Est. 1841 | YORK ST JOHN UNIVERSITY

Gornale, Shivanand, Kamat, Priyanka, Hiremath, Prakash and Siddalingappa, Rashmi (2025) A Hybrid Ensemble of Denoising Autoencoders and Deep Learning Models for Fetal Image Analysis. Cureus Journal Of Computer Science.

Downloaded from: https://ray.yorksj.ac.uk/id/eprint/12901/

The version presented here may differ from the published version or version of record. If you intend to cite from the work you are advised to consult the publisher's version: https://doi.org/10.7759/s44389-025-09506-x

Research at York St John (RaY) is an institutional repository. It supports the principles of open access by making the research outputs of the University available in digital form. Copyright of the items stored in RaY reside with the authors and/or other copyright owners. Users may access full text items free of charge, and may download a copy for private study or non-commercial research. For further reuse terms, see licence terms governing individual outputs. Institutional Repositories Policy Statement

RaY

Research at the University of York St John

For more information please contact RaY at ray@yorksj.ac.uk

Cureus Journal of Computer Science

Open Access Original Article

A Hybrid Ensemble of Denoising Autoencoders and Deep Learning Models for Fetal Image Analysis

Received 08/10/2025
Review began 08/24/2025
Review ended 09/19/2025
Published 09/22/2025

© Copyright 2025
Gornale et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI:
https://doi.org/10.7759/s44389-025-09506-x

Shivanand S. Gornale [1] , Priyanka C. Kamat [1] , Prakash S. Hiremath [2] , Rashmi Siddalingappa [3]

1. Department of Computer Science, Rani Channamma University, Belagavi, IND 2. Department of Computer Applications, KLE Technological University, Hubli, IND 3. Department of Computer Science, York St John University, London, GBR

Corresponding author: Priyanka C. Kamat, priya.kamat25@gmail.com

Abstract

Medical image analysis, particularly ultrasonography, has involved increasing attention in computer science and engineering due to its potential for automated and scalable interpretation. Ultrasound imaging is widely used in prenatal care because of its non-invasive nature and cost-effectiveness. Automated analysis of fetal ultrasound images can improve diagnostic accuracy and reduce inter-observer variability. However, challenges such as speckle noise, low contrast, and anatomical variations across trimesters make automated interpretation difficult, requiring robust preprocessing, segmentation, and classification methods.

This study proposes a hybrid ensemble deep learning framework for analyzing fetal ultrasound images. The framework integrates a denoising autoencoder for noise reduction and image enhancement, as well as seven segmentation architectures (U-Net, DeepLabV3+, DenseNet-U-Net, MFP-UNet, Attention U-Net, MobileNet-U-Net, and ResNet-U-Net), and five ensemble strategies (maximum voting, majority voting, weighted voting, confidence-based fusion, and averaging) to enhance segmentation performance. A multi-input classification approach is also introduced, combining individual and ensemble segmentation outputs in a fine-tuned DenseNet121 for trimester categorization (first, second, and third trimesters) based on head circumference and femur length.

The framework is evaluated using Dice score, mean intersection over union, accuracy, precision, recall, and F1-score. Experimental results show that ensemble strategies significantly improve segmentation. The multi-input classification achieves 92.50% accuracy for head circumference and 90.60% for femur length on the custom dataset, as well as 83.68% on the HC18 dataset, outperforming individual models.

The main contributions include (1) a hybrid ensemble strategy for robust segmentation and (2) a multi-input trimester classification method. The proposed framework is generalizable and can be extended to other medical imaging applications beyond fetal ultrasound analysis.

Categories: Image Processing and Analysis, Medical Expert systems, Deep Learning
Keywords: ultrasound medical image, medical image analysis, image segmentation, ensemble approach, trimester-based image classification, multi-input classification, deep learning techniques, feature extraction

Introduction

Ultrasound imaging is a crucial diagnostic modality in prenatal healthcare due to its affordability, portability, and non-invasive nature without ionizing radiation [1]. Clinicians monitor fetal development across three trimesters (0-13, 14-26, and 27-40 weeks) using biometric parameters such as Head Circumference (HC), Femur Length (FL), Abdominal Circumference (AC), and Crown-Rump Length to estimate gestational age and assess developmental progression [2-5]. However, automated fetal ultrasound analysis is challenging due to high intra-class variability caused by inconsistent anatomical views, fetal movement, and gestational changes, along with imaging artifacts such as speckle noise, acoustic shadowing, and low tissue contrast. These factors obscure anatomical boundaries and increase diagnostic errors, necessitating robust computational techniques for accurate structure detection and biometric measurement [6-9].

Traditional computer vision approaches, including Gaussian and median filtering for denoising, Otsu's thresholding, edge detection, and region-growing for segmentation, have shown limited success because they rely on handcrafted features and rigid assumptions, often leading to over-smoothing and loss of anatomical details [10-12]. Similarly, classical machine learning methods, such as Support Vector Machines and Random Forests, require manual feature extraction and fail to capture complex spatial relationships in fetal anatomy, making them vulnerable to intensity variations and artifacts in ultrasound images [13-15]. Although deep learning-based models have improved segmentation accuracy, most

existing studies are trimester-agnostic, training on mixed-trