLO Detectors

Object detection is located between classification and segmentation. Bounding boxes are much less expensive to annotate than masks; however, they still help localize the ovary and even individual follicles at a fine-grained level. In their study, Sowmiya et al. used YOLOv8 for follicle detection; post-processing was done using fuzzy c-means active contour, and GLCM texture features were extracted and compared to the classifiers. They achieved 95% accuracy with F-Net on 100 subjects and 97.5% on 200 subjects [13]. Zhao et al. trained YOLOv11 on a prospective two-center intracavitary cohort comprising 933 patients (781 in the first center and 152 in the second) and 1,722 images. In one pass, the model detected and classified the ovary, achieving an mAP of 95.7%-97.8%, an F1 of 95.0%-96.9%, and an AUC of 0.953-0.973 on internal and external sets, while reading images 50 times faster than an experienced sonographer [14].

The detection model has evolved very fast too. YOLOv10 made consistent dual-label assignment for training without NMS [39]. YOLOv11 incorporated the C3k2 and C2PSA position-sensitive attention [40], while YOLOv12 applied area attention with residual aggregation [41]. The end-to-end design becomes the default in YOLOv26, which excludes the DFL from the box head, gets rid of the NMS, balances the loss components incrementally during training (ProgLoss), applies label assignment with the small target-aware principle (STAL), and uses the MuSGD optimizer [18], [19]. Fig. 3 depicts the differences between YOLOv8/11 and YOLOv26 pipelines, while Table II describes the three generations.

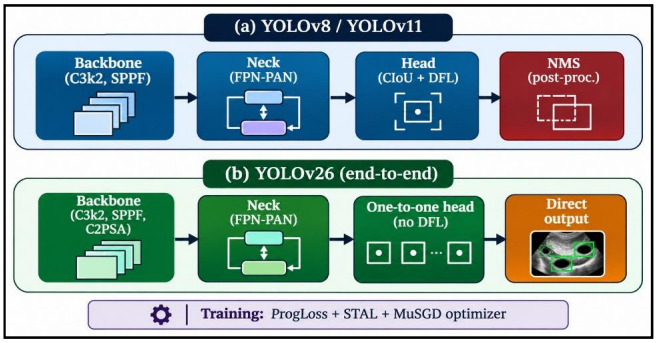

Fig. 3. Comparison of detection pipelines. (a) YOLOv8 and YOLOv11 use a DFL-based box head followed by NMS. (b) YOLOv26 uses an end-to-end one-to-one head without DFL or NMS and is trained with ProgLoss, STAL, and MuSGD. Adapted from [18], [40].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

TABLE II: COMPARISON OF YOLO GENERATIONS RELEVANT TO OVARIAN ULTRASOUND

Feature	YOLOv8	YOLOv11	YOLOv26
Key blocks	C2f, SPPF	C3k2, SPPF, C2PSA [40]	C3k2, SPPF, C2PSA [18]
Box loss	CIoU + DFL	CIoU + DFL	IoU loss, no DFL; ProgLoss
Assignment	Task-aligned	Task-aligned	STAL (small targets)
Post-processing	NMS	NMS	None (end-to-end)
Optimiser	SGD / AdamW	SGD / AdamW	MuSGD [19]
PCOS use	Follicles [13]	Ovary, two centers [14]	Not yet reported

Three characteristics make YOLOv26 suitable for use in ovarian ultrasound. STAL is designed to handle situations in which a couple of millimeter-sized follicles occupy only a few pixels. Eliminating NMS prevents the suppression of neighboring true positives, which is one of the challenges encountered when follicles overlap. The lightweight head works very efficiently on CPU to enable running the detector on the PC associated with the ultrasound machine without having a GPU [18]. As far as the literature goes, we have not come across any peer-reviewed PCOS study utilizing YOLOv26 at the time of writing.

E. Multimodal and Clinical-Data Fusion

Since PCOS is a syndrome and not an imaging pattern, some groups use ultrasound in combination with clinical information. Agirsoy and Oehschlaeger used XGBoost to train on Rotterdam-compatible clinical, biochemical, and ultrasound data, achieving an AUC of 0.9947 and an accuracy of 95.53% when AMH was included, along with SHAP values for interpretability [15]. AutoPCOS is a stratified risk assessment method that considers clinical, laboratory, and ultrasound information as it becomes available, using random forests at each step [16]. Both models use sonographer-obtained count and volume measurements from ultrasound, but not the raw pixel data from the scans. The combination of image representations with endocrine measurements in an end-to-end manner remains relatively untapped.

