Scientific Reports

https://doi.org/10.1038/s41598-026-41635-8

Article in Press

A multi-task deep learning and radiomics framework for fetal anatomical structure detection and classification in ultrasound imaging

Received: 15 December 2025

Accepted: 22 February 2026

Published online: 02 March 2026

Cite this article as: Zhou X., Wan J., Sun F. *et al.* A multi-task deep learning and radiomics framework for fetal anatomical structure detection and classification in ultrasound imaging. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-41635-8

Xuan Zhou, Jie Wan, Fengjie Sun, Ruxin Wang, Yafei Yan, Pin Li & Cuihua Wang

A multi-task deep learning and radiomics framework for fetal anatomical structure detection and classification in ultrasound imaging

Xuan Zhou[1#], Jie Wan[1#], Fengjie Sun[1], Ruxin Wang[2], Yafei Yan[1], Pin Li[1*], Cuihua Wang[1*]

Xuan Zhou and Jie Wan contributed equally to this article and should be considered co-first authors.

[1] Department of Ultrasound Diagnosis, Affiliated Hospital of Hebei University, Baoding City, Hebei Province, 071000, China

[2] Department of Radiology, Affiliated Hospital of Hebei University, Baoding City, Hebei Province, 071000, China

[#] Xuan Zhou and Jie Wan contributed equally to this article and should be considered co-first authors.

[*]**Corresponding Authors**

Pin Li
Department of Ultrasound Diagnosis, Affiliated Hospital of Hebei University, Baoding City, Hebei Province, 071000, China
Email: wch18233277919@sina.com

Cuihua Wang
Department of Ultrasound Diagnosis, Affiliated Hospital of Hebei University, Baoding City, Hebei Province, 071000, China
Email: 17633127586@sina.com

A multi-task deep learning and radiomics framework for fetal anatomical structure detection and classification in ultrasound imaging

Abstract

Objective: To develop a comprehensive deep learning and radiomics-based multi-task pipeline for the detection and classification of key fetal anatomical structures in first-trimester ultrasound images, using a diverse multi-center dataset to ensure high variability, reproducibility, and generalizability.

Materials and Methods: A total of 4,532 fetal ultrasound scans (gestational age 11–14 weeks), retrospectively collected from nine medical centers, were included in this study. Two detection models, You Only Look Once version 11 (YOLOv11) and Shifted Window Transformer (Swin Transformer), were trained to localize nine fetal brain and craniofacial structures. From each detected region, 215 radiomic features and 1,792 deep features were extracted. Feature stability was ensured through intra-class correlation coefficient (ICC) filtering (threshold ≥ 0.75), Pearson correlation analysis, and Least Absolute Shrinkage and Selection Operator (LASSO) regression. The selected features were then used to train a Transformer-based model for tabular data (TabTransformer) to classify fetal anatomical structures into clinically defined categories based on their sonographic appearance. Model performance was evaluated using Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), and Sensitivity across training, internal validation, and external test datasets.

Results: Fusion models integrating radiomic and deep features consistently outperformed single-modality models in both detection and classification. On the external test set, classification accuracy reached 96.1%, with AUCs up to 96.89%, and sensitivity exceeding 95% for key anatomical structures. Swin Transformer achieved superior localization performance compared to YOLOv11, with Intersection over Union (IoU) values up to 0.97 and F1-scores ≥ 0.94. Feature reproducibility remained above 75% across centers. The TabTransformer classifier demonstrated strong generalization and robustness, effectively leveraging the fused feature space for high-precision classification.

Conclusion: This study presents the fully integrated, multi-task framework for fetal anatomical structure detection and classification using multi-center ultrasound data. The proposed approach demonstrates high reproducibility and diagnostic performance, offering strong clinical potential for early and objective fetal anomaly screening in the first trimester.

Keywords: deep learning, fetal anatomy, first trimester, machine learning, radiomics, ultrasound imaging, visual transformer

1.Introduction

Ultrasound imaging allows detailed visualization of fetal anatomy throughout gestation, particularly the developing central nervous system [1,2]. Evaluation of key structures such as the thalami, midbrain, cisterna magna, nuchal translucency, and facial features, including the nasal bone and palate, is essential for early detection of congenital and genetic anomalies [3,4]. However, diagnostic accuracy in fetal neurosonography depends heavily on operator expertise and image quality, resulting in considerable inter- and intra-observer variability [5,6]. The subjective nature of interpretation further complicates differentiation between normal and abnormal development, especially across centers with varying imaging protocols.

In addition to the difficulties in visual identification, the classification of detected fetal structures into clinically meaningful categories, such as determining whether a structure is normal, hypoplastic, or absent, requires nuanced anatomical knowledge and consistent interpretation criteria [7,8]. Such classification tasks are critical for guiding clinical decision-making but remain underexplored in terms of automated, objective systems. These

limitations underscore the need for robust, scalable, and interpretable computational approaches that can assist clinicians in both detection and classification tasks with minimal operator dependency [9–11]. Furthermore, existing classification efforts often lack generalizability across multi-center datasets due to differences in image quality, acquisition protocols, and fetal positioning [12–14]. Manual labeling of such data is labor-intensive and time-consuming, limiting the scalability of traditional methods [15–17]. Current solutions rarely incorporate per-structure classification after detection, leaving a gap in fine-grained, structure-specific diagnostic support [18–21]. Bridging this gap requires end-to-end frameworks capable of handling both spatial localization and semantic interpretation of fetal structures across varied imaging conditions [22,23].

Recent advances in deep learning have introduced powerful solutions for automated medical image analysis. In particular, detection models such as You Only Look Once version 11 (YOLOv11) [24,25] and the Shifted Window Transformer (Swin Transformer) have demonstrated strong performance in localizing anatomical structures within complex imaging modalities [26,27]. These models offer significant advantages in terms of speed, precision, and their ability to generalize across heterogeneous data sources. Complementing these are radiomics [28,29] and deep learning–based feature extraction techniques [30], which have emerged as prominent tools for characterizing image regions based on quantitative descriptors. Radiomics enables the extraction of high-dimensional handcrafted features that capture

texture, shape, and intensity, while models like the Vision Transformer (ViT) can generate deep representations that encode complex image patterns in a data-driven manner [31,32].

The integration of advanced feature extraction and machine learning classifiers has shown strong performance in medical imaging [33–36]. When combined with feature selection methods such as intra-class correlation (ICC) [37], correlation filtering, and LASSO regression [38], these approaches enhance feature stability and generalizability. Models like TabTransformer further improve adaptability and interpretability for structured image-derived data.

In this study, we propose a unified deep learning and radiomics-based framework for fetal brain structure assessment in first-trimester ultrasound. The pipeline integrates detection, feature extraction, and classification using multi-center data to ensure reproducibility and robustness.

Its main contributions include: (1) employing Swin Transformer and YOLOv11 for multi-structure detection; (2) extracting and refining radiomic and deep features through ICC, correlation analysis, and LASSO; (3) combining both feature types via hybrid fusion for improved classification; and (4) implementing an end-to-end ViT for fully automated structure classification. This integrated approach enhances objective, clinically relevant decision support in prenatal ultrasound imaging.

2.Materials and methods

2.1. Data collection

Ultrasound data for this study were retrospectively collected from nine independent medical imaging centers to construct a heterogeneous and generalizable dataset for the detection and classification of fetal brain and craniofacial structures. These centers included tertiary care hospitals and specialized prenatal imaging facilities, each following institutional clinical protocols and adhering to ethical standards for data privacy and patient confidentiality. The initial dataset consisted of 12,569 ultrasound examinations performed on pregnant patients over a three-year period. All imaging data were de-identified prior to analysis to ensure compliance with data protection regulations. Figure 1 presents the complete architecture of the proposed multi-faceted analysis pipeline, integrating deep learning and radiomics methodologies for the detection and classification of fetal brain and craniofacial structures.

This retrospective multi-center study was conducted in accordance with the Declaration of Helsinki and relevant national regulations on human subject research. Ethical approval was reviewed and formally waived by the Institutional Review Board of the Affiliated Hospital of Hebei University (Baoding, China) because the study involved retrospective analysis of fully anonymized ultrasound data, with no direct patient contact and no use of identifiable personal information. Due to the retrospective nature of the study

and complete data anonymization, the requirement for written informed consent was waived by the same Institutional Review Board, in accordance with institutional policies and national ethical guidelines governing secondary use of medical data.

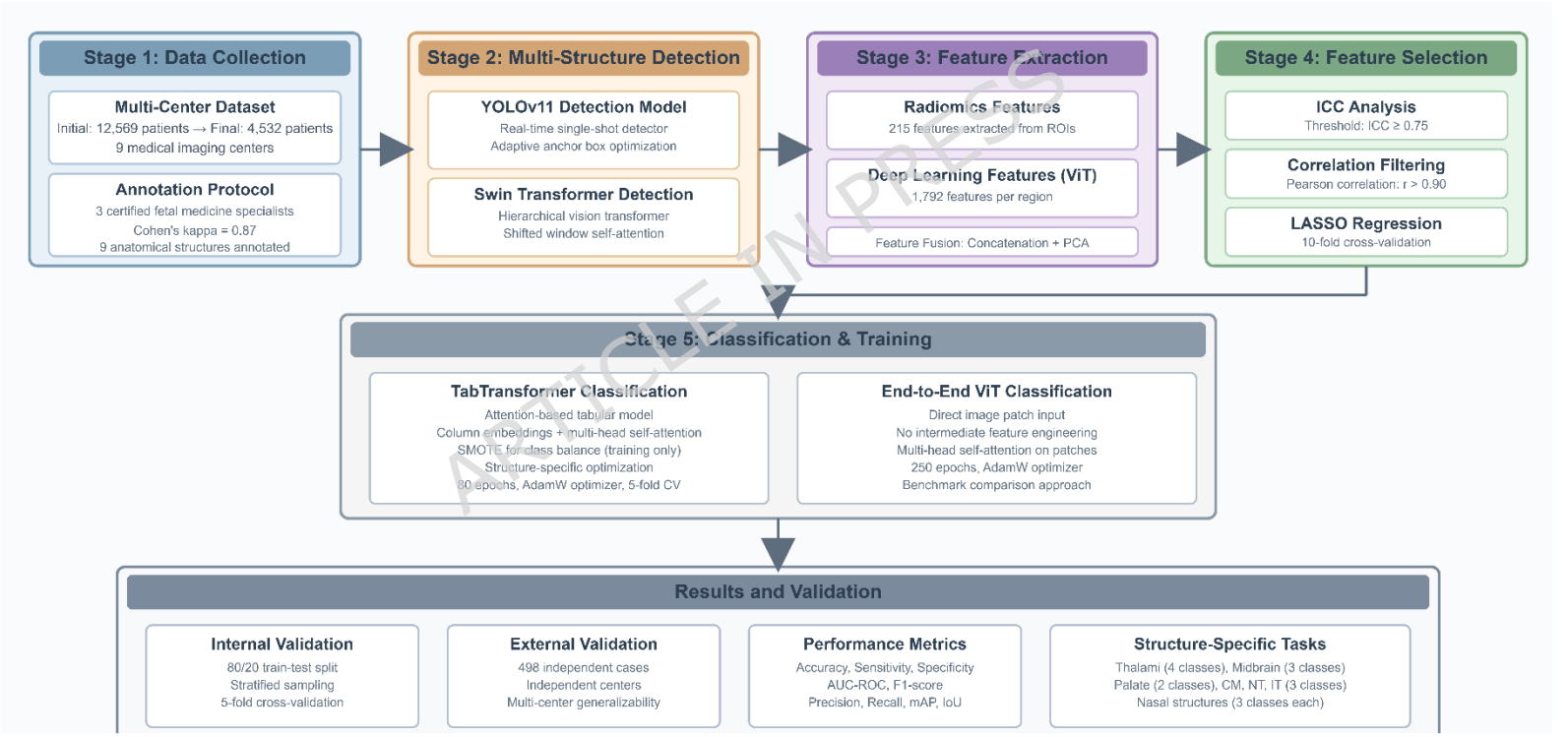

Figure 1. Overview of the multi-faceted deep learning and radiomics pipeline for fetal brain structure analysis in first-trimester ultrasound imaging

Multi-center ultrasound dataset

Each participating center employed high-resolution ultrasound equipment from various manufacturers, resulting in diversity in imaging characteristics such as resolution, bit depth, and frame rate. This variability contributes to the robustness of the proposed deep learning models by ensuring their exposure to a wide range of imaging conditions. The detailed specifications of the ultrasound systems used across the nine participating centers have been provided in Supplementary Table S1.

