

AUTOMATED LOCALIZATION OF AMNIOTIC FLUID POCKETS IN OBSTETRIC ULTRASOUND IMAGES USING YOLOV8: A DOCUMENTED DETECTION BASELINE

Malik Muhammad Arif^{*1}, Sadiq Ali², Arzoo Bangash³, Daniyal Ahmed Khan⁴,
Muhammad Amin⁵

^{*1, 2, 4, 5}Electrical Engineering Department, UET, Peshawar, Pakistan

³Peshawar, Pakistan

^{*1}engrmaalikarif@uetpeshawar.edu.pk, ²sadiqali@uetpeshawar.edu.pk, ³arzoo.bangash@kmc.edu.pk,

⁴17pwmec4340@uetpeshawar.edu.pk, ⁵muhammadamin0334@gmail.com

DOI: <https://doi.org/10.5281/zenodo.22844042>

Keywords

amniotic fluid; obstetric ultrasound; YOLOv8; object detection; deep learning; fetal assessment

Article History

Received: 28 January 2026

Accepted: 13 March 2026

Published: 27 March 2026

Copyright @Author

Corresponding Author: *

Malik Muhammad Arif

Abstract

Amniotic fluid volume (AFV) is a routine, non-invasive indicator of fetal well-being, but conventional ultrasound assessment depends strongly on sonographer skill and pocket selection. In this work, we report a deep-learning baseline for localizing visible amniotic-fluid pockets in two-dimensional obstetric ultrasound images. The primary public dataset comprised 3,335 images: 1,570 from the Zenodo Fetal Planes Database and 1,765 from a Roboflow fluid-detection dataset. A further approximately 120 locally collected clinical images were also available, intended primarily for external robustness assessment rather than for training. The 3,335 public images were partitioned using a 70:20:10 split into 2,331 training images, 667 validation images, and 337 independent test images, and a YOLOv8s model was trained on 640×640 -pixel inputs with a batch size of 16. Validation precision reached 0.9973, recall 1.0000, mAP50 0.9950, and mAP50-95 0.7678. The model output is fluid-pocket localization only: pixel-level segmentation, calibrated Single Deepest Pocket measurement, and clinical classification were not established as validated outcomes in the available experimental record. Accordingly, the contribution here is a documented amniotic-fluid localization baseline built on the full 3,335-image public dataset, with a clear path toward independent testing, segmentation, and clinical measurement, rather than a complete automated AFV-assessment system.

I. INTRODUCTION

Amniotic fluid is the protective medium surrounding the fetus throughout gestation. It contributes to mechanical protection, temperature stability, immune support, and the muscular and pulmonary movements required for normal fetal development. Its volume reflects fetal swallowing and urination, placental function, and membrane integrity, and is consequently

used as a non-invasive component of prenatal surveillance [10], [11]–[13].

Clinical assessment commonly uses the Maximum Vertical Pocket (MVP), Amniotic Fluid Index (AFI), and Single Deepest Pocket (SDP) methods. AFI sums the deepest vertical pocket measurements from four uterine quadrants, whereas SDP selects the single deepest suitable pocket. Prior evidence indicates that SDP can provide screening performance comparable to AFI while reducing false-positive

oligohydramnios diagnoses and unnecessary interventions [11]. These procedures nevertheless depend on correct pocket selection and accurate caliper placement, which is where inter-observer variability and measurement uncertainty creep in.

Reduced and excessive amniotic fluid can both be clinically important. Oligohydramnios is associated with cord compression, fetal growth concerns, and, in severe or prolonged cases, pulmonary hypoplasia; polyhydramnios is associated with preterm birth, malpresentation, delivery complications, and cord-related risk [10], [12], [13]. Improved reproducibility at the image-analysis stage could assist clinical workflows without replacing expert interpretation.

Deep-learning methods have expanded fetal-ultrasound automation from handcrafted image processing and classical machine learning to convolutional neural networks, U-Net-based segmentation, Mask R-CNN, and YOLO-family detection [1], [3]–[7]. But most prior studies use different task definitions, datasets, and validation protocols, which makes cross-study comparison difficult. The intended end-to-end pipeline covers segmentation, SDP estimation, and clinical classification; the present paper reports the experimentally supported stage of that pipeline, using the full 3,335-image public corpus under a 70:20:10 training/validation/test split, and does not draw on any other, differently sized experimental run. The key contributions of this work are: (1) a documented YOLOv8s baseline for localizing amniotic-fluid pockets in two-dimensional obstetric ultrasound, trained on the training subset of the full 3,335-image public corpus and internally evaluated on the validation subset under a 70:20:10 split, with an independent test subset reserved for final evaluation; (2) a report of the documented training configuration and internal-validation results for the experiment; and (3) a clear separation between demonstrated detection performance and the downstream segmentation, measurement, and classification tasks that remain to be validated.

II. RELATED WORK

Prior amniotic-fluid studies have mainly followed two directions: pixel-level segmentation and end-to-end assessment. U-Net-style networks are a standard approach to biomedical

segmentation, while Mask R-CNN provides instance-level delineation of multiple structures [3], [4]. Prior work also reviews three-dimensional models for placenta, fetus, and amniotic-fluid segmentation, demonstrating the potential of deep learning when detailed masks and volumetric data are available [9], [10].

YOLO-based methods emphasize efficient localization. A related YOLOv8 study combined fluid detection with contour analysis and SDP estimation, reporting precision 0.84, recall 0.764, and mAP50 0.84 on 3,335 images [2]. These figures are not directly comparable with the present experiment, since the task, dataset, annotation format, and pipeline all differ. In contrast, the present study isolates the bounding-box localization stage and reports its internal validation performance independently of segmentation, SDP estimation, and downstream clinical classification, rather than proposing a new detector architecture.

Related fetal-monitoring research also includes wearable sensing with accelerometers, gyroscopes, acoustic sensors, and piezoelectric sensors. Prior work summarizes an IoT-enabled accelerometer/gyroscope system with 90.0% sensitivity and an F1-score of 88.56%, while noting class imbalance and maternal-perception-based labeling [8]. This line of work is relevant to longer-term multimodal maternal-fetal monitoring but sits outside the image-detection task implemented here. Explainable deep-learning approaches that fuse multiple physiological data modalities have also shown strong diagnostic performance in other clinical imaging domains, such as cardiovascular disease detection from ECG images, underscoring the broader relevance of interpretable multimodal pipelines to medical imaging tasks [14].

A central methodological gap in this literature is the lack of separation between pipeline stages. Detection locates an approximate region; segmentation determines the true fluid boundary; SDP measurement requires physical calibration; and clinical classification requires validated patient-level reference labels. The present paper addresses this gap directly, by reporting only the stage actually evaluated rather than converting a detection result into an unsupported clinical claim.

III. PROPOSED METHODOLOGY

A. Dataset / Materials

The primary model-development corpus comprised 3,335 two-dimensional obstetric ultrasound images: 1,570 from the Zenodo Fetal Planes Database and 1,765 from Roboflow. A further approximately 120 locally collected clinical images were also available, giving

approximately 3,335 images in the broader study pool; these local images were not included in the training, validation, or test metrics reported in this paper and are noted here only as a resource for future external robustness evaluation. The 3,335 public images were divided at a 70:20:10 ratio into 2,331 training images, 667 validation images, and 337 independent test images.



Fig. 1a. Representative obstetric ultrasound frames from the training corpus, illustrating variability in amniotic-fluid pocket appearance, probe orientation, and image quality.

The dataset preparation protocol included duplicate removal, exclusion of corrupted or poor-quality images, and patient- or examination-level grouping before splitting, to reduce information leakage. This matters because multiple frames from a single examination can otherwise inflate performance estimates in a way that looks better than it is. The dataset protocol specifies that partitioning should occur at the patient or ultrasound-examination level, although a complete patient-level audit table is not included.

The full dataset methodology describes orientation correction, grayscale conversion, intensity normalization, contrast enhancement, and controlled augmentation applied only to the training subset. Images were standardized to 640×640 pixels. Training-only augmentation was intended to simulate realistic acquisition variability without touching the validation and test distributions.

