SonoAudit: A Framework for Auditing Caliper Placement in Fetal Ultrasound Biometry

Vasanth Ambrose[1] Manasee Dnyaneshwar Dhore[2], Shyam Ayyasamy[2], Keerthi Ram[3], Manojkumar Lakshmanan[2], and Mohanasankar Sivaprakasam[1]

[1] Department of Electrical Engineering, Indian Institute of Technology, Madras, India
[2] Healthcare Technology Innovation Centre (HTIC), IITM, India
[3] Sudha Gopalakrishnan Brain Centre, IITM, India
vasanthambrose@htic.iitm.ac.in, mohan@ee.iitm.ac.in

Abstract. Accurate fetal biometry measurements from ultrasound examinations are critical for estimating clinical indicators like gestational age (GA), monitoring fetal growth, and assessing overall fetal well-being. Clinicians annotate the required measurements by placing calipers on the standard fetal planes. However, the reliability of these measurements depends heavily on the correct placement of calipers by sonographers. Even small deviations in caliper positioning can introduce clinically significant errors in biometric measurements and clinical indicators. While expert sonographers achieve high measurement consistency, their availability is limited, and less experienced operators are prone to annotation variability and placement errors. In this work, we present **SonoAudit**, an automated caliper auditing framework for fetal ultrasound biometry that evaluates the quality of clinician-placed calipers and provides objective feedback on measurement correctness. The proposed system localizes caliper annotations in fetal ultrasound images and segments the relevant anatomical structures, enabling assessment of caliper placement against clinically established measurement guidelines to determine whether the annotations satisfy standard measurement criteria. The proposed framework further estimates biometric measurements and corresponding clinical indicators, enabling quantitative assessment of annotation quality and its impact on fetal growth evaluation. Validation on patient studies demonstrated the feasibility of automated caliper auditing and its potential as a quality-control tool for fetal ultrasound examinations. The proposed framework achieved robust caliper detection and localization, with an average Precision of 0.88, a mean localization error of 3.05 ± 1.61 pixels, and 88.1% PCK@5, indicating accurate caliper placement within a 5-pixel tolerance of expert annotations. And attain an average MAE of 1.13 mm for biometric measurements.

Keywords: Fetal Ultrasound · Caliper Placement Analysis · Biometry

1 Introduction

In current clinical practice, sonographers manually place measurement calipers on ultrasound standard planes to obtain biometric measurements. Despite the existence of standardized measurement guidelines [10], the quality and accuracy of the final measurements depend heavily on the operator's expertise. Variability is particularly evident among less experienced operators, trainees, and in high-volume clinical environments where maintaining consistent measurement quality can be challenging [12, 11]. Even small deviations in caliper placement can introduce clinically significant measurement errors, potentially leading to inaccurate growth assessment. Furthermore, expert fetal sonographers capable of performing highly reliable measurements are limited in number, creating a need for systems that can assist in quality control and training.

Recent advances in artificial intelligence have significantly improved automated fetal ultrasound analysis, including standard plane detection [2, 7, 4] and anatomical structure segmentation [6, 1]. However, most existing methods focus on generating measurements directly from images rather than evaluating the correctness of measurements already performed by clinicians. As a result, there remains a critical gap in automated quality assurance systems that can objectively assess whether clinician-placed calipers satisfy established clinical measurement guidelines [3]. Automated auditing of biometric annotations presents several challenges due to the calipers being small visual markers, while the validity of a measurement depends not only on the caliper locations [13] but also on their geometric relationships and anatomical consistency with surrounding fetal structures and the auditing criteria differ across biometric measurements, requiring measurement-specific validation rules that closely reflect clinical practice.