Inclusion and exclusion criteria

Figure 2 summarizes the multi-step inclusion and exclusion process used to curate the final dataset of 4,532 first-trimester ultrasound exams. Cases were removed for incomplete imaging, poor image quality, out-of-range gestational age, or missing clinical metadata, ensuring a high-quality and clinically consistent dataset for subsequent detection and classification tasks. The detailed clinical characteristics of the final study cohort are provided in Supplementary Table S2.

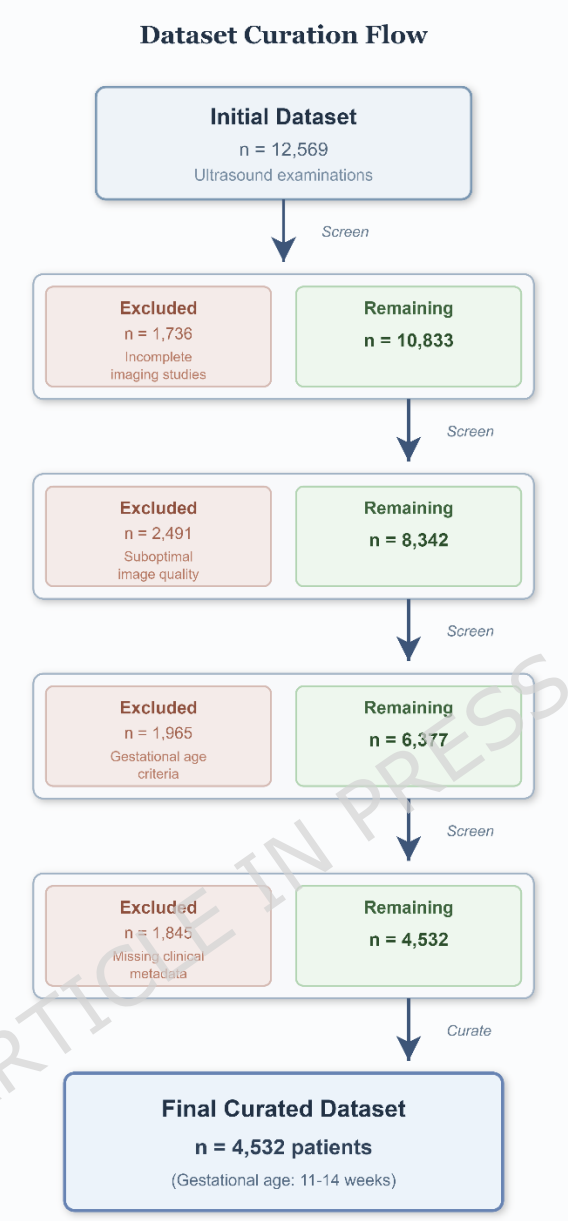

Figure 2. Inclusion and exclusion criteria for ultrasound dataset curation from multi-center retrospective study

2.2. Annotation protocol

Annotations of fetal structures were conducted by a panel of three certified fetal medicine specialists, each with over a decade of experience in prenatal

ultrasound and fetal neurosonography. Each fetal medicine specialist performed independent annotations while being blinded to the assessments of the other reviewers to prevent inter-observer influence. Following the independent annotation phase, inter-rater agreement was calculated using Cohen's kappa coefficient ($\kappa = 0.87$), reflecting substantial consistency among the three annotators. Any discrepancies were subsequently resolved through a consensus discussion to generate the final reference labels used for model training and validation. Nine anatomical structures were manually annotated using bounding boxes in standard ultrasound planes, with consensus labels established after independent review of 300 cases. Discrepancies were resolved by discussion, and the final annotations showed substantial inter-observer agreement ($\kappa = 0.87$).

Structure-specific classification tasks

Each of the nine annotated fetal structures was categorized into clinically meaningful classes based on established sonographic criteria, reflecting normal development and pathological variants observed between 11 and 14 weeks of gestation. The classification schema accounted for both morphological and biometric indicators, such as structural symmetry, echogenic contours, and threshold-based measurements (e.g., for NT, CM, and nasal bone). Table 1 summarizes the class definitions, diagnostic criteria, and case distribution for each structure, serving as the foundation for the subsequent multi-class classification tasks.

Table 1. Classification schema and distribution of annotated fetal anatomical structures (n = 4,532) used for multi-class labeling and model training across nine defined regions

Structure	Classes (n)	Class Labels	Definition / Criteria	Number of Cases
Thalami	4	Immature, Moderately mature, Mature, Abnormal	Echogenic differentiation and medial separation; dysmorphology	480 / 1020 / 2800 / 232
Midbrain	3	Normal, Asymmetrical, Abnormal	Bilateral symmetry, contour integrity, and midline morphology	3600 / 512 / 420
Palate	2	Normal, Cleft present	Continuity of echogenic palatal line	4375 / 157
Cisterna Magna (CM)	3	Hypoplastic, Normal, Enlarged	Size-based: <2 mm (hypoplastic), 2–10 mm (normal), >10 mm (enlarged)	370 / 3940 / 222
Nuchal Translucency (NT)	3	Normal, Borderline, Increased	≤2.5 mm (normal), 2.6–3.0 mm (borderline), >3.0 mm (increased)	3900 / 400 / 232

Nasal Tip	3	Normal, Hypoplastic, Absent	Shape and echogenic prominence on mid-sagittal view	3820 / 518 / 194
Nasal Skin	3	Normal, Hypoplastic, Absent	Skin thickness and continuity over nasal bridge	3860 / 500 / 172
Nasal Bone	3	Normal, Hypoplastic, Absent	Bone length: ≥2.5 mm (normal), 1.5–2.4 mm (hypoplastic), not visible (absent)	3880 / 464 / 188
Intracranial Translucency (IT)	3	Normal, Narrowed, Absent	Anechoic space behind brainstem: normal (>1 mm), narrowed (<1 mm), or absent	3760 / 500 / 272

2.3. Detection of fetal structures

You only look once version 11 model for detection

Automated detection was performed using YOLOv11, a fast single-shot detector optimized for small anatomical targets. Trained on annotated bounding boxes for nine fetal structures, the model leveraged multi-scale features and adaptive anchors to achieve accurate localization. Training used IoU-based box regression and cross-entropy classification losses, resulting in precise spatial delineation for downstream classification.

Shifted Window Transformer-based detection

The Swin Transformer is a hierarchical vision transformer that processes input images as fixed-size non-overlapping patches and applies shifted window-based self-attention within and across local regions [39–42]. This design enables multi-scale representation learning while preserving spatial continuity, providing superior performance in capturing both global and local contextual information. Within our pipeline, the Swin Transformer functions as a primary detection model, generating precise bounding boxes for fetal structures by integrating contextual cues across different feature resolutions.

The selection of YOLOv11 and Swin Transformer was guided by their recent advancements and suitability for medical imaging tasks involving small, variable structures. YOLOv11 introduces adaptive anchor box optimization, enhanced feature pyramids, and cross-stage partial connections, offering superior detection accuracy and computational efficiency compared to earlier architectures such as YOLOv8 or EfficientDet [24,43,44]. The Swin Transformer was chosen for its hierarchical self-attention mechanism using shifted windows, which captures both local and global contextual features critical for delineating subtle fetal structures.

2.4. Evaluation metrics for detection performance

Performance metrics were selected according to established standards in medical image analysis [45]. For detection tasks, precision, recall

(sensitivity), and F1-score quantify the balance between true positive and false positive detections, while Intersection over Union (IoU) measures the overlap between predicted and ground-truth bounding boxes. The mean Average Precision (mAP) summarizes detection performance across multiple IoU thresholds (0.5–0.95). These metrics enabled quantitative comparison between YOLOv11 and Swin Transformer performance, training was conducted for 250 epochs.

2.5. Feature extraction

Radiomics and deep feature extraction

Radiomic features (215 IBSI-compliant descriptors) were extracted from ROIs generated by the detection models, capturing first-order intensity properties and higher-order texture patterns. Complementing these handcrafted metrics, the ViT model produced 1,792 deep features per region by encoding multi-scale spatial and contextual information through patch embeddings and multi-head self-attention. Together, these feature sets provided both quantitative texture descriptors and rich data-driven representations of fetal anatomy.

Feature fusion strategy (Radiomics + Deep)

To leverage the complementary strengths of handcrafted and learned features, a feature fusion strategy was implemented. Feature fusion was implemented through direct feature-level concatenation of normalized

radiomic and deep feature vectors, forming a single hybrid representation for each annotated anatomical region. Prior to concatenation, both feature sets were z-score normalized to ensure comparable scaling and projected into harmonized subspaces through Principal Component Analysis (PCA), preventing dominance of one feature domain due to dimensional imbalance. The resulting fused vector captures complementary information: radiomic features provide handcrafted descriptors quantifying textural and morphological properties, while deep features encode higher-order contextual and spatial patterns derived from hierarchical transformer representations. This combined embedding thus integrates low-level handcrafted descriptors with semantically rich, data-driven deep features, yielding a comprehensive and discriminative representation of each fetal structure.

$$r \in R^p,\ d \in R^q$$

Mathematically, if $r \in R^p$ and $d \in R^q$represent normalized radiomic and deep feature vectors, respectively, the fused representation is expressed as

$$f = [r] \mid [d],$$

where [.]| [.]denotes concatenation along the feature dimension. The fused representation f was subsequently processed through the same reproducibility and dimensional optimization pipeline as individual feature sets, including ICC filtering (threshold ≥ 0.75), correlation reduction (r >

0.90), and LASSO regression for sparse feature selection. This ensured that the final hybrid feature space retained only robust, non-redundant, and maximally informative features for classification by the TabTransformer model.

Radiomics features and deep ViT features were concatenated at the feature level for each annotated region. Prior to fusion, both feature sets were normalized, and Principal Component Analysis (PCA) was applied to align the dimensionality of radiomic and deep representations, ensuring that neither feature domain dominated the combined vector due to scale or dimensional imbalance. The fused feature set was then subjected to the same preprocessing pipeline as the individual sets, including ICC filtering, correlation reduction, and LASSO-based feature selection. This hybrid representation provided a balanced and comprehensive characterization of fetal structure phenotypes for subsequent classification.

2.6. Intraclass correlation coefficient analysis, Correlation, and Least Absolute Shrinkage and Selection Operator regression

Feature refinement involved three steps: (1) ICC filtering using a two-way random-effects model, retaining only features with ICC ≥ 0.75 to ensure inter-center and inter-observer reproducibility; (2) Pearson correlation filtering ($r > 0.90$) to remove redundant features; and (3) LASSO regression with 10-fold cross-validation to select the most discriminative features for

each structure. This pipeline produced a compact, stable, and highly informative feature set for downstream classification.

2.7. Classification of detected structures

Structure-specific classification tasks

Each fetal structure detected in the previous stage was subjected to an individual classification task based on its clinically relevant categories. The annotated structures, thalami, midbrain, palate, CM, NT, nasal tip, nasal skin, nasal bone, and IT, were each classified into two to four categories, depending on structural complexity and diagnostic criteria. The class definitions were established based on current clinical standards and sonographic thresholds, as previously detailed in Table 1. For each structure, classification tasks were modeled independently to capture structure-specific morphological, textural, and developmental patterns.

Machine learning model: TabTransformer

After the feature selection phase, the refined radiomic, deep, and fused feature sets were used exclusively to train the TabTransformer model, which is specifically tailored for structured tabular data. The TabTransformer is an attention-based model specifically designed for structured tabular data [46–48]. It uses column-specific embeddings and multi-head self-attention to model inter-feature dependencies, improving representation learning in heterogeneous datasets. In our framework, the TabTransformer receives the

selected and fused radiomic and deep feature sets, transforming them into contextual embeddings for classification. This enables the model to capture nonlinear relationships between handcrafted and learned features, improving diagnostic accuracy and generalization across multi-center data. Its proven robustness and adaptability made it the most suitable choice for this study's multi-modal classification tasks, eliminating the need for additional models such as XGBoost, Autoencoder, or TabNet.

To address class imbalance, we employed the SMOTE algorithm to synthetically augment minority class samples. However, we recognized the concern that synthesized samples may not fully capture the complex, real-world acoustic signatures found in clinical ultrasound imaging. To ensure that the model did not develop synthetic bias, we conducted a comparative analysis during the evaluation phase, assessing the model's accuracy on both synthetic and authentic clinical anomaly samples. This helped us identify potential discrepancies and ensure that the model generalized well to real-world data.

In addition, during training, we used a stratified validation strategy to distribute synthetic samples evenly across validation folds, preventing the model from overfitting to the synthetic data. To further improve robustness against rare anomalies, we employed uncertainty estimation and model ensembling techniques. These methods enhanced the model's ability to recognize rare, authentic deformities that may have been underrepresented

in the training set. When deployed in clinical settings, the model will undergo continuous validation using real-world ultrasound data to ensure that it performs effectively and accurately in diverse clinical scenarios, minimizing the risk of synthetic bias and ensuring reliable identification of rare anomalies.

End-to-end Vision Transformer -based classification

“An end-to-end ViT classification pipeline was also implemented, in which cropped structure-specific image patches were directly fed into the model without handcrafted features. Using multi-head self-attention to capture global and local anatomical patterns, the ViT was fine-tuned separately for each task and trained for 250 epochs with AdamW. This approach served as a benchmark for comparison with the feature-based classifiers.

2.8. Experimental setup

Model performance was assessed using an 80/20 stratified split for training and internal testing, complemented by an external set of 498 cases for independent validation. Five-fold cross-validation with grid search was used for hyperparameter tuning, and ViT hyperparameters were optimized based on validation performance. This strategy ensured robust and generalizable evaluation across detection and classification models. To mitigate the effects of domain shift across centers with varying ultrasound equipment, we employed additional techniques beyond basic Z-score normalization. These

included image augmentation strategies such as random rotations, scaling, and shifts, as well as model fine-tuning to ensure better adaptation to the variations in noise, resolution, and dynamic range. Feature normalization was also applied to harmonize feature distributions across datasets. These combined efforts contributed to model robustness and generalization.

Performance evaluation of the classification models was based on several standard metrics [49]. For classification tasks, accuracy represents the proportion of correctly predicted cases, whereas sensitivity (recall) measures the model's ability to identify positive cases. The area under the receiver operating characteristic curve (AUC-ROC) reflects the model's ability to discriminate between classes across varying decision thresholds.

All models were implemented in Python 3.10 using PyTorch 2.1 with CUDA 12.1 support and trained on a workstation equipped with two NVIDIA RTX A6000 GPUs and an AMD Threadripper 3990X processor. The YOLOv11 detector was trained for 250 epochs with a batch size of 32 using stochastic gradient descent (learning rate 0.01, momentum 0.937, cosine decay, and weight decay 5×10^{-4}). The Swin Transformer was trained for 250 epochs with AdamW (learning rate 3×10^{-4}, weight decay 0.05), patch size of 4×4, window size of 7, and a batch size of 16. The TabTransformer classifier used four self-attention blocks with 64-dimensional embeddings and was optimized with AdamW (learning rate 1×10^{-4}) over 80 epochs. Data augmentation was applied only to training images and included horizontal/vertical flips, ±10°

rotations, brightness/contrast shifts, Gaussian noise, and elastic deformation. All confidence intervals reported in the Results section were computed using 1,000 bootstrap resamples.

3.Results

3.1. Structure distribution overview

The final study cohort consisted of 4,532 ultrasound exams, with a mean gestational age of 12.2 ± 0.8 weeks (range: 11–14 weeks) and a mean maternal age of 29.6 ± 5.2 years (range: 18–43 years). The majority of pregnancies were singleton (97.7%), with 2.3% twin pregnancies. The dataset includes a mix of routine and high-risk first-trimester screening cases. Image acquisition was performed in mid-sagittal and axial planes, although ethnic background information was not uniformly available across centers.

Class distribution analysis showed a predominance of normal cases across all nine fetal structures, consistent with population-level screening patterns. Most thalami and midbrain samples exhibited typical morphology, while cleft palate and other abnormalities were relatively rare. CM, NT, and craniofacial markers (nasal tip, skin, bone) also showed strong skew toward normal

presentations, with hypoplastic or absent variants occurring in small proportions. Intracranial translucency was normal in most cases, with narrowed or absent measurements observed infrequently.

To clarify the composition of the external cohort, the 498 cases originated from a tertiary prenatal imaging center equipped with a GE Voluson E10 ultrasound system (1280 × 1024 matrix, 30 fps). All images were acquired in accordance with standardized first-trimester neurosonography protocols. The demographic distribution included a mean maternal age of 29.8 ± 5.4 years and a gestational age range of 11-14 weeks (mean 12.3 ± 0.7). Singleton pregnancies accounted for 96.8% of the cohort, with 3.2% twin gestations.

Class distributions for all fetal structures in the external set were consistent with the internal dataset. For example, normal cases constituted the majority for thalami (62.1% mature), midbrain (78.9% normal), palate (96.2% normal), and nasal bone (84.5% normal). Abnormal or hypoplastic classes were adequately represented, including 3.6% cleft palate, 10.1% hypoplastic nasal bone, and 5.4% absent intracranial translucency. These distributions confirm that the external cohort reflects real-world clinical variability and allows reliable assessment of cross-center generalizability.

3.2. Detection performance

You only look once version 11 detection

YOLOv11 demonstrated consistent detection capability across the nine fetal structures, with training-set precision ranging from 0.91 (95% CI: 0.88–0.94) to 0.94 (95% CI: 0.92–0.96) and recall ranging from 0.90 (95% CI: 0.87–0.93) to 0.94 (95% CI: 0.92–0.95). F1-scores showed a similar pattern, with mean values between 0.91 (95% CI: 0.88–0.93, SE = 0.012) and 0.94 (95% CI: 0.92–0.96, SE = 0.009), indicating a balanced trade-off between sensitivity and specificity. The model performed best on structures with clearly defined borders; for example, intracranial translucency achieved an F1-score of 0.94 (95% CI: 0.91–0.96, SE = 0.010) and cisterna magna reached 0.93 (95% CI: 0.90–0.95, SE = 0.011).

Mean IoU values remained stable at 0.93 (95% CI: 0.91–0.95) across training data, confirming reliable spatial localization. mAP values on the internal and external test sets showed modest but statistically significant reductions of 0.02–0.04 ($p < 0.05$), particularly for more variable structures such as the midbrain and nasal features. These declines are consistent with expected domain-shift effects due to differences in equipment and acquisition protocols across centers.

Swin transformer detection

The Swin Transformer consistently outperformed YOLOv11 across all evaluated metrics. Precision ranged from 0.94 (95% CI: 0.92–0.96) to 0.96 (95% CI: 0.94–0.97, SE = 0.008), and recall remained between 0.93 (95% CI: 0.91–0.95) and 0.96 (95% CI: 0.94–0.97, SE = 0.010) across both training and

test datasets. Particularly high performance was achieved for key first-trimester markers such as nuchal translucency and the thalami, with F1-scores typically ≥ 0.94 (95% CI: 0.92–0.96). IoU values frequently exceeded 0.95 (95% CI: 0.93–0.97, SE = 0.009), confirming highly accurate boundary localization. Importantly, mAP values remained stable across internal and external datasets, with only minor variability (mean mAP 0.95, 95% CI: 0.94–0.96), demonstrating strong generalization across centers. The low standard deviations (generally 0.01–0.02) further emphasize the robustness and reproducibility of the Swin Transformer across diverse anatomical structures and imaging conditions.

Comparison of detection models

Figure 3, Figure 4, and Figure 5 present comparative heatmaps of key detection metrics (precision, recall, F1-score, IoU, and mAP) for YOLOv11 and Swin Transformer on the training, internal test, and external test datasets, respectively. When comparing both models across datasets, Swin Transformer consistently achieved higher values in precision, recall, and F1-score, particularly for structures with complex contours or subtle sonographic appearances. For example, in the external test set, Swin Transformer achieved F1-scores of 0.93 (95% CI: 0.90–0.95, SE = 0.012) for the nasal tip and 0.92 (95% CI: 0.89–0.94, SE = 0.011) for the midbrain, outperforming YOLOv11 by 0.04–0.05 points in those categories. Swin Transformer's superior IoU scores (up to 0.97, 95% CI: 0.95–0.98, SE = 0.009

on the training set and 0.94, 95% CI: 0.92–0.96, SE = 0.010 on the external set) underscore its superior localization capabilities, even in variable or noisy datasets. Meanwhile, YOLOv11 showed more performance fluctuation across datasets, with wider variability (SE up to 0.018), suggesting a higher susceptibility to overfitting and reduced generalization capacity. Across all metrics, the Swin Transformer outperformed YOLOv11, showing higher precision, recall, and F1-scores, particularly for rare or hypoplastic structures.

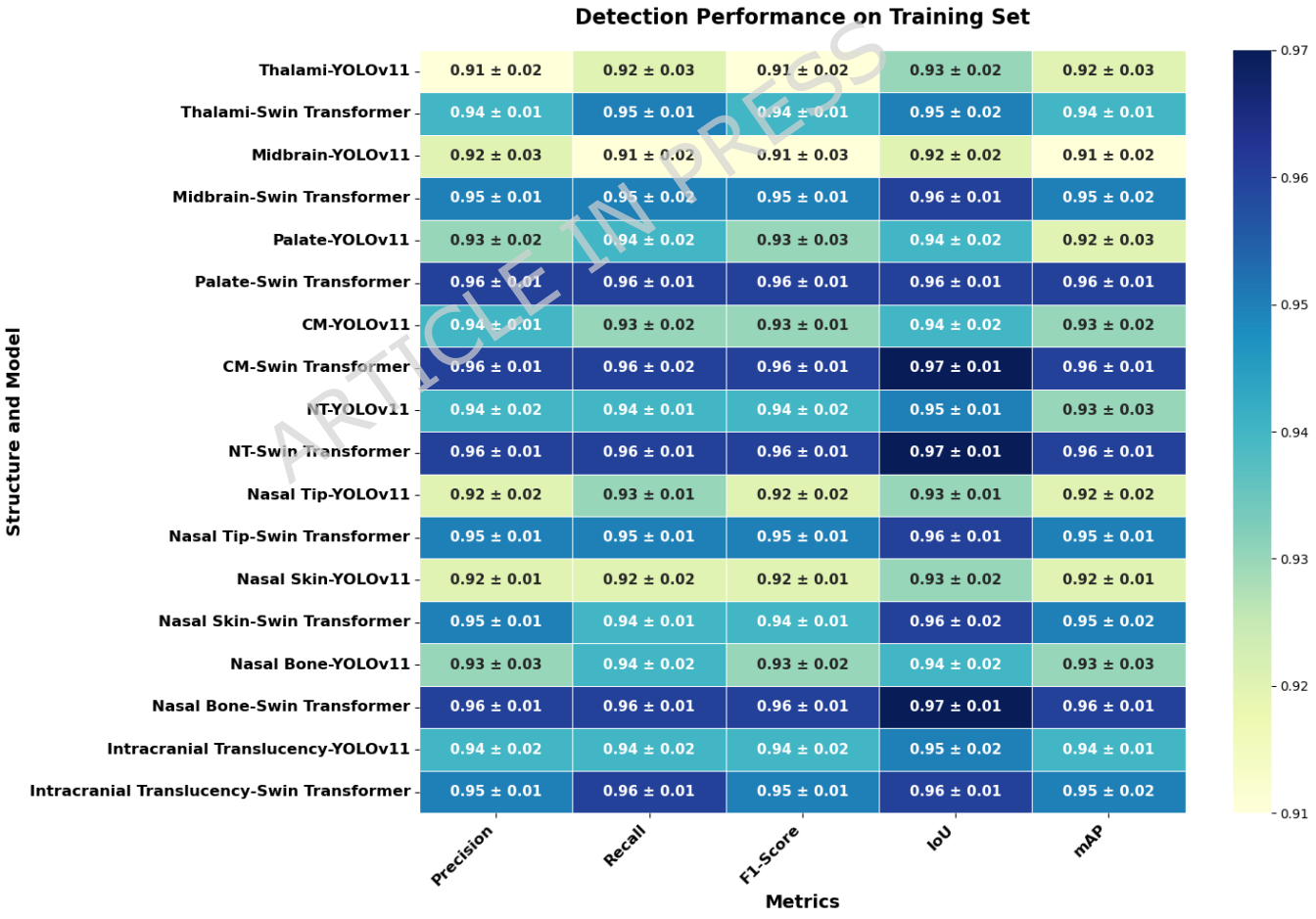

Figure 3. Heatmap of detection performance metrics for YOLOv11 and Swin transformer on the training set