IV. EVALUATION PROTOCOLS AND COMPARATIVE ANALYSIS

Comparison of results from different studies is only possible if the metrics and validation methodology are known. In this part, first, the metrics used in the articles discussed will be defined; second, their results will be compared; and third, the validation methodology will be analyzed.

A. Performance Metrics

Classification evaluation methods derive their metrics from the counts of TP, TN, FP, and FN. The accuracy, precision, and recall can be calculated using formulas (1) and (2).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP}, \ \text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

The F1 score shown in (3) is the compromise between the two. Moreover, the Matthews correlation coefficient, as given in (4), incorporates all four elements of the confusion matrix and remains informative even in the presence of class imbalance, which is common in PCOS datasets.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

Detection studies first match predicted and ground-truth boxes by their intersection over union (IoU), defined in (5) for a predicted box Bp and a ground-truth box Bg. Average precision is the area under the precision-recall curve of each class, and the mean over N classes gives mAP in (6).

$$\text{IoU} = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \quad (5)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

$$\mathrm{AP}_c = \int_0^1 P_c(R)\,dR\,,\ \ \mathrm{mAP} = \frac{1}{N}\sum_{c=1}^{N} \mathrm{AP}_c \quad (6)$$

Accuracy, mAP, and AUC assess distinct aspects; therefore, it is not possible to compare 99% classifier accuracy to 96% mAP for a detector. Comparing the measures is fairer when reporting the AUC along with confidence intervals and MCC.

B. Comparative Analysis of Reported Results

Table III outlines the representative papers, the corresponding task, model, data, best performance, and validation scheme. Fig. 4 depicts the accuracy achieved by the image-based classification and detection-assisted approaches. There are three common trends. First, the accuracies on the binary Kaggle and FigShare benchmarks fall within 95%-99.9% range, allowing very little room for differentiation between the approaches. Second, the only approach that developed a novel three-class dataset with clinical labels reported 76.2% [27], which is significantly lower than other publicly available benchmarking results. Finally, there is not much benefit in using deeper ensembles. EffiDenseGenOp requires two large backbones and genetic optimization to achieve 99.58% [9], while WomenSafe achieves 95.31% with 2.89 million parameters [10].

TABLE III: SUMMARY OF REPRESENTATIVE DEEP LEARNING STUDIES ON PCOS ULTRASOUND

Study	Task	Method	Data	Best result	Validation
Shanmugavadivel [23]	Class.	VGG16 transfer learning; SVM on clinical data	Kaggle, 3,856 images	Acc. 98.29%	Internal split
Moral [35]	Class.	CystNet with multilevel thresholding	Kaggle; PCOSGen	Acc. 97.75%; 94.39%	Two public sets
Sowmiya [13]	Det. + class.	YOLOv8 follicles, active contour, GLCM, F-Net	100 and 200 subjects	Acc. 97.5%	Internal
Sarkar [11]	Seg.	DC-UNet double contraction path	USOVA3D, ≈42,000 slices	Automatic follicle masks	Internal
Ghosh [9]	Class.	FIS + EfficientNetB7 + DenseNet201, GA tuning	Kaggle, 3,856 images	Acc. 99.58%, AUC 98.97%	K-fold CV
Borna [27]	Class.	AlexNet, DenseNet121, ResNet18/34	New 3-class set	Acc. 76.2%, F1 78.2%	Internal
Zhao [14]	Det.	YOLOv11 ovary detection and classification	933 patients, 2 centers	mAP 95.7–97.8%, AUC 0.953–0.973	External + reader study
Natarajan [38]	Seg.	FCAU-Net attention U-Net	3,800 images	Acc. 99.89%	Internal
Agirsoy [15]	Multimodal	XGBoost on clinical, US, and AMH features	Tabular cohort	AUC 0.9947	Internal and external
Srikanth [10]	Class.	MobileNetV2-SE + MViT-SE (WomenSafe)	FigShare, 11,784 images	Acc. 95.31%, MCC 89.08%	80:20 split
Saha [24]	Class.	HOG + VGG19 fusion (PCOSFusionNet)	3,856 and 12,680 images	Acc. 98.49%	5-fold CV
Kranthi [33]	Class.	Five CNNs + stacked autoencoder	FigShare, 9,000 images	Acc. 98.68%	Internal
Tasnim [34]	Class.	ResNet-18 + GoogLeNet fusion	Public ultrasound set	Acc. 99.19%	Internal
Shah [36]	Class.	Swin features + RF/GB ensemble	Ultrasound images	Acc. 99.87%	Internal