To address these challenges, we propose **SonoAudit**, an automated caliper auditing framework for fetal ultrasound biometry. The proposed system employs a Yolo based detection model [8] to localize clinician-placed caliper arrowheads from fetal ultrasound images followed by a lightweight UNet based segmentation model [9] for anatomical structure segmentation and subsequently evaluates the quality of the measurements using clinically motivated auditing rules. In addition to auditing measurement quality, the proposed framework computes biometric measurements and corresponding clinical indicators. Validation on inhouse and online data demonstrates the feasibility of automated caliper auditing and highlights its potential as a quality assurance and sonographer training tool. Our main contributions are as follows:

- **SonoAudit:** A fully automated framework for auditing clinician-performed fetal biometric measurements from fetal ultrasound images.
- **Clinical Guidelines Checker:** A module that evaluates the geometric and anatomical correctness of caliper placement and biometric measurements from a sonographer's perspective.
- **Clinical Validation:** Extensive evaluation on in-house and publicly available datasets demonstrates a caliper detection precision of 0.88 and a biometric measurement MAE of 1.13 mm.

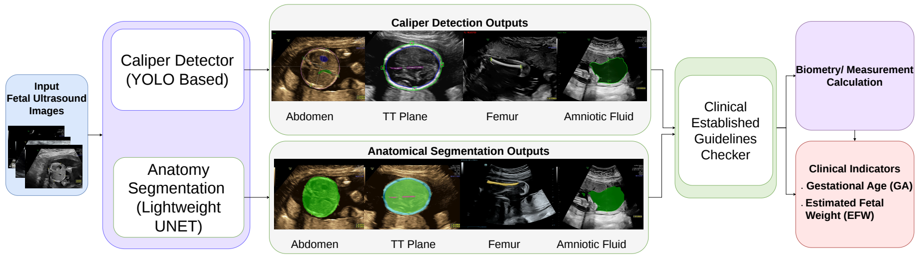

Fig. 1. SonoAudit Framework Overview

2 Methodology

2.1 Overview

Our proposed framework **SonoAudit** as shown in Fig. 1. performs caliper localization through the Yolo v8 variant detection model when a fetal ultrasound image containing clinician-placed biometric calipers is given as an input. Caliper localization is formulated as an object detection problem, where each caliper arrowhead is detected independently. As an initial quality-control step, the framework verifies whether the expected number of caliper points has been detected for a given biometric measurement. Cases with missing or excess caliper detections are automatically flagged as potential annotation errors and excluded from subsequent auditing stages. While caliper localization provides the geometric information required for measurement estimation, several clinical auditing criteria require anatomical context. Therefore, a lightweight UNET based segmentation network is employed to delineate anatomical structures relevant to each biometric measurement. The detected caliper locations and segmentation masks are subsequently analyzed using measurement-specific auditing algorithms designed to emulate established clinical measurement guidelines. Furthermore, the estimated biometric measurements are used to derive clinically relevant indicators, and they are subsequently mapped to gestational-age-specific Hadlock reference centile curves, enabling assessment of fetal growth relative to normative population standards.

2.2 Dataset

The proposed framework was trained and evaluated on an in-house dataset comprising of 1,000 and 400 patient studies used for training and testing respectively collected from multiple scanning centers. To further assess the generalizability of the proposed approach, additional evaluation was performed on 424 images from the publicly available UCL fetal ultrasound dataset [5]. The dataset contains examinations acquired using ultrasound systems from different manufacturers, resulting in substantial variability in image appearance, acquisition settings, and caliper annotation styles as shown in Fig.2. Such heterogeneity is particularly

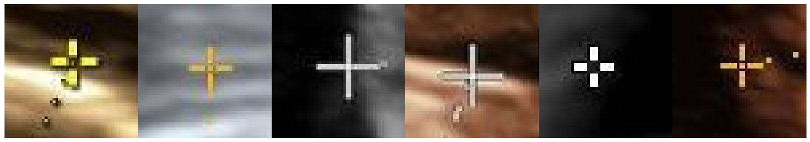

Fig. 2. Caliper Types

important for developing robust caliper auditing systems that generalize across imaging platforms.