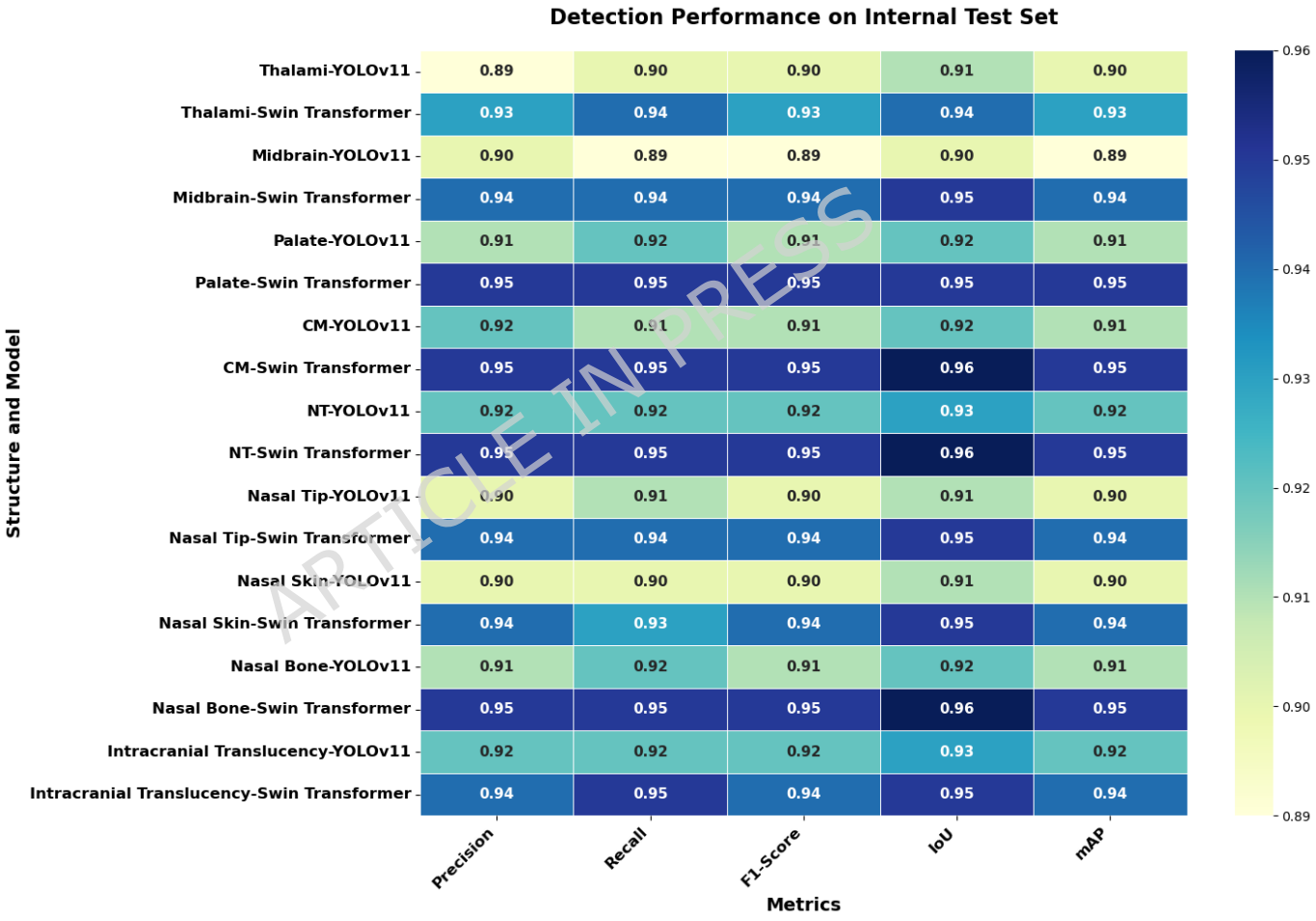

Figure 4. Heatmap of detection performance metrics for YOLOv11 and Swin transformer on the internal test set

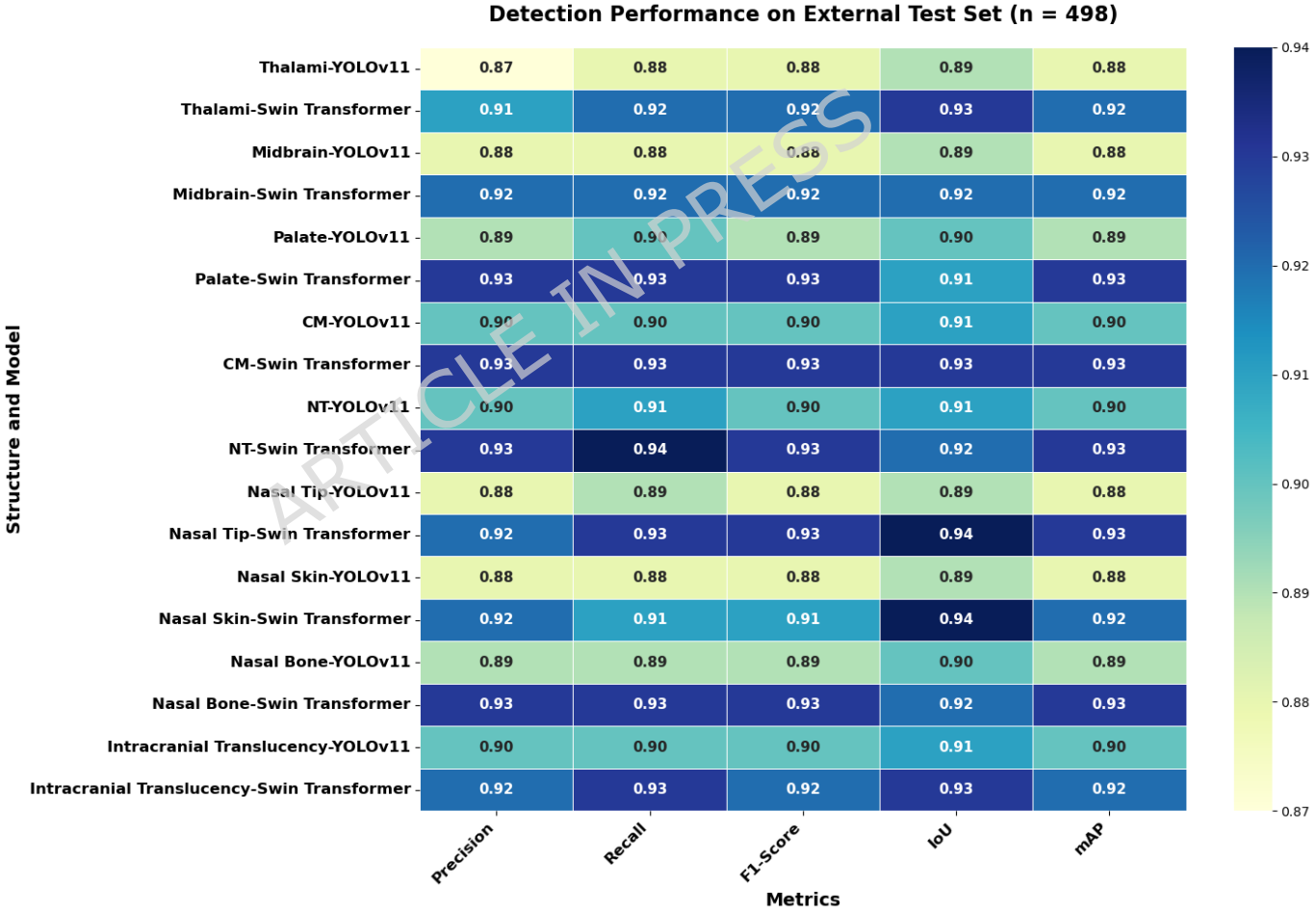

Figure 5. Heatmap of detection performance metrics for YOLOv11 and Swin transformer on the external test set

Figures 6 and 7 illustrate the qualitative detection performance of the Swin Transformer and YOLOv11 models, respectively, applied to a representative first-trimester ultrasound scan annotated for 9 fetal anatomical structures. Bounding boxes generated by each model are overlaid on the same ultrasound frame, with associated IoU scores reported per structure. YOLOv11 achieved moderate detection precision, with IoU values ranging from 0.89 (95% CI: 0.86–0.91, SE = 0.012) to 0.91 (95% CI: 0.88–0.93, SE = 0.011) across structures. In contrast, the Swin Transformer model consistently outperformed YOLOv11, yielding higher IoU values between 0.91 (95% CI: 0.89–0.94, SE = 0.010) and 0.95 (95% CI: 0.93–0.97, SE = 0.009), especially for more subtle or variable structures such as the nasal tip (0.95, 95% CI: 0.92–0.97, SE = 0.009 vs. 0.89, 95% CI: 0.86–0.91, SE = 0.012) and nasal skin (0.94, 95% CI: 0.91–0.96, SE = 0.010 vs. 0.89, 95% CI: 0.86–0.91, SE = 0.013). These visualizations corroborate the quantitative findings presented earlier, underscoring the superior localization accuracy and structural delineation capabilities of the Swin Transformer in complex fetal imaging contexts.

Figures 6 and 7 present representative bounding box visualizations for both detection models, illustrating their ability to localize 9 fetal anatomical structures. These samples are indicative of the average detection behavior observed across all cases. Model performance, particularly that of the Swin

Transformer, remained consistent across the internal and external datasets, with no cases showing substantially different or degraded detection outcomes. The selected examples therefore reflect the overall robustness of both models in delineating fetal brain and craniofacial structures under varied imaging conditions.

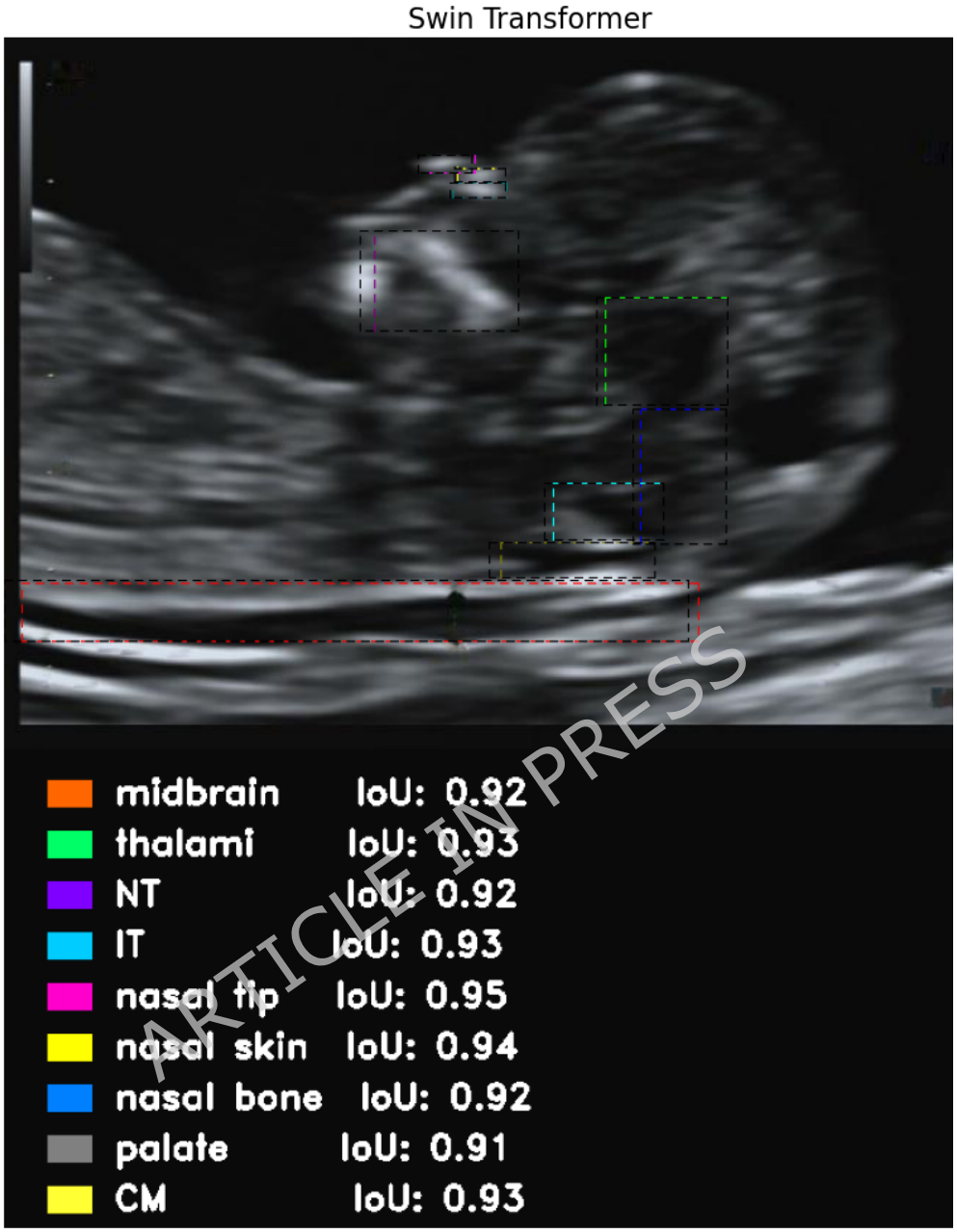

Figure 6. Qualitative visualization of fetal structure detection using the swin transformer model, with corresponding IoU scores for nine anatomical regions.