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

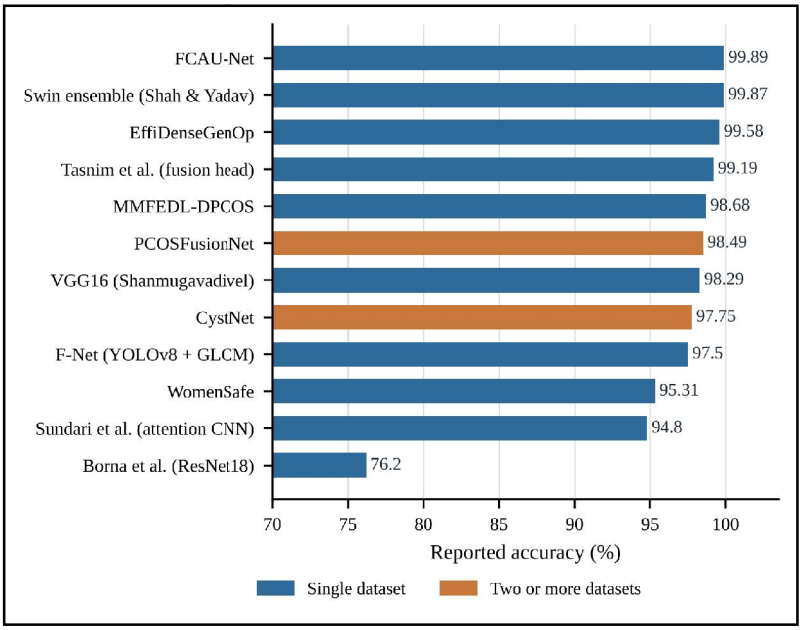

Fig. 4. Accuracy reported by representative image-based PCOS studies, compiled from [9], [10], [13], [23], [24], [27], [28], [33]–[36], [38].

C. Validation Practices and Data Leakage

The vast majority of reviewed studies randomly partition images into training and test datasets. When multiple images belong to the same patient or are taken from the same video clip, or when data augmentation happens before partitioning, then near-identical images will appear in both partitions. Bhuiyan et al. measured the extent of this bias for PCOSGen. Using perceptual hashing, the authors determined that excluding identical images from the 4,668 pooled images reduced the number to 3,471, while excluding near-identical images further reduced it to 225 images – a 95.2% decrease [17]. Such measurements have not been made for the Kaggle and FigShare datasets, which are built using overlapping datasets [22], [25]. Accordingly, near-perfect accuracies on these datasets should be considered upper bounds. By splitting patients rather than images and using the second center dataset for external validation, Zhao et al. managed to avoid this issue [14].

V. OPEN CHALLENGES AND FUTURE DIRECTIONS

The critique highlights certain challenges that hinder the adoption of these models in clinical practice. These challenges relate to the data used to train the models, the validation and explanation processes, and the environment in which they will finally be deployed. These challenges are linked to corresponding recommendations from the literature in Table IV.

TABLE IV: OPEN CHALLENGES AND SUGGESTED RESEARCH DIRECTIONS

Challenge	Effect	Direction
Duplicate and near-duplicate images	Inflated test accuracy	Hash-based deduplication; patient-level splits [17]
Folder-level, non-guideline labels	Proxy task, poor clinical match	Rotterdam/2023 labels per patient [5]
Single-source data	Unknown generalisation	Multicenter external tests [14]
Small follicles, speckle	Missed and merged targets	STAL, NMS-free detection [18]
Heavy ensembles	Hard to deploy	Lightweight hybrids, CPU export [10]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

Image-only input	Ignores endocrine criteria	Image-clinical fusion [15], [16]

A. Data Curation and Guideline-Aligned Labeling

Multicentric intracavitary image datasets annotated according to the 2023 guideline must be used for further research on a per-patient basis, and transducer frequency, patient age, and device age should be documented [5]. Perceptual hashing is an essential part of pre-processing and should be included in all research articles [17]. Annotating only the box levels of the ovaries and follicles, if possible, serves as a practical compromise between image labeling and masking [13], [14].

B. Generalization, Interpretability and Clinical Workflow

External validation at a second site, ideally prospective, is the minimal requirement [14]. Confidence intervals should be used, and correlated AUCs should be compared formally. Grad-CAM visualization is widespread [28], [29]. Still, its relevance to the area explored by the sonographer is seldom validated, and a heat map of the ovary does not mean that the network is counting follicles. The detection boxes are more amenable to auditing, as each box can either be validated or rejected by the clinician. The accuracy and time measures used by Zhao et al. provide the strongest evidence of clinical utility.