2.3 Clinical Guideline Checker Module

Different biometric measurements require different anatomical landmarks for auditing. Consequently, a lightweight UNet based segmentation model is trained to identify the structures relevant to each measurement plane. The extracted contours from the predicted segmentation masks are then used to verify whether the detected caliper locations align with the corresponding anatomical boundaries, as defined by clinical measurement guidelines.

Trans-Thalamic Plane: For auditing Biparietal Diameter (BPD) and Occipitofrontal Diameter (OFD) measurements, the segmentation model predicts masks corresponding to the midline falx and inner and outer calvarial boundaries. These structures provide the anatomical reference required to verify whether the calipers are positioned correctly on the fetal skull according to standard measurement protocols. OFD boundary compliance requires the left and right caliper points to lie on the outer calvarial boundary, whereas BPD requires the top and bottom calipers to lie on the outer and inner calvarial boundaries, respectively. Furthermore, the BPD and OFD measurements must be orthogonal within 30° tolerance, with the BPD passing through the OFD midpoint and the OFD intersecting the midline falx.

Abdominal Plane: For abdominal measurements, the segmentation model predicts the fetal skin boundary, portal vein, and stomach bubble. These structures serve as anatomical landmarks for validating the placement of calipers used to obtain the Anteroposterior Abdominal Diameter (APAD) and Transverse Abdominal Diameter (TAD), ensuring consistency with established clinical guidelines. The TAD and APAD measurements are validated by assessing the angle between the two diameters computed from the abdominal circumference, requiring orthogonality within a tolerance of 30°.Furthermore, the APAD is required to pass through the portal vein, the TAD through the stomach bubble region, and the TAD length being greater than the APAD length.

Lower Limbs Plane: For femur length (FL) auditing, the femur is segmented and the detected caliper locations are verified against the extrema points of the

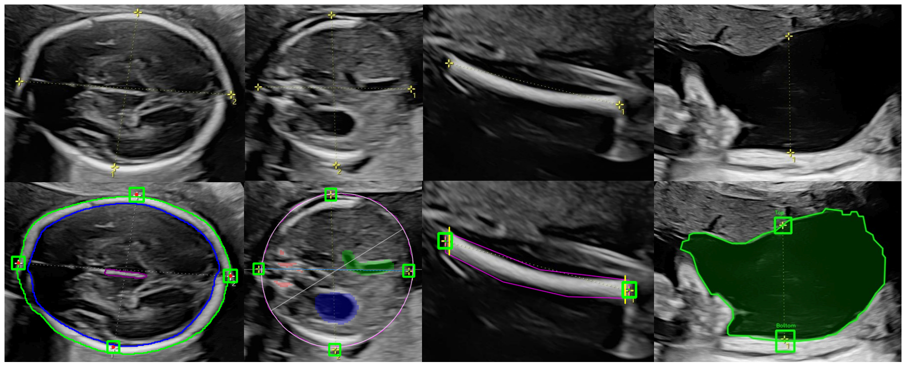

Fig. 3. Caliper Detection TT, AC, FL, SDVP (from left to right)

segmented femur mask. Additionally, because the femur exhibits strong acoustic boundaries in ultrasound images, local image gradients around each caliper location are analyzed to ensure alignment with the corresponding femoral bone boundaries. Measurements are considered valid when the calipers coincide with both the anatomical extrema of the segmented femur and the gradient-shift region.

Fetal Environment Plane: The framework is also extended to audit amniotic fluid measurements, where the Single Deepest Vertical Pocket (SDVP) is assessed. The contour extracted from the segmented amniotic fluid mask is verified to ensure that it intersects the detected caliper locations. Additionally, the verticality of the SDVP measurement is evaluated to confirm compliance with clinical measurement guidelines.