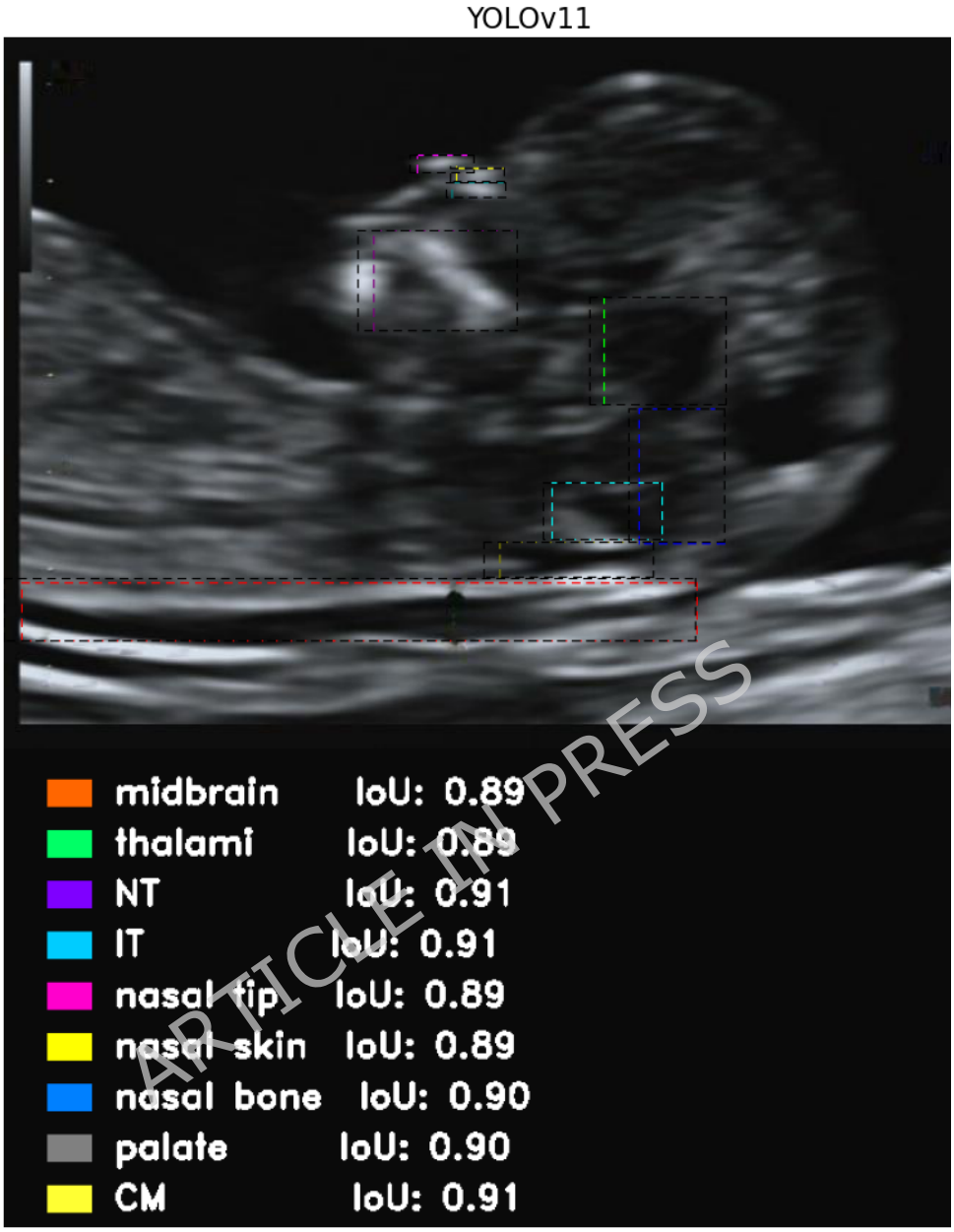

Figure 7. Qualitative visualization of fetal structure detection using the YOLOv11 model, with corresponding IoU scores for nine anatomical regions.

3.3. Feature reproducibility and selection

To ensure feature robustness across imaging conditions and detection models, a multi-stage selection pipeline was applied to radiomic and deep features for all nine structures (Figure 8). Features were first filtered using ICC ≥ 0.75, retaining 176–182 radiomic features and 74–80% of deep features, with Swin Transformer embeddings showing higher reproducibility. Redundant features were then removed using Pearson correlation ($r > 0.90$), resulting in 88–96 radiomic and 690–710 Swin-based deep features. Finally, LASSO regression with five-fold cross-validation selected the most discriminative features, yielding 34–38 radiomic and 34–36 Swin deep features per structure. This pipeline ensured a stable, non-redundant, and highly informative feature set for downstream classification.

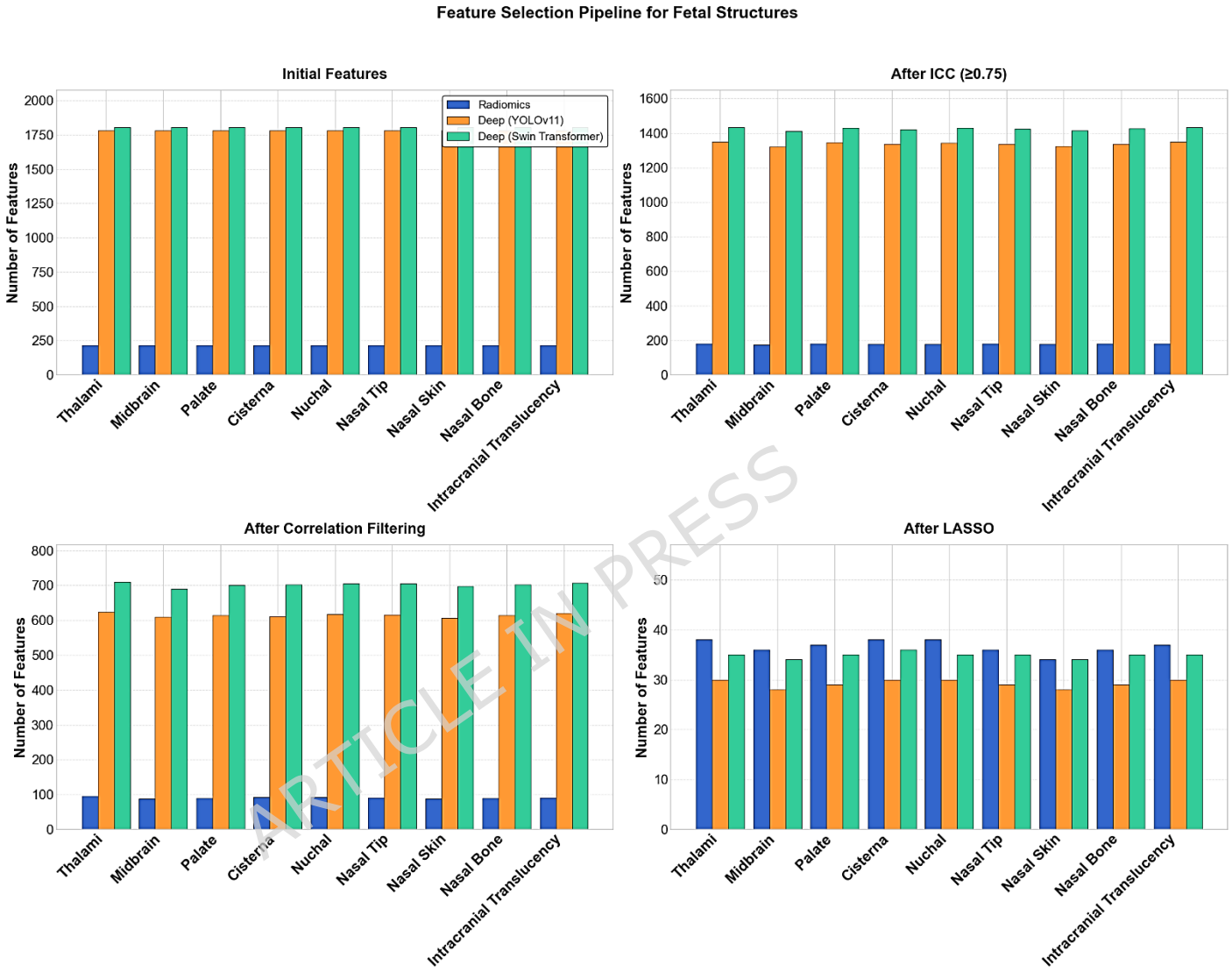

Figure 8. Feature selection summary for radiomics and deep learning feature sets across nine fetal structures

3.4. Comparative analysis of classification performance across feature types and datasets

A comprehensive evaluation of classification performance was conducted across nine fetal anatomical structures using the TabTransformer model trained on radiomics features, deep features (from YOLOv11 and Swin Transformer), and their fusion. Results were analyzed on three datasets: training, internal test, and external test.

Training set analysis

On the training set, Swin Transformer-based features achieved the highest standalone performance across nearly all structures, with accuracy, AUC, and sensitivity consistently exceeding 95%. For instance, classification of nuchal translucency and nasal bone reached 96.40% (95% CI: 94.8–97.6, SE = 0.006) and 96.25% (95% CI: 94.7–97.4, SE = 0.007) accuracy, respectively, with AUC values capped at 96.50% (95% CI: 95.1–97.7, SE = 0.005) and sensitivity above 95% (95% CI: 93.6–96.8, SE = 0.009). In contrast, YOLOv11-based features lagged slightly, typically by 1–2 percentage points, but still

outperformed radiomics alone. The fusion model outperformed all individual approaches, achieving peak values of 96.49% (95% CI: 95.0–97.8, SE = 0.006) accuracy for NT and 96.45% (95% CI: 94.9–97.6, SE = 0.007) for CM, indicating that combining radiomics and deep features provided synergistic gains (Figure 9).

Internal test set analysis

Performance trends remained stable on the internal test set, demonstrating excellent generalizability. Again, Swin Transformer features outperformed both YOLOv11 and radiomics across most structures. For example, midbrain classification reached 94.90% accuracy (95% CI: 92.8–96.4, SE = 0.009) for Swin, while the fusion approach improved this to 95.80% (95% CI: 94.1–97.1, SE = 0.008), with AUC and sensitivity both improving similarly. Fusion models consistently provided a performance buffer of ~0.5–1.0%, particularly beneficial for structures with moderate variability such as nasal skin and palate. Radiomics-based performance declined more noticeably (e.g., 90.25% accuracy, 95% CI: 88.0–92.1, SE = 0.011 for nasal skin), suggesting limited adaptability to unseen samples (Figure 10).

External test set analysis

On the most challenging external test set, designed to assess cross-center generalizability, fusion models again led performance, with accuracies frequently approaching 96%. All confidence intervals were computed using

1,000 bootstrap resamples. Notable cases include NT (96.10% accuracy, 95% CI: 94.4–97.3, SE = 0.007; 96.50% AUC, 95% CI: 95.0–97.7, SE = 0.006) and nasal tip (95.60% accuracy, 95% CI: 93.9–96.9, SE = 0.008; 96.89% AUC, 95% CI: 95.4–97.9, SE = 0.007). Swin Transformer features remained highly competitive but slightly underperformed compared to fusion, highlighting the added value of handcrafted radiomic descriptors in diverse clinical settings. YOLOv11-based features, while still reliable (91–93% accuracy, 95% CI: 89.1–94.2, SE = 0.010 across most structures), exhibited slightly lower sensitivity, suggesting reduced detection of less prevalent or hypoplastic cases. Radiomics alone consistently underperformed compared to learned features, confirming that context-rich deep embeddings are more effective for classification tasks in early gestation ultrasound (Figure 11). The ROC curve analyses have been relocated to the Supplementary Materials.