C. Efficient and Multimodal Deployment

Point-of-care requires compact models working with CPUs and embedded systems. Both lightweight hybrid solutions [10] and NMS-less detectors [18], [19] align with this direction. Exporting the YOLOv26 detection system to ONNX or a similar format, and measuring latency directly on the workstation with an ultrasound device, would convert efficiency claims into measurable facts. Outside imaging, the combination of detector embeddings and AMH, androgen, and menstrual cycle measurements can provide complete Rotterdam-criteria-based decision-making in a single model [15], [16].

VI. CONCLUSION

The present review is devoted to deep learning approaches for PCOS detection in ovarian ultrasound images during 2024-2026. Progress in the literature is shifting from single CNN-based classifiers to ensembles, hybrid CNN-transformers, segmentation models, YOLO detectors, and multimodal architectures. Accuracy scores on publicly available datasets range from 95% to 99.9%. These accuracy rates depend, however, on a few publicly available datasets, which have no patient and acquisition information, and at least one is highly contaminated with near-duplicate images. Clinical and externally obtained accuracy rates are lower and less common. Only one study validated on another dataset and compared the results with sonographers; it was a detection, not classification, study.

The result of this is a strong bias towards detection approaches which localize the ovary and the follicles within it, which are trained using guideline-based labeling of intracavitary scans partitioned by patients and validated using external groups. The use of YOLOv26, with its NMS-free head, small-target-aware label assignments, and CPU efficiency, is the obvious approach for this purpose, despite the lack of any peer-reviewed PCOS research that has used it to date. Future progress in the coming few years is unlikely to come from new backbones but rather from improved data and validation.

REFERENCES

[1] L. A. Owens and S. Franks, “Polycystic ovary syndrome: Origins and implications: The impact of polycystic ovary syndrome on reproductive health: A narrative review,” Reproduction, vol. 169, no. 5, Art. no. e240485, 2025, doi: 10.1530/REP-24-0485.

[2] N. Salari, A. Nankali, A. Ghanbari, S. Jafarpour, H. Ghasemi, S. Dokaneheifard, and M. Mohammadi, “Global prevalence of polycystic ovary syndrome in women worldwide: A comprehensive systematic review and meta-analysis,” Arch. Gynecol. Obstet., vol. 310, no. 3, pp. 1303–1314, Sep. 2024, doi: 10.1007/s00404-024-07607-x.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