B. Model / Algorithm

YOLOv8 is a single-stage, anchor-free detector designed for efficient multi-scale object

localization [1]. In the present application, the model predicts a single class, fluid, corresponding to visible amniotic-fluid pockets. Transfer learning is used to initialize the network from pretrained weights before fine-tuning on the ultrasound training set. The larger training corpus is intended to expose the detector to greater variation in anatomy, image quality, pocket appearance, and acquisition conditions.

The reported experiment uses the full public dataset: a 70:20:10 partition of 3,335 images, corresponding to 2,331 training, 667 validation, and 337 test images. The model was trained with batch size 16 on 640×640 -pixel inputs, as stated in the Abstract and Dataset subsection. The remaining training parameters (optimizer, learning-rate schedule, epoch count, augmentation policy, random seed, and software/hardware versions) were not available in the experimental record and are therefore not reported.

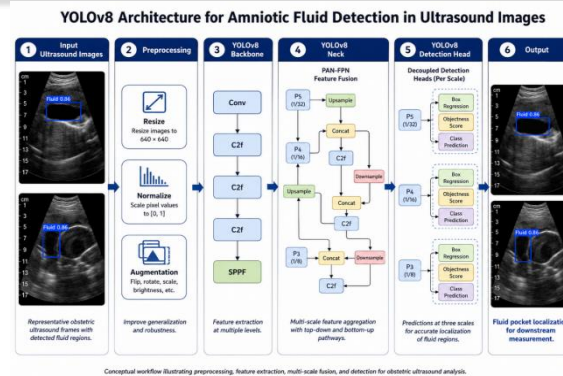


Fig. 1. YOLOv8s processing flow for amniotic-fluid detection in ultrasound images.

C. Experimental Setup

Model development was performed in a GPU-enabled Google Colab environment. The training session used approximately 12.7 GB of system RAM and two CPU cores; the specific GPU model and software-library versions used for the reported 3,335-image experiment are not documented in the experimental record. Precision, recall, mAP50, and mAP50-95 are the principal reported detection metrics, with validation values of 0.9973, 1.0000, 0.9950, and 0.7678, respectively, for the 667-image

validation split. The independent 337-image test set is defined by the same dataset protocol, but its final numerical metrics are not available in the experimental record.

Because the experimentally supported task is localization, Dice coefficient, mask IoU, Hausdorff distance, SDP measurement error, and clinical classification accuracy are reserved for later segmentation and measurement stages. The present paper therefore evaluates amniotic-fluid detection, not a complete automated AFV assessment.

IV. RESULTS AND DISCUSSION

A. Dataset Summary

TABLE I. FINAL IMAGE COUNTS FOR THE REPORTED EXPERIMENT

Subset	Images	Role
Training	2,331	Model fitting
Validation	667	Checkpoint selection / internal evaluation
Test	337	Independent test set; metrics not reported
Total public	3,335	Public model-development corpus

B. Detection Performance

The selected YOLOv8s model achieved a validation precision of 0.9973, recall of 1.0000, mAP@0.5 of 0.9950, and mAP@0.5:0.95 of 0.7678. Under the final dataset description,

these validation results correspond to the approximately 667-image validation subset of the 3,335-image public corpus, with the documented training split comprising approximately 2,331 images.

TABLE II. INTERNAL VALIDATION PERFORMANCE

Metric	Value	Metric	Value
Precision	0.9973	Recall	1.0000
mAP50	0.9950	mAP50-95	0.7678

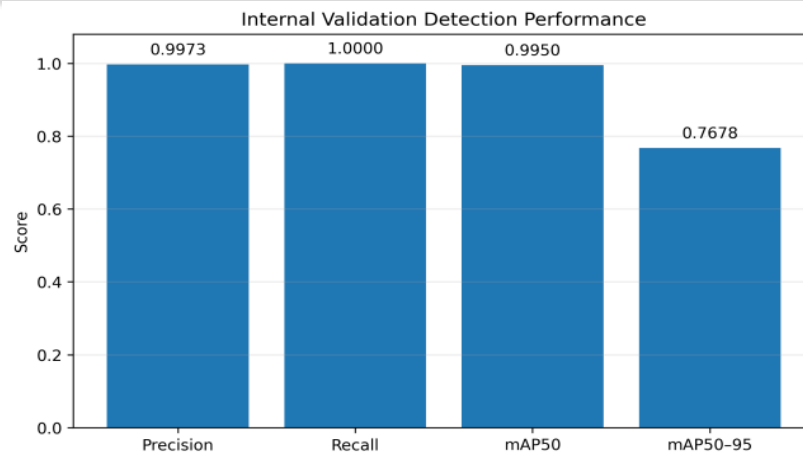


Fig. 2. Reported validation precision, recall, mAP50, and mAP50-95 for the YOLOv8s model on the approximately 667-image validation subset.

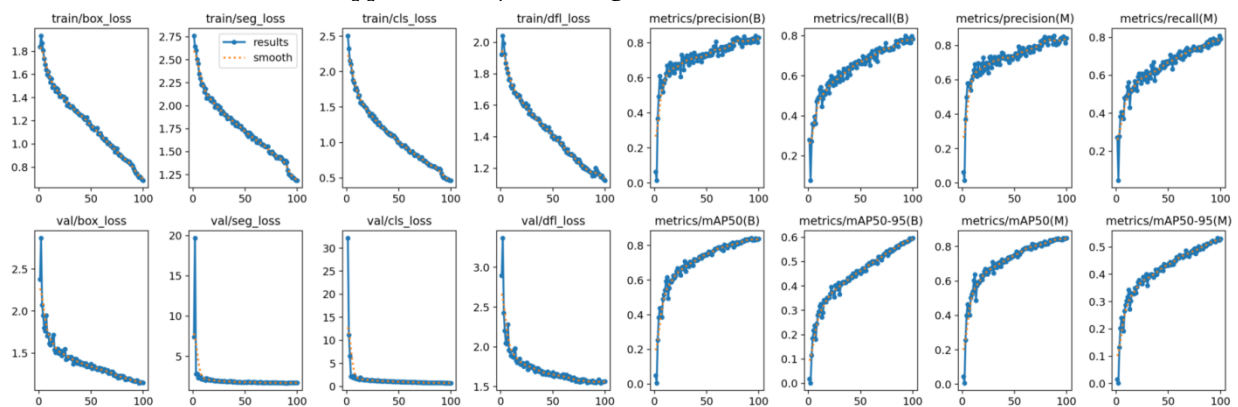


Fig. 3. Training and validation loss and metric curves (box, segmentation, classification, and DFL losses; precision, recall, mAP50, and mAP50-95) over 100 epochs.

C. Interpretation

The reported validation performance is strong, and is substantially more informative when read against the documented 667-image validation set rather than the earlier 29-image description. The high precision and recall indicate strong

localization performance on the internal validation distribution, although independent test results are required to establish generalization, and patient- or examination-level separation should be confirmed as specified in the dataset protocol.