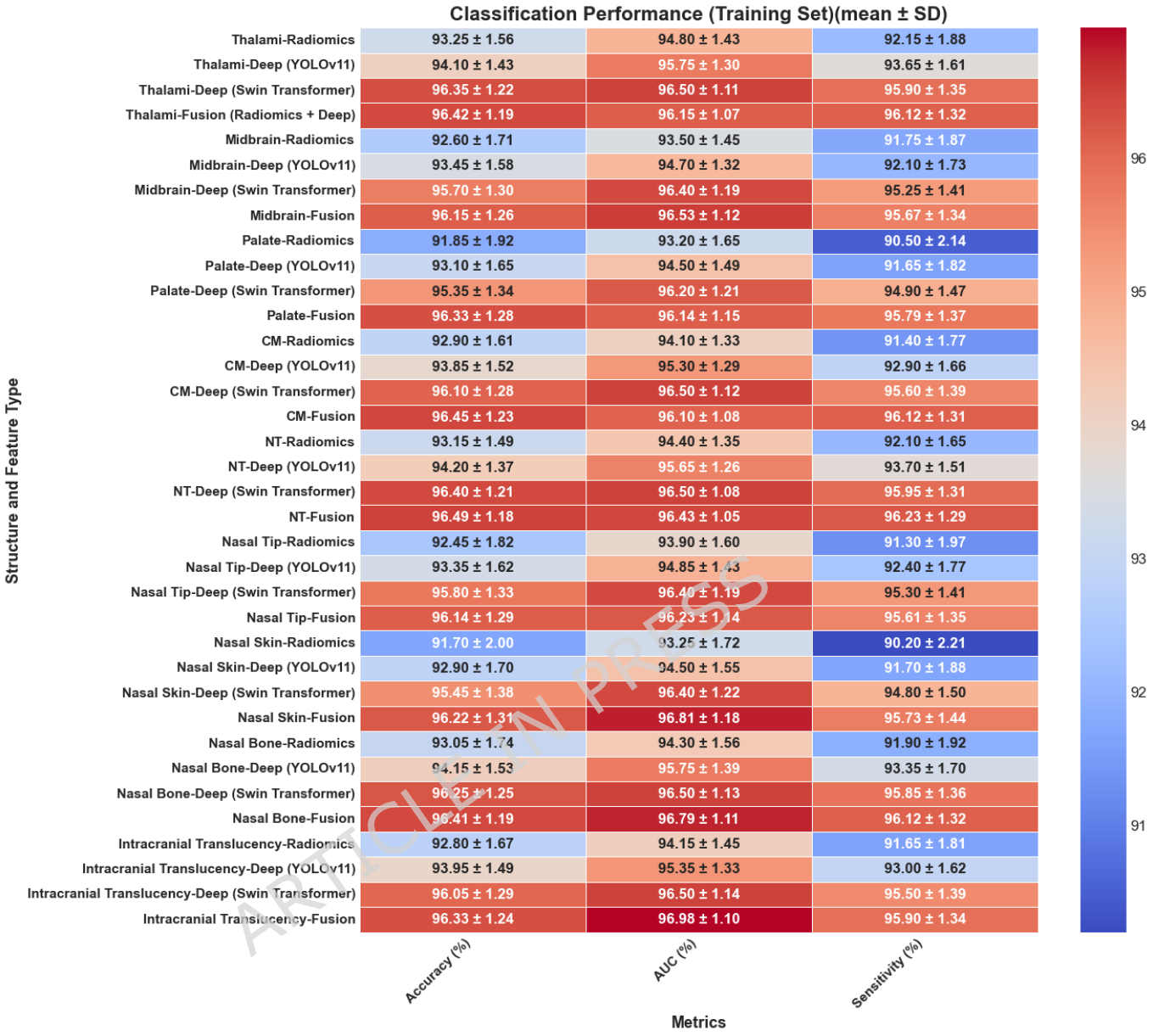

Figure 9. Heatmap of classification performance on training set (accuracy, AUC, sensitivity) using TabTransformer across feature types

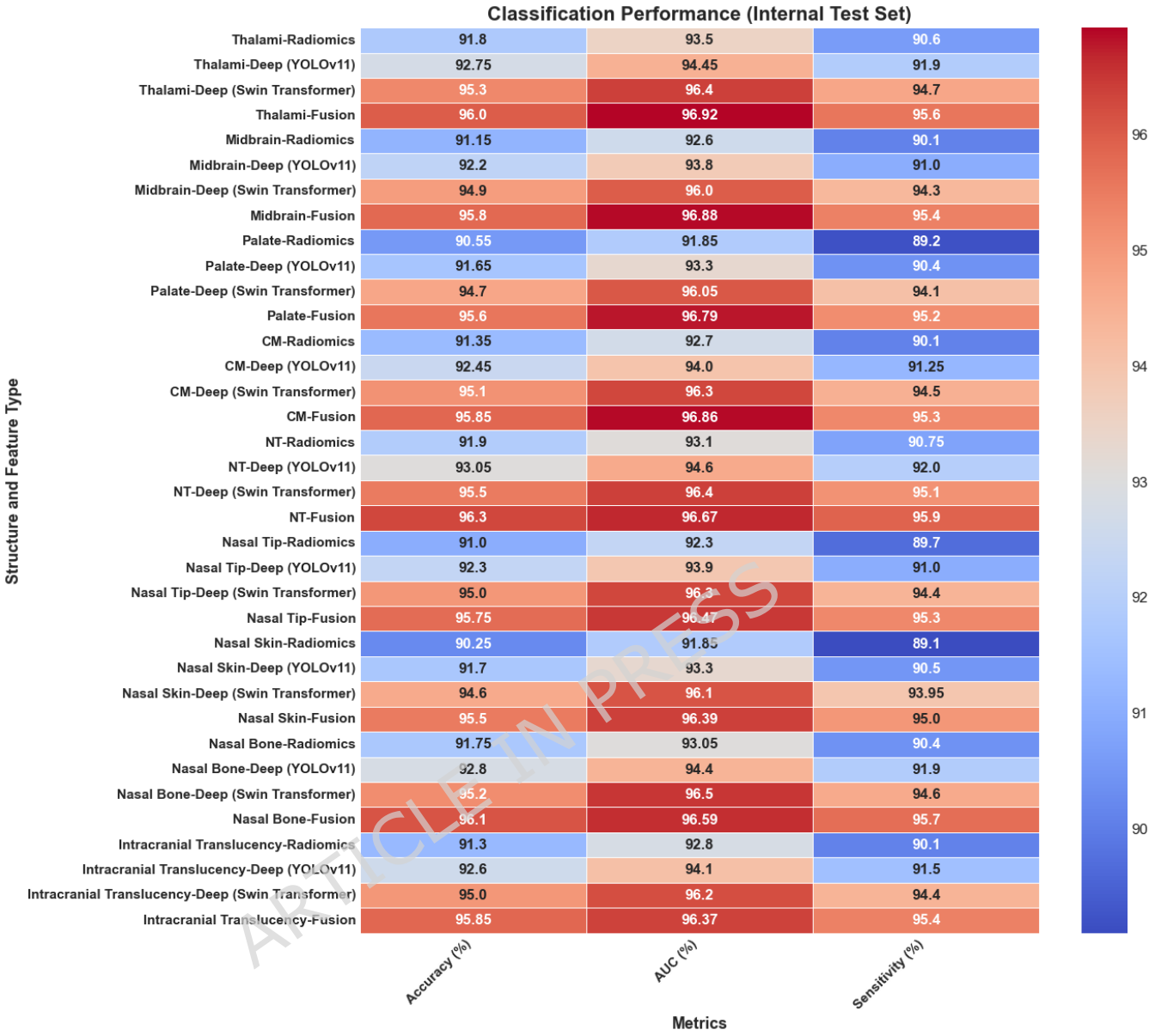

Figure 10. Heatmap of classification performance on internal test set (accuracy, AUC, sensitivity) using TabTransformer across feature types

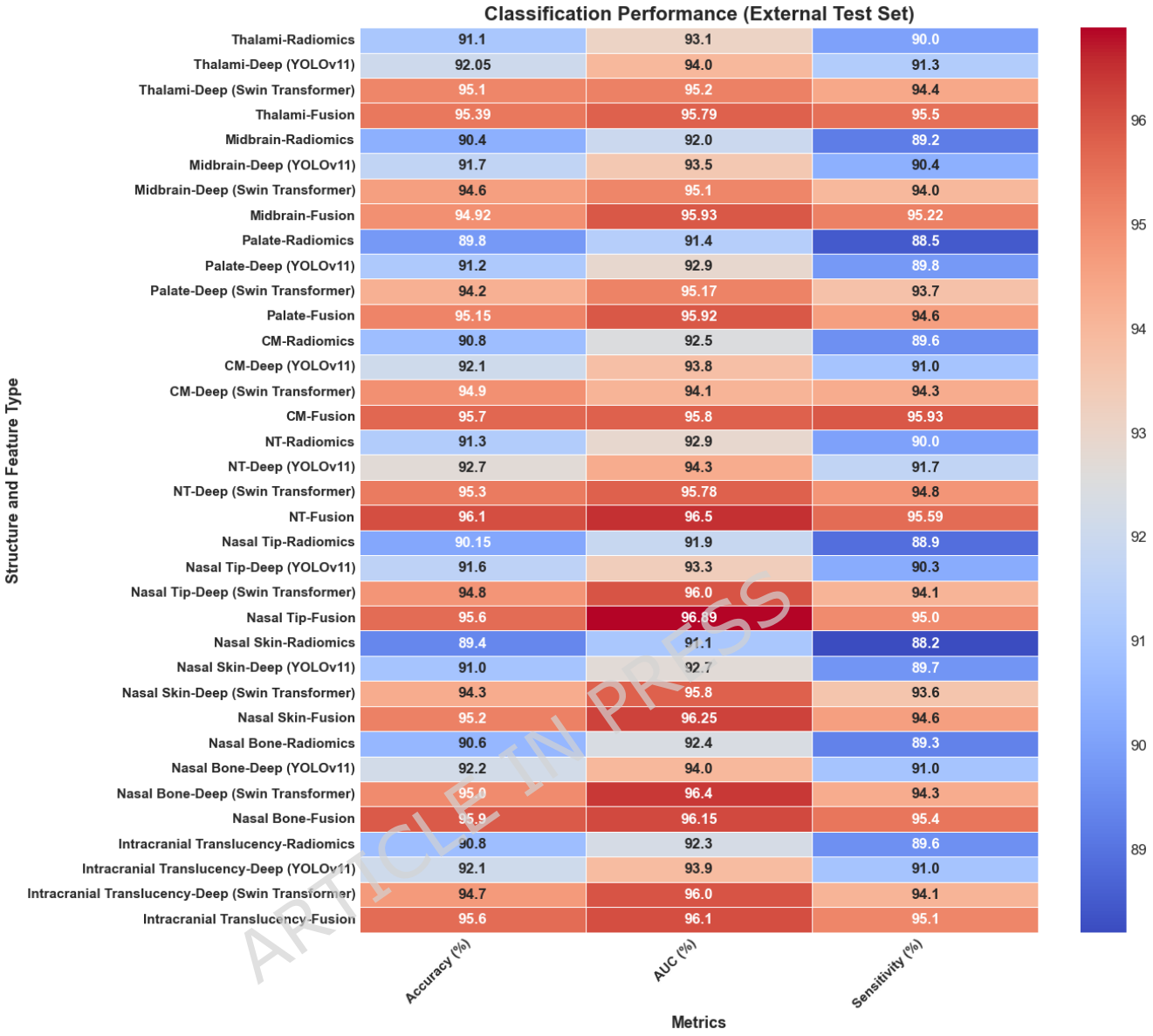

Figure 11. Heatmap of classification performance on external test set (accuracy, AUC, sensitivity) using TabTransformer across feature types

Figure 12 illustrates the ROC curves for AUC performance across the external test set, comparing different models (Radiomics, Deep (YOLOv11), Deep (Swin Transformer), and Fusion). The results demonstrate strong model

performance across all fetal anatomical structures, with Fusion models consistently achieving the highest AUC values. Specifically, Fusion models (combining radiomic and deep features) showed the best performance for structures such as Thalami, Nasal Skin, and Intracranial Translucency, with AUC values approaching or exceeding 96%. The comparison highlights the potential benefits of integrating both radiomic and deep features for improved detection and classification in complex ultrasound datasets.