2.4 Measurement and Clinical Indicators Estimation

For standard-planes that satisfy the auditing criteria, the detected caliper locations are used to estimate the corresponding biometric measurements, which are subsequently employed to derive clinically relevant indicators, including Gestational Age (GA), Estimated Fetal Weight (EFW), and Cephalic Index (CI). In addition to the estimated composite GA, measurement-specific GA is also computed independently from HC, BPD, AC, and FL using established fetal growth equations. Furthermore, each biometric measurement is mapped to gestational-age-specific reference charts using Hadlock regression equations to determine its corresponding percentile, enabling assessment of fetal growth relative to normative population standards. By combining accurate caliper localization, anatomical segmentation, and clinically informed auditing rules as shown in Fig.3, the proposed framework provides an automated and interpretable quality-assurance mechanism for fetal ultrasound biometry.

3 Results and Discussion

The proposed framework was evaluated on 400 in-house and 424 publicly available datasets encompassing fetal head, abdominal, femur, and amniotic fluid measurements. For caliper detection, performance was assessed using Intersection over Union (IoU) and Precision. Caliper localization accuracy was evaluated using Mean Localization Error, Root Mean Square Error (RMSE), and Percentage of Correct Keypoints (PCK) at a tolerance threshold of 5 pixels. Biometric measurement accuracy was quantified using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Relative Error (MRE). Agreement between automated and reference measurements was assessed using the Intraclass Correlation Coefficient (ICC), Lin's Concordance Correlation Coefficient (CCC), and Pearson's correlation coefficient (r).

Detection and Localization of Calipers: Table 1 summarizes the caliper detection and localization results. The proposed detector achieves an average IoU of 0.70 and a precision of 0.88 across all anatomical planes. The mean localization error was 3.05±1.61 pixels, with an overall RMSE of 3.45 pixels. These results demonstrate robust detection and localization of caliper arrowheads despite variations in ultrasound image quality, anatomical appearance, and device-specific annotation styles. The highest localization accuracy was observed for the lower limbs plane, achieving a precision of 0.93 and a PCK@5 of 93.1%, while fetal environment measurements exhibited comparatively lower performance due to increased variability in caliper placement and structure appearance.

Table 1. Caliper detection and localization performance evaluation.

Measurement	**IoU**	**Precision**	**Mean(px)**	**RMSE (px)**	**PCK@5 (%)**
TransThalamic Plane	0.715	0.912	2.87 ± 1.52	3.25	91.2
Abdominal Plane	0.691	0.879	3.10 ± 1.57	3.48	87.9
Lower Limbs Plane	0.735	0.931	2.59 ± 1.30	2.90	93.1
Fetal Environment	0.660	0.799	3.62 ± 2.03	4.15	80.0
Average	**0.700**	**0.880**	$\mathbf{3.05 \pm 1.61}$	**3.45**	**88.1**

Biometry Measurement: Table 2 presents the biometric measurement and clinical indicator evaluation results. Across all biometric measurements, the proposed framework achieved an average MAE of 1.13 mm, RMSE of 1.48 mm, and MRE of 2.7%, indicating high measurement accuracy. The low standard deviation of the MRE across all measurements indicates consistent performance of the proposed framework and demonstrates its robustness across different anatomical structures, imaging conditions, and acquisition settings. Among the evaluated measurements, biparietal diameter (BPD) and occipitofrontal diameter (OFD)

Table 2. Biometry Measurement Performance Evaluation

Measurement	**MAE**	**RMSE**	**MRE**	**ICC**	**CCC**	**Pearson's(r)**
BPD	1.11 ± 1.32	1.70 ± 1.32	0.021 ± 0.022	0.972	0.971	0.982
OFD	1.11 ± 1.56	1.89 ± 1.56	0.016 ± 0.020	0.971	0.970	0.973
AC	1.77 ± 3.46	1.48 ± 3.46	0.050 ± 0.027	0.933	0.933	0.989
FL	0.74 ± 0.80	1.08 ± 0.80	0.022 ± 0.026	0.965	0.964	0.966
SDVP	0.92 ± 0.95	1.25 ± 0.95	0.028 ± 0.031	0.958	0.957	0.961
Average	$\mathbf{1.13 \pm 1.62}$	$\mathbf{1.48 \pm 1.62}$	$\mathbf{0.027 \pm 0.025}$	**0.960**	**0.959**	**0.974**
Clinical Indicators						
GA	3.22 ± 2.33	3.95 ± 2.33	2.233 ± 1.625	0.917	0.914	0.950
CI	1.61 ± 1.56	2.22 ± 1.56	2.152 ± 2.065	0.792	0.784	0.822

demonstrated the highest agreement, with ICC values exceeding 0.97. Gestational age (GA) estimation achieved an MAE of 3.22 days and an ICC of 0.917, while the clinical index (CI) achieved an MAE of 1.61 and an ICC of 0.792. These results suggest that the proposed automated framework can provide clinically reliable biometric measurements and derived fetal growth indicators.

Error Identification Cases: Fig. 4 shows representative examples of measurement annotations flagged by the proposed framework. In the BPD example (leftmost), the topmost caliper is not placed on the outer edge of the calvarium. In the AC example, the measurement axis fails the orthogonality constraint and the topmost caliper is incorrectly positioned. In the FL example, the caliper endpoints do not align with the femoral bone-edge gradient transitions, indicating inaccurate endpoint placement. In the amniotic fluid (SDVP) example, the inferior caliper is positioned above the fluid boundary instead of on the boundary itself. These cases highlight the capability of the proposed framework to automatically detect diverse caliper placement errors that could lead to inaccurate biometric measurements.

Measurement Centiles: To evaluate the clinical relevance of the estimated biometric measurements, each measurement was mapped to Hadlock Equation based GA-specific centile curves (Fig. 5). The majority of BPD, HC, AC, and FL measurements were located near the 50th percentile and predominantly within

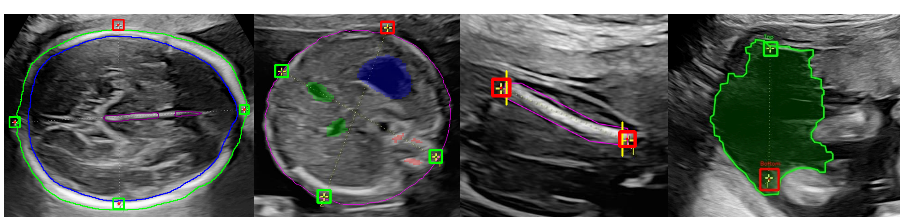

Fig. 4. Error Identification Cases TT, AC, FL, SDVP (from left to right)

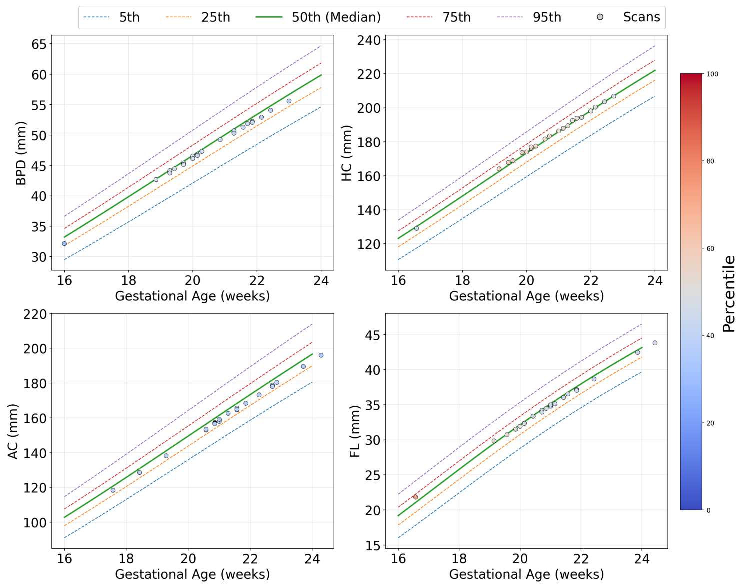

Fig. 5. GA specific centile curve plot

the 25th–75th percentile range. Additionally, all measurements remained within the 5th–95th percentile reference limits, suggesting that the audited measurements are consistent with expected fetal growth patterns. These centile plots provide an intuitive visualization of fetal growth status, enabling identification of measurements that may deviate from normal growth trajectories.