[3] M. Amiri, S. Hatoum, R. P. Buyalos, A. Sheidaei, and R. Azziz, "The influence of study quality, age, and geographic factors on PCOS prevalence: A systematic review and meta-analysis," J. Clin. Endocrinol. Metab., vol. 110, no. 7, pp. 2082–2103, Jul. 2025, doi: 10.1210/clinem/dgae917.
[4] Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome," Fertil. Steril., vol. 81, no. 1, pp. 19–25, Jan. 2004, doi: 10.1016/j.fertnstert.2003.10.004.
[5] H. J. Teede, C. T. Tay, J. Laven, et al., "Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome," J. Clin. Endocrinol. Metab., vol. 108, no. 10, pp. 2447–2469, Oct. 2023, doi: 10.1210/clinem/dgad463.
[6] A. C. H. Neven, M. Forslund, S. Ranashinha, A. Mousa, C. T. Tay, A. Peña, S. Oberfield, S. Witchel, H. Teede, and J. A. Boyle, "Prevalence and accurate diagnosis of polycystic ovary syndrome in adolescents across world regions: A systematic review and meta-analysis," Eur. J. Endocrinol., vol. 191, no. 4, pp. S15–S27, 2024, doi: 10.1093/ejendo/lvae125.
[7] J. Pea, J. Bryan, C. Wan, et al., "Ultrasonographic criteria in the diagnosis of polycystic ovary syndrome: A systematic review and diagnostic meta-analysis," Hum. Reprod. Update, vol. 30, no. 1, pp. 109–130, 2024, doi: 10.1093/humupd/dmad027.
[8] S. Ahmed, M. S. Rahman, I. Jahan, M. S. Kaiser, A. S. M. S. Hosen, D. Ghimire, and S.-H. Kim, "A review on the detection techniques of polycystic ovary syndrome using machine learning," IEEE Access, vol. 11, pp. 86522–86543, Aug. 2023, doi: 10.1109/ACCESS.2023.3304536.
[9] A. Ghosh and K. Srinivasan, "EffiDenseGenOp: Ensemble transfer learning with hyperparameter tuning using genetic algorithm optimization for PCOS detection from ultrasound sonography images," IEEE Access, vol. 13, pp. 54285–54313, Mar. 2025, doi: 10.1109/ACCESS.2025.3553895.
[10] P. Srikanth, M. Gatika, R. S. Rao, and M. R. Holla, "WomenSafe: An ultra-lightweight convolution-vision transformer fusion model for PCOS diagnosis," IEEE Access, vol. 14, pp. 24841–24853, Feb. 2026, doi: 10.1109/ACCESS.2026.3664790.
[11] M. Sarkar, A. Mandal, and A. Tudu, "DC-UNet: Looking for follicles in the ovarian ultrasound images," Franklin Open, vol. 8, Art. no. 100149, Sep. 2024, doi: 10.1016/j.fraope.2024.100149.
[12] M. Sha, "Segmentation of ovarian cyst in ultrasound images using AdaResU-net with optimization algorithm and deep learning model," Sci. Rep., vol. 14, no. 1, Art. no. 18868, Aug. 2024, doi: 10.1038/s41598-024-69427-y.
[13] S. Sowmiya, S. Umapathy, O. Alhajlah, F. Almutairi, S. Aslam, and R. K. Ahalya, "F-Net: Follicles Net an efficient tool for the diagnosis of polycystic ovarian syndrome using deep learning techniques," PLOS ONE, vol. 19, no. 8, Art. no. e0307571, Aug. 2024, doi: 10.1371/journal.pone.0307571.
[14] B. Zhao, L. Wen, Y. Huang, Y. Fu, S. Zhou, J. Liu, M. Liu, and Y. Li, "A deep learning-based automatic recognition model for polycystic ovary ultrasound images," Balkan Med. J., vol. 42, no. 5, pp. 419–428, 2025, doi: 10.4274/balkanmedj.galenos.2025.2025-5-114.
[15] M. Agirsoy and M. A. Oehlschlaeger, "A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features," Sci. Rep., vol. 15, no. 1, Art. no. 33638, 2025, doi: 10.1038/s41598-025-10453-9.
[16] J. Hou, Y. Duan, W. Zhao, Q. Sun, and J. Wang, "AutoPCOS: A stepwise multimodal intelligent framework for polycystic ovary syndrome risk stratification and diagnostic support," Front. Endocrinol., vol. 17, Art. no. 1760541, 2026, doi: 10.3389/fendo.2026.1760541.
[17] S. T. Bhuiyan, R. Rahman, S. Wasi, H. Khatun, A. Islam, A. K. M. M. Rahman, S. B. Alam, and M. A. Amin, "PelviNeXt: A modality-agnostic hybrid network for pelvic imaging in women's health," arXiv:2608.20144, 2026, doi: 10.48550/arXiv.2608.20144.
[18] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection," arXiv:2509.25164, 2025, doi: 10.48550/arXiv.2509.25164.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