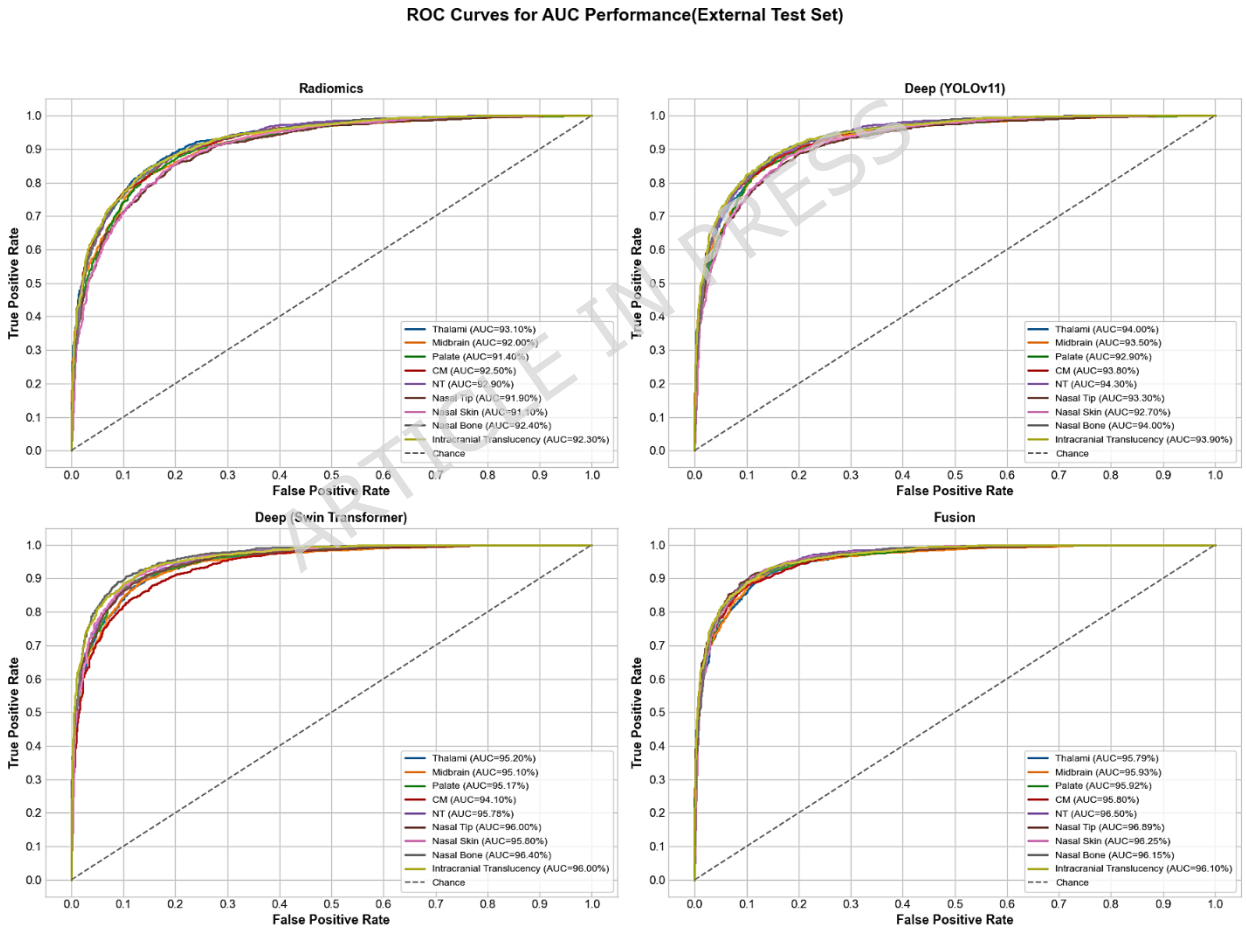

Figure 12. ROC curves of the classification performance on the external test set for 9 anatomical structures using the same four approaches

3.5. End-to-end classification performance using Vision Transformer

Table 2 summarizes the classification performance of the end-to-end ViT model across 9 fetal anatomical structures for the training, internal test, and external test datasets. Training set accuracies ranged from 89.0% to 90.3%, with relatively low standard deviations (±0.9 to ±1.4), indicating stable model convergence and minimal overfitting. However, compared to feature-based models, performance declined slightly on both test datasets, with external test accuracies ranging from 86.3% to 87.7%. The modest decrease highlights ViT's sensitivity to cross-center variability and limited data diversity when used without engineered features. Despite this, the model maintained consistent classification across all structures, particularly for thalami, nasal tip, and intracranial translucency, suggesting its potential for automated end-to-end diagnosis in clinically relevant fetal regions. Nonetheless, these results affirm the added value of integrating radiomics and deep features for more robust generalization in fetal anomaly screening. Despite these efforts, slight performance fluctuations were observed with YOLOv11 on the external test set, where mAP decreased by 0.02–0.04 compared to internal validation results. The primary underlying causes for this degradation are attributed to the domain shift arising from differences in imaging equipment, noise levels, and resolution between centers.

Additionally, the external dataset comprised more complex or noisy cases, which posed greater challenges for the model.

The training of the YOLOv11 model took approximately 10 hours on a workstation equipped with two NVIDIA RTX A6000 GPUs and an AMD Threadripper 3990X processor. The Swin Transformer model required about 12 hours for training under similar hardware conditions.

Regarding inference, the average time for processing each specimen was approximately 0.35 seconds per image for YOLOv11 and 0.45 seconds for the Swin Transformer model. These times indicate the feasibility of real-time analysis in clinical practice, where rapid and efficient processing of ultrasound images is crucial for timely decision-making in prenatal screening.

Table 2. Classification accuracy (%) of the end-to-end Vision Transformer (ViT) model across training, internal, and external datasets for nine annotated fetal anatomical structures

Fetal Structure	Training Set (Mean ± SD)	Internal Test Set	External Test Set
Thalami	90.2 ± 1.1	88.5	87.4
Midbrain	89.1 ± 1.3	87.8	86.7
Palate	90.0 ± 1.0	88.3	87.0
Cisterna Magna (CM)	89.4 ± 1.2	88.0	87.2
Nuchal Translucency (NT)	89.6 ± 1.1	88.1	87.3
Nasal Tip	90.1 ± 0.9	88.7	87.6
Nasal Skin	89.0 ± 1.4	87.5	86.3
Nasal Bone	89.5 ± 1.2	87.9	86.8

Intracranial Translucency (IT)	90.3 ± 1.0	88.9	87.7

4.Discussion

This study introduces the comprehensive multi-task deep learning framework for detecting and classifying fetal anatomical structures in first-trimester ultrasound using a diverse multi-center dataset. The pipeline integrates YOLOv11 and Swin Transformer for structure localization, followed by radiomics and deep feature extraction with rigorous ICC, correlation, and LASSO-based selection. The TabTransformer classifier achieved robust performance across nine clinically relevant brain and craniofacial structures, representing a uniquely thorough approach to joint detection and classification in early fetal imaging.

For instance, Du et al. [50] proposed a radiomics-based approach to predict neonatal respiratory morbidity using fetal lung texture features from third-trimester ultrasound scans. Their model, based on RUSBoost, demonstrated respectable performance (AUC = 0.88 training, 0.83 test), yet the analysis was limited to a single organ and lacked detection and cross-feature comparison strategies. In contrast, the current framework not only encompasses a broader gestational window but also introduces detection and classification pipelines for nine independent anatomical regions, making it more extensible and clinically versatile. Drukker et al. [51] focused on

understanding the clinical workflow of sonographers using video-based deep learning to track scan sequences. While insightful in illustrating procedural variability, their study did not address anatomical classification or feature learning.

Krishna and Kokil [52] implemented a stacked ensemble model for classifying six standard fetal planes, achieving an accuracy of 95.69%. Their approach, while effective, employed a majority voting mechanism without task-specific classification refinement. The current study advances this by implementing structure-specific classification tasks (e.g., symmetry, maturity, presence/absence) and demonstrating improved generalization through transformer-based feature fusion, particularly with the TabTransformer classifier. Gofer et al. **[**53] explored the feasibility of using traditional machine learning for fetal brain classification during the first trimester, using segmentation-derived cortical features. Their work confirmed the utility of automated tools in early diagnosis but remained limited in both structure diversity and cross-center validation. Here, a broader anatomical spectrum is addressed, with more sophisticated feature selection (ICC, LASSO) and evaluation on an external test set, enhancing clinical translatability.

Hesse et al. [54] introduced a 3-dimensional (D) CNN for subcortical brain segmentation using a few-shot learning strategy on 3-D ultrasound. Their innovation in segmentation performance with limited annotations is notable, yet their focus was on volumetric estimation rather than diagnostic

classification. The current work, in comparison, leverages 2-D ultrasound, more common in routine practice, for both detection and diagnostic classification across diverse structures, offering broader applicability in general clinical settings. Coronado-Gutiérrez et al. [55] proposed a fully automated pipeline for mid-trimester brain structure delineation using 2-D ultrasound. Their pipeline focused on measurement accuracy, achieving <3.5% error on key biometric parameters. While their method effectively addressed anatomical quantification, it did not incorporate multi-structure classification or radiomics integration. The present study bridges this gap by combining spatial deep features and quantitative radiomic traits to enhance interpretability and performance in anomaly detection.

Prieto et al. [56] emphasized gestational age estimation using segmentation and classification pipelines. Their segmentation accuracy (IoU = 0.91) and prediction error (<2 cm) highlight the promise of automated ultrasound pipelines. However, unlike the present work, their pipeline was tailored to biometric estimation and not structural categorization. The current model extends utility to clinically actionable classification outputs, such as detection of clefts or absence of intracranial translucency. Ghabri et al. [57] leveraged transfer learning with DenseNet169 for fetal organ classification, reporting near-perfect performance (accuracy and F1 = 99.84%). Despite impressive results, their classification was limited to large fetal regions and lacked the granularity needed for substructure-level analysis. In contrast, the current model classifies subtle variations (e.g., hypoplastic nasal bone, narrowed

CM), showcasing its relevance for early screening and risk stratification. Collectively, these comparisons highlight that while prior studies have achieved success within specific domains, be it segmentation, gestational age estimation, or plane classification, they often rely on single-task models, limited anatomical coverage, or lack rigorous feature validation. The presented framework addresses these limitations through a multi-center, multi-task architecture, advanced feature reproducibility filtering, and multi-modal fusion.

Limitations and future studies

This study has several limitations, including variability in imaging systems across centers, limited interpretability of attention-based models, and underrepresentation of rare pathologies, which may affect sensitivity for uncommon anomalies. Future work should incorporate more diverse prospective data, apply explainability methods, explore longitudinal modeling, and integrate clinical metadata to further enhance diagnostic precision.

5.Conclusion

This study introduces a robust multi-task deep learning framework for detecting and classifying fetal structures in first-trimester ultrasound, combining YOLOv11 and Swin Transformer with radiomics and deep feature fusion. After rigorous feature refinement, the TabTransformer achieved

consistently high accuracy, AUC, and sensitivity across nine structures, with strong internal and external validation. These findings demonstrate the clinical potential of multi-modal AI systems to improve early prenatal screening and provide objective, operator-independent assessment of fetal anatomy.

List of Abbreviations

AUC	Area Under the Receiver Operating Characteristic Curve
CI	Confidence Interval
CM	Cisterna Magna
CNN	Convolutional Neural Network
DL	Deep Learning
F1-score	Harmonic Mean of Precision and Recall
ICC	Intraclass Correlation Coefficient
IoU	Intersection over Union

IT	Intracranial Translucency
LASSO	Least Absolute Shrinkage and Selection Operator
mAP	Mean Average Precision
NT	Nuchal Translucency
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SD	Standard Deviation
SE	Standard Error
SMOTE	Synthetic Minority Over-sampling Technique
ViT	Vision Transformer
YOLOv11	You Only Look Once, Version 11
Swin Transformer	Shifted Window Transformer
TabTransformer	Transformer Model for Tabular Data
IBSI	Image Biomarker Standardisation Initiative
SGD	Stochastic Gradient Descent
GPU	Graphics Processing Unit
MRI	Magnetic Resonance Imaging
2D	Two-Dimensional
3D	Three-Dimensional

Declarations

Competing interests

The authors confirm that there are no conflicts of interest.