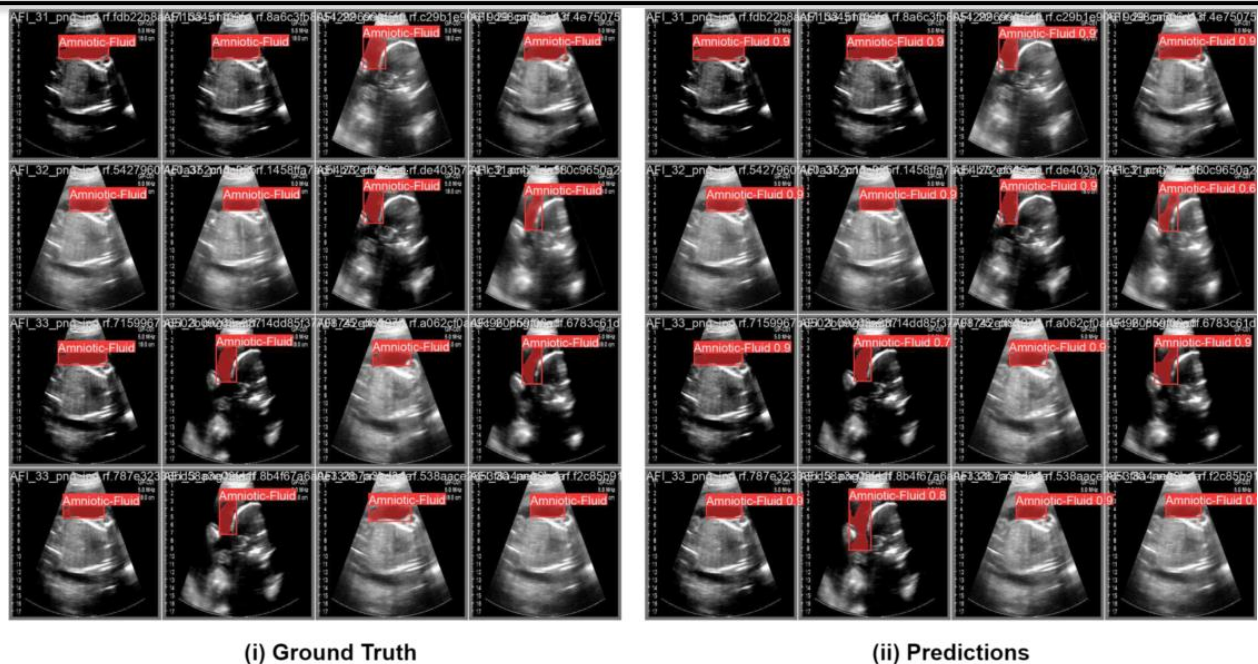


Fig. 4. Qualitative detection results: ground-truth bounding-box annotations (left columns) versus YOLOv8s predictions with confidence scores (right columns) on held-out ultrasound frames.

The gap between mAP50 (0.9950) and mAP50-95 (0.7678) is particularly informative here. Detection remains strong at IoU = 0.50, while stricter overlap requirements reduce performance, which points to less consistent bounding-box alignment. This matters for future SDP estimation, because physical measurement depends on precise pocket boundaries rather than approximate localization.

The independent test set contains 337 public images under the 70:20:10 split, but its final performance metrics are not available in the experimental record and should be obtained from the evaluation log before conference submission, rather than inferred from validation performance.

D. Discussion and Limitations

- This result is narrower in scope than end-to-end systems that combine segmentation, SDP measurement, and clinical classification [2]–[4], so a direct numerical comparison would not be appropriate without matched datasets and task definitions. The strength of the present study is methodological transparency: detection performance is reported separately from functions that were designed but not actually validated.

Dataset scale: the final public corpus contains 3,335 images, divided into approximately 2,331 training, 667 validation, and 337 test images.

- Patient-level provenance: the experimental record does not include a patient-level split audit that independently verifies the absence of leakage.

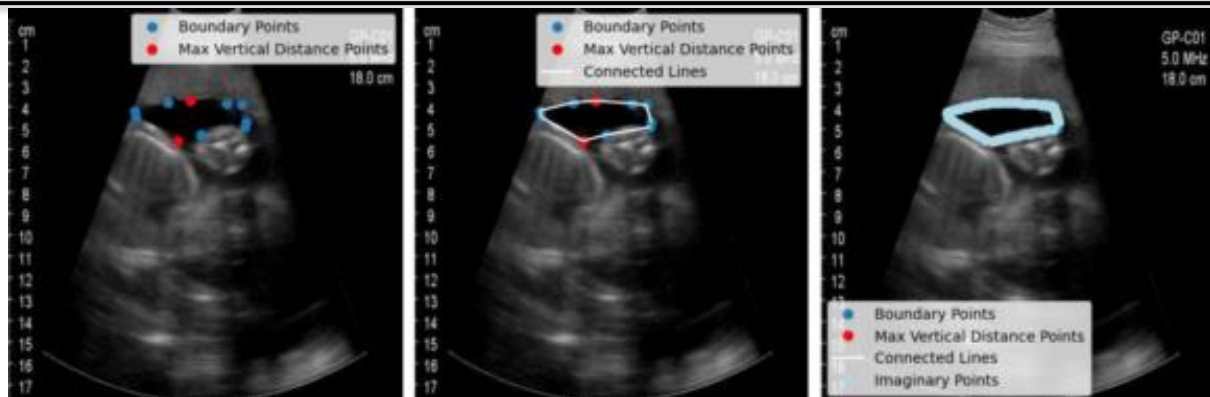


Fig. D1. Conceptual boundary-point and maximum-vertical-distance approach for single deepest pocket (SDP) measurement, proposed as a future extension beyond the detection stage reported here.

External validation: no independent multi-centre or cross-device test result is currently available.

- Task scope: the model produces bounding boxes, not pixel-level masks, and therefore does not itself establish SDP measurement or clinical classification.
- Uncertainty: confidence intervals or standard deviations were not supplied alongside the reported point estimates.

V CONCLUSION

Produce research established a YOLOv8s baseline for localizing visible amniotic-fluid pockets, using the full 3,335-image public dataset, split 2,331/667/337 for training, validation, and independent testing. This experiment achieved a validation precision of 0.9973, recall of 1.0000, mAP@0.5 of 0.9950, and mAP@0.5:0.95 of 0.7678.

These results support the feasibility of an automated detection stage for ultrasound assistance, but they do not demonstrate pixel-level segmentation, calibrated SDP measurement, or clinical classification. The next experimental stage should validate instance segmentation against expert masks, establish reliable physical calibration, compare automated SDP measurements against clinicians, and only then evaluate clinical categories on verified patient-level labels. Larger, patient-separated, multi-centre datasets and independent testing are also needed.

REFERENCES

- G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics: YOLOv8 real-time object detection and segmentation framework," Ultralytics Technical Documentation, 2023.
- M. K. Rafi et al., "Automated amniotic fluid volume assessment using YOLOv8 for enhanced fetal health diagnosis in ultrasound imaging," Research Square, 2025, doi: 10.21203/rs.3.rs-4878126/v1.
- O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234–241.
- K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proc. IEEE ICCV, 2017, pp. 2961–2969.
- H. Kim and C. Kim, "Deep learning-based segmentation and classification of amniotic fluid in ultrasound images," Computers in Biology and Medicine, 2021.
- I. Ullah Khan et al., "Deep learning-based computer-aided classification system for amniotic fluid assessment using ultrasound images," Biomedical Signal Processing and Control, 2024.
- S. Sun et al., "Automatic segmentation of amniotic fluid in ultrasound images using deep learning approaches," IEEE Access, 2020.

- A. Rattanasak et al., "An IoT-enabled wearable device for fetal movement detection using accelerometer and gyroscope sensors," *Sensors*, vol. 25, no. 5, p. 1552, 2025, doi: 10.3390/s25051552.
- D. Csillag et al., "AmnioML: Deep learning-based segmentation and volume estimation of amniotic fluid in fetal MRI," *Scientific Reports*, 2025.
- A. Costanzo et al., "Automated amniotic fluid segmentation from fetal magnetic resonance imaging using deep learning," *Medical Image Analysis*, 2022.
- A. F. Nabhan and Y. K. Abdelmoula, "Amniotic fluid index versus single deepest vertical pocket as a screening test for preventing adverse pregnancy outcome," *Cochrane Database of Systematic Reviews*, CD006593, 2024.
- P. F. Chamberlain et al., "Ultrasound evaluation of amniotic fluid volume: I. The relationship of marginal and decreased amniotic fluid volumes to perinatal outcome," *American Journal of Obstetrics and Gynecology*, 1984.
- E. F. Magann, S. P. Chauhan, W. Washington, N. S. Whitworth, J. N. Martin, and J. C. Morrison, "Ultrasound estimation of amniotic fluid volume using the largest vertical pocket containing umbilical cord: measure to or through the cord?" *Ultrasound in Obstetrics & Gynecology*, vol. 20, no. 5, pp. 464-467, 2002, doi: 10.1046/j.1469-0705.2002.00802.x.
- D. A. Khan, S. Ali, A. N. Khan et al., "Explainable Swin transformer with clinically constrained ECG-vital signs fusion for cardiovascular disease detection from real-world ECG images," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-67374-4.