4 Conclusion

In this work, we presented SonoAudit, an automated framework for auditing fetal ultrasound caliper annotations. By detecting calipers and checking them against clinical guidelines, the method assesses annotation quality and estimates biometric measurements for fetal growth evaluation. Validation on in-house and public datasets showed good performance, with 0.88 Precision, 88.1% PCK@5, and 1.13 mm MAE for biometric measurements demonstrating its potential as a quality-assurance and training tool. Future work will extend validation to larger multi-center datasets, extending to additional biometric measurements and integrating real-time feedback mechanisms to assist sonographers.

Prospects of Application: **SonoAudit** can be deployed as a quality control and training support tool in clinical fetal ultrasound systems. By automatically auditing biometric annotations and providing objective feedback, it has the potential to improve measurement consistency, reduce operator-dependent variability, and assist both experienced and trainee sonographers in routine fetal growth assessment.

References

1. Avisdris, N., Joskowicz, L., Dromey, B., David, A.L., Peebles, D.M., Stoyanov, D., Ben Bashat, D., Bano, S.: Biometrynet: landmark-based fetal biometry estimation from standard ultrasound planes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 279–289. Springer (2022)
2. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE transactions on medical imaging **36**(11), 2204–2215 (2017)
3. Bilardo, C., Chaoui, R., Hyett, J., Kagan, K., Karim, J., Papageorghiou, A., Poon, L., Salomon, L., Syngelaki, A., Nicolaides, K.: Isuog practice guidelines (updated): performance of 11–14-week ultrasound scan. Ultrasound in Obstetrics and Gynecology **61**(1) (2023)
4. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020)
5. Di Vece, C., Mao, Z., Avisdris, N., Dromey, B., Napolitano, R., Ben Bashat, D., Vasconcelos, F., Stoyanov, D., Joskowicz, L., Bano, S.: A multicentre benchmark dataset for comprehensive landmark-based fetal ultrasound biometry. Scientific Reports (2026)
6. Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S.: A review on deep-learning algorithms for fetal ultrasound-image analysis. Medical image analysis **83**, 102629 (2023)
7. Maani, F., Saeed, N., Saleem, T., Farooq, Z., Alasmawi, H., Diehl, W., Mohammad, A., Waring, G., Valappi, S., Bricker, L., et al.: Fetalclip: A visual-language foundation model for fetal ultrasound image analysis. arXiv preprint arXiv:2502.14807 (2025)
8. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
10. Salomon, L., Alfirevic, Z., Da Silva Costa, F., Deter, R., Figueras, F., Ghi, T.a., Glanc, P., Khalil, A., Lee, W., Napolitano, R., et al.: Isuog practice guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound in obstetrics & gynecology **53**(6), 715–723 (2019)
11. Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D., Papageorghiou, A., Fetal, I., for the 21st Century (INTERGROWTH-21st), N.G.C.: Intra-and interobserver variability in fetal ultrasound measurements. Ultrasound in obstetrics & gynecology **39**(3), 266–273 (2012)
12. Sarris, I., Ioannou, C., Dighe, M., Mitidieri, A., Oberto, M., Qingqing, W., Shah, J., Sohoni, S., Al Zidjali, W., Hoch, L., et al.: Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. Ultrasound in obstetrics & gynecology **38**(6), 681–687 (2011)
13. Venturini, L., Budd, S., Farruggia, A., Wright, R., Matthew, J., Day, T.G., Kainz, B., Razavi, R., Hajnal, J.V.: Whole examination ai estimation of fetal biometrics from 20-week ultrasound scans. NPJ Digital Medicine **8**(1), 22 (2025)