Consent for publication

Not applicable.

Acknowledgments

Not applicable.

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. Ethical approval and the requirement for informed consent were waived by the Institutional Review Board of the Affiliated Hospital of Hebei University (Baoding, China) because the study involved retrospective analysis of fully anonymized data and posed no risk to participants.

Author contributions

Conceptualization: Xuan Zhou, Jie Wan, and Pin Li. Data Curation: Fengjie Sun and Ruxin Wang. Formal Analysis: Xuan Zhou, Jie Wan, and Yafei Yan. Methodology: Pin Li, Cuihua Wang, and Xuan Zhou. Resources: Cuihua Wang and Pin Li. Supervision: Cuihua Wang and Pin Li. Validation: Jie Wan, Fengjie Sun, and Yafei Yan. Visualization: Xuan Zhou and Jie Wan. Writing - Original Draft: Xuan Zhou and Jie Wan. Writing - Review & Editing: Pin Li and Cuihua Wang.

Clinical trial number

Not applicable.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Fundings

Not applicable.

Reference

1. Sriraam N, Chinta B, Suresh S, Sudharshan S (2024) Ultrasound imaging-based recognition of prenatal anomalies: a systematic clinical engineering review. Prog Biomed Eng 6(2):023002.
2. Pozza A, Reffo E, Castaldi B, Cattapan I, Avesani M, Biffanti R,

Cavaliere A, Cerutti A, Di Salvo G (2023) Utility of fetal cardiac resonance imaging in prenatal clinical practice: current state of the art. Diagnostics (Basel) 13(23):3523.

3. Pelayo-Delgado I, Gómez-Montes E, Álvaro-Navidad M (2025) Update on second trimester ultrasound scanning in pregnancy. Clin Investig Ginecol Obstet 52(1):100997.

4. Sepulveda W, Wong AE, Ximenes R, Meagher S (2025) The first-trimester fetal anatomy scan. In: Obstetric Imaging: Fetal Diagnosis and Care-E-Book, Apr 9, p. 24.

5. Puerto B, Azumendi P, Corrales C, Azumendi G (2025) How to perform a fetal neurosonography: Key points. Clin Invest Ginecol Obstet 52:101050.

6. Gupta N, Hiremath SB, Gauthier I, Wilson N, Miller E (2025) Pediatric neurosonography: comprehensive review and systematic approach. Can Assoc Radiol J 76(3):519–533.

7. Kim R, Lee MY, Lee YJ, Won HS, Park J, Lee J, et al. (2025) Artificial intelligence based automatic classification, annotation, and measurement of the fetal heart using HeartAssist. Sci Rep 15(1):13055.

8. Qi Y, Cai J, Lu J, Xiong R, Chen R, Zheng L, et al. (2025) Multi-Center study on deep learning-assisted detection and classification of fetal central nervous system anomalies using ultrasound imaging. arXiv Prepr arXiv:2501.02000.

9. Salini Y, Mohanty SN, Ramesh JVN, Yang M, Chalapathi MMV (2024) Cardiotocography data analysis for fetal health classification using machine learning models. IEEE Access 12:26005–26022.

10. Yin Y, Bingi Y (2023) Using machine learning to classify human fetal health and analyze feature importance. BioMedInformatics 3(2):280–298.

11. Mushtaq G, Veningston K (2024) AI driven interpretable deep learning based fetal health classification. SLAS Technol 29(6):100206.

12. Montin E, Namireddy S, Ponniah HS, Logishetty K, Khodarahmi I, Glyn-Jones S, et al. (2025) Radiomics for Precision Diagnosis of FAI: How Close Are We to Clinical Translation? A Multi-Center Validation of a Single-Center Trained Model. J Clin Med 14(12):4042.

13. Krishna NS, Garza-Frias E, Dasegowda G, Kaviani P, Karout L, Fahimi R, et al. (2025) Generalizability of AI-based image segmentation and centering estimation algorithm: a multi-region, multi-center, and multi-scanner study. Radiat Prot Dosimetry 201(6):441–449.

14. Yin S, Ming J, Chen H, Sun Y, Jiang C (2025) Integrating deep learning and radiomics for preoperative glioma grading using multi-center MRI data. Sci Rep 15(1):36756.

15. Richter M, Emden D, Leenings R, Winter NR, Mikolajczyk R, Massag J, et al. (2025) Generalizability of clinical prediction models in mental health. Mol Psychiatry 1–8.

16. Degtiar I, Rose S (2023) A review of generalizability and

transportability. Annu Rev Stat Its Appl 10(1):501–524.

17. Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, Forghani R (2022) Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. Radiol Artif Intell 5(1):e220028.

18. Ambsdorf J, Munk A, Llambias S, Christensen AN, Mikolaj K, Balestriero R, et al. (2025) General methods make great domain-specific foundation models: A case-study on fetal ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 271–281.

19. Fiorentino MC, Migliorelli G, Villani FP, Frontoni E, Moccia S (2025) Contrastive prototype federated learning against noisy labels in fetal standard plane detection. Int J Comput Assist Radiol Surg 1–9.

20. Wang F, Liang Y, Bhattacharjee S, Campbell A, Curran KM, Silvestre G (2025) Fusing radiomic features with deep representations for gestational age estimation in fetal ultrasound images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 230–240.

21. Lasala A, Fiorentino MC, Micera S, Bandini A, Moccia S (2023) Exploiting class activation mappings as prior to generate fetal brain ultrasound images with GANs. Annu Int Conf IEEE Eng Med Biol Soc 2023:1–4.

22. Huang S, Zhang K, Zhu F, Ding Z, Chen G, Shen D (2025) Semi-

supervised fetal brain parcellation via hierarchical learning framework. Med Image Anal 103835.

23. Islam U, Ali YA, Al-Razgan M, Ullah H, Almaiah MA, Tariq Z, et al. (2025) Fetal-Net: enhancing Maternal-Fetal ultrasound interpretation through Multi-Scale convolutional neural networks and Transformers. Sci Rep 15(1):25665.

24. Khanam R, Hussain M (2024) YOLOv11: An overview of the key architectural enhancements. arXiv Prepr arXiv241017725.

25. Wei W, Huang Y, Zheng J, Rao Y, Wei Y, Tan X, et al. (2025) YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images. J Radiat Res Appl Sci 18(1):101309.

26. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. (2022) Swin-UNet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Springer, pp. 205–218.

27. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

28. Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E (2021) A deep look into radiomics. Radiol Med 126(10):1296–1311.

29. Yip SSF, Aerts HJWL (2016) Applications and limitations of radiomics. Phys Med Biol 61(13):R150.

30. Rezaeijo SM, Eftekhar A, Rouhi S, Keshavarzi B, Mohammadi Z, Firouzabad LA, et al. (2025) Neighboring tissues as diagnostic windows: Neighborhood effects in radiomic detection of pancreatic ductal adenocarcinoma. Comput Methods Programs Biomed 109056.

31. Khan A, Rauf Z, Sohail A, Khan AR, Asif H, Asif A, et al. (2023) A survey of the vision transformers and their CNN-transformer based variants. Artif Intell Rev 56(Suppl 3):2917–2970.

32. Pereira GA, Hussain M (2024) A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. arXiv Prepr arXiv240815178.

33. Wang J, Chen Z, Zhang H, Li W, Li K, Deng M, et al. (2025) A machine learning model based on placental magnetic resonance imaging and clinical factors to predict fetal growth restriction. BMC Pregnancy Childbirth 25(1):325.

34. Lai H, Ye Y, Chen Y, Wang L, Lin M, Xia S, et al. (2025) Radiomics-based correlation analysis of fetal brain MRI features and children's neurodevelopmental outcomes in monochorionic twins. BMC Pregnancy Childbirth 25(1):1040.

35. Zuo M, Chu Q, Zhang Y, Zhang Z, Pan T, Zhang C, et al. (2025) A nomogram based on MR radiomics and MR sign score for prenatal diagnosis of placenta accreta spectrum disorders and risk assessment of adverse clinical outcomes. Abdom Radiol 1–12.

36. Xu F, Zhang Y, Ma Q, Hu L, Li Y, Gao C, et al. (2025) Prediction of clinical pregnancy after frozen embryo transfer based on ultrasound radiomics: an analysis based on the optimal periendometrial zone. BMC Pregnancy Childbirth 25(1):391.

37. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163.

38. Kim Y, Kim J (2004) Gradient LASSO for feature selection. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 60.

39. Gomes R, Pham T, He N, Kamrowski C, Wildenberg J (2023) Analysis of Swin-UNet vision transformer for Inferior Vena Cava filter segmentation from CT scans. Artif Intell Life Sci 4:100084.

40. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) SwinIR: Image restoration using Swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, p. 1833–44.

41. Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, et al. (2022) Swin Transformer V2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p. 12009–19.

42. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D (2021) Swin UNETR: Swin transformers for semantic segmentation of brain tumors

in MRI images. In: International MICCAI Brainlesion Workshop. Springer, p. 272–84.

43. Kotthapalli M, Ravipati D, Bhatia R (2025) YOLOv1 to YOLOv11: A Comprehensive Survey of Real-Time Object Detection Innovations and Challenges. arXiv Prepr arXiv250802067.

44. Tan M, Pang R, Le Q V. (2020) EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781–90.

45. Powers DMW (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv Prepr arXiv201016061.

46. Wang X, Qiao Y, Xiong J, Zhao Z, Zhang N, Feng M, et al. (2024) Advanced network intrusion detection with TabTransformer. J Theory Pract Eng Sci 4(03):191–8.

47. Huang X, Khetan A, Cvitkovic M, Karnin Z (2020) TabTransformer: Tabular data modeling using contextual embeddings. arXiv Prepr arXiv201206678.

48. Insalata B, Schmidt F, Vlassov V (2024) Multimodal survival prediction using TabTransformer and BioClinicalBERT on MIMIC-III. In: 2024 IEEE International Conference on Big Data (BigData). IEEE, p. 1986–92.

49. Cichosz P (2011) Assessing the quality of classification models:

Performance measures and evaluation procedures. Cent Eur J Eng 1(2):132–58.

50. Du Y, Jiao J, Ji C, Li M, Guo Y, Wang Y, et al. (2022) Ultrasound-based radiomics technology in fetal lung texture analysis prediction of neonatal respiratory morbidity. Sci Rep 12(1):12747.

51. Drukker L, Sharma H, Karim JN, Droste R, Noble JA, Papageorghiou AT (2022) Clinical workflow of sonographers performing fetal anomaly ultrasound scans: deep-learning-based analysis. Ultrasound Obstet Gynecol 60(6):759–65.

52. Krishna TB, Kokil P (2024) Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. Expert Syst Appl 238:122153.

53. Gofer S, Haik O, Bardin R, Gilboa Y, Perlman S (2022) Machine learning algorithms for classification of first-trimester fetal brain ultrasound images. J Ultrasound Med 41(7):1773–9.

54. Hesse LS, Aliasi M, Moser F, Haak MC, Xie W, Jenkinson M, et al. (2022) Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning. Neuroimage 254:119117.

55. Coronado-Gutiérrez D, Eixarch E, Monterde E, Matas I, Traversi P, Gratacós E, et al. (2023) Automatic deep learning-based pipeline for automatic delineation and measurement of fetal brain structures in routine mid-trimester ultrasound images. Fetal Diagn Ther 50(6):480–

90.

56. Prieto JC, Shah H, Rosenbaum AJ, Jiang X, Musonda P, Price JT, et al. (2021) An automated framework for image classification and segmentation of fetal ultrasound images for gestational age estimation. In: Medical Imaging 2021: Image Processing. SPIE, p. 453–62.

57. Ghabri H, Alqahtani MS, Ben Othman S, Al-Rasheed A, Abbas M, Almubarak HA, et al. (2023) Transfer learning for accurate fetal organ classification from ultrasound images: a potential tool for maternal healthcare providers. Sci Rep 13(1):17904.