[19] G. Jocher and J. Qiu, "Ultralytics YOLO26," Ultralytics, Jan. 2026. [Online]. Available: https://github.com/ultralytics/ultralytics
[20] Z. Chen, C. Zhang, Z. Li, J. Yang, and H. Deng, "Automatic segmentation of ovarian follicles using a deep neural network combined with edge information," Front. Reprod. Health, vol. 4, Art. no. 877216, Aug. 2022, doi: 10.3389/frph.2022.877216.
[21] Auto-PCOS Classification Challenge Organizers, "PCOSGen-train dataset (v2)," Delhi Technological Univ. and Indira Gandhi Delhi Tech. Univ. for Women, Zenodo, 2025, doi: 10.5281/zenodo.14592001.
[22] A. Choudhari and A. Korde, "PCOS detection using ultrasound images," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/anaghachoudhari/pcos-detection-using-ultrasound-images
[23] K. Shanmugavadivel, M. S. Murali Dhar, T. R. Mahesh, T. Al-Shehari, N. A. Alsadhan, and T. E. Yimer, "Optimized polycystic ovarian disease prognosis and classification using AI based computational approaches on multi-modality data," BMC Med. Inform. Decis. Mak., vol. 24, no. 1, Art. no. 281, 2024, doi: 10.1186/s12911-024-02688-9.
[24] D. Saha, A. Mandal, S. K. Biswas, B. Das, A. Bhattacharya, and A. K. Das, "PCOSFusionNet: Hybrid deep feature fusion network for PCOS classification from ultrasound images of ovaries," Ultrason. Imaging, vol. 48, no. 5, pp. 358–377, 2026, doi: 10.1177/01617346261416509.
[25] A. Indirani, "PCOS dataset," figshare, Nov. 2024. [Online]. Available: https://figshare.com/articles/dataset/PCOS_Dataset/27682557
[26] A. Divekar and A. Sonawane, "Leveraging AI for automatic classification of PCOS using ultrasound imaging," arXiv:2501.01984, Dec. 2024, doi: 10.48550/arXiv.2501.01984.
[27] M.-R. Borna, H. Saadat, M. M. Sepehri, H. Torkashvand, L. Torkashvand, and S. Pilehvari, "AI-powered diagnosis of ovarian conditions: Insights from a newly introduced ultrasound dataset," Front. Physiol., vol. 16, Art. no. 1520898, Jul. 2025, doi: 10.3389/fphys.2025.1520898.
[28] M. S. Sundari et al., "Transfer learning-enhanced CNN model for integrative ultrasound and biomarker-based diagnosis of polycystic ovarian disease," Sci. Rep., vol. 15, no. 1, Art. no. 34519, 2025, doi: 10.1038/s41598-025-17711-w.
[29] S. Abrar, S. Rahman, I. J. Momo, M. T. Samiha, B. M. S. Alam, M. T. Noor, and N. T. Niloy, "Smart diagnosis and early intervention in PCOS: A deep learning approach to women's reproductive health," arXiv:2602.04944, 2026, doi: 10.48550/arXiv.2602.04944.
[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[31] J. Kermanshahchi, A. J. Reddy, J. Xu, G. K. Mehrok, and F. Nausheen, "Development of a machine learning-based model for accurate detection and classification of polycystic ovary syndrome on pelvic ultrasound," Cureus, vol. 16, no. 7, Art. no. e65134, Jul. 2024, doi: 10.7759/cureus.65134.
[32] P. Jeyashanker et al., "Advanced holographic convolutional dense networks and Tangent runner optimization for enhanced polycystic ovarian disease classification," Sci. Rep., vol. 15, no. 1, Art. no. 15719, 2025, doi: 10.1038/s41598-025-98873-5.
[33] S. Kranthi, Y. Sandeep, K. Pranathi, E. L. Lydia, G. P. Joshi, and W. Cho, "A multimodal feature fusion with deep representation learning approach for polycystic ovary syndrome diagnosis using ultrasound images," Sci. Rep., vol. 16, no. 1, Art. no. 9977, 2026, doi: 10.1038/s41598-026-40718-w.
[34] M. Tasnim, S. Das, S. S. Tonoy, and T. M. S. Sazzad, "PCOS detection from ovarian ultrasound images using feature fusion and hybrid classification approaches," Health Inf. Sci. Syst., vol. 14, no. 1, Art. no. 75, 2026, doi: 10.1007/s13755-026-00472-9.
[35] P. Moral, D. Mustafi, A. Mustafi, and S. K. Sahana, "CystNet: An AI driven model for PCOS detection using multilevel thresholding of ultrasound images," Sci. Rep., vol. 14, no. 1, Art. no. 25012, Oct. 2024, doi: 10.1038/s41598-024-75964-3.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

[36] I. B. Shah and P. K. Yadav, "Automated PCOS detection using fine-grained deep feature extraction and explainable AI: A transformer-based ensemble approach," Int. J. Image, Graph. Signal Process., vol. 18, no. 1, pp. 88–105, Feb. 2026, doi: 10.5815/ijigsp.2026.01.06.

[37] M. M. Hoque, M. M. Hassain, M. Rahaman, M. T. Islam, S. Rani, and M. S. Mollah, "Vision models for medical imaging: A hybrid approach for PCOS detection from ultrasound scans," arXiv:2601.15119, 2026, doi: 10.48550/arXiv.2601.15119.

[38] Y. Natarajan, K. R. Sri Preethaa, and M. Shyamala Devi, "Feature fusion context attention gate UNet for detection of polycystic ovary syndrome," Sci. Rep., vol. 15, no. 1, Art. no. 42386, 2025, doi: 10.1038/s41598-025-26314-4.

[39] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 37, 2024, doi: 10.48550/arXiv.2405.14458.

[40] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," arXiv:2410.17725, 2024, doi: 10.48550/arXiv.2410.17725.

[41] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 38, 2025, doi: 10.48550/arXiv.2502.12524.