**Appendix 8: Validation methods**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset Name** | **Validation Method** | **Number of Experts Involved** | **Metrics Used for Validation** |
| FPUS23: An Ultrasound Fetus Phantom Dataset with Deep Neural Network Evaluations for Fetus Orientations, Fetal Planes, and Anatomical Features | Models evaluated using accuracy, F1-score, mAP, and mAR metrics | Not specified | Accuracy, F1-score, mean Average Precision (mAP), mean Average Recall (mAR) |
| Large-scale annotation dataset for fetal head biometry in ultrasound images | Two-step validation process involving medical experts | Senior Attending Physician and a Radiologic Technologist | Intraclass Correlation Coefficients (ICC), Jaccard similarity indices (JS) |
| The JNU-IFM dataset for segmenting pubic symphysis-fetal head | Labels validated by two experienced radiologists | Two experienced radiologists | Not mentioned |
| The construction and application of an ultrasound and anatomical cross-sectional database of structural malformations of the fetal heart | Confirmation of congenital heart defects through postnatal echocardiography, surgery, or pathological specimens | Not specified | Not specified |
| PSFHS: Intrapartum ultrasound image dataset for AI-based segmentation of pubic symphysis and fetal head | Double annotation, reviewed by expert physicians | Expert physicians | Dice coefficient for consistency |
| How much can AI see in early pregnancy: A multi-center study of fetus head characterization in week 10–14 in ultrasound using deep learning | Internal and external testing, comparison with classic deep learning models and human experts | Not explicitly mentioned | Sensitivity, specificity, area under the curve (AUC), accuracy, precision, F1-Score |
| Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries | Models evaluated using AUC, accuracy, recall, and precision metrics | Not specified | Area Under the Curve (AUC), accuracy, recall, precision |
| Automated measurement of fetal head circumference using 2D ultrasound images | Comparison with manual annotations by experienced sonographer and medical researcher, statistical analysis | Experienced sonographer and medical researcher | Difference (DF), absolute difference (ADF), Hausdorff distance (HD), Dice similarity coefficient (DSC), mean difference in GA |
| Fetal Abdominal Structures Segmentation Dataset Using Ultrasonic Images | Rigorous quality control process overseen by two team members | Two team members | Quality control criteria ensuring exclusion of images with calipers or acoustic shadows |
| Automated annotation and quantitative description of ultrasound videos of the fetal heart | Leave-one-out cross-validation | Not specified | Classification error rate, orientation error, phase error |
| Real-time diameter of the fetal aorta from ultrasound | Comparison with state-of-the-art methods, synthetic data validation | Not specified | Mean squared error (MSE), relative error (RE) |
| Values and validity of fetal parameters by ultrasound and Doppler as markers of fetal lung maturity | Comparison with neonatal outcomes (presence of RDS) | Not specified | Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy |
| Generative Diffusion Model Bootstraps Zero-shot Classification of Fetal Ultrasound Images In Underrepresented African Populations | Comparison with baseline methods, ablation studies | Not specified | Accuracy, Recall, Precision, F-score, Area Under the Curve (AUC) |
| Automatic detection of complete and measurable cardiac cycles in antenatal pulsed-wave Doppler signals | Leave-one-subject-out cross-validation | Not specified | Accuracy of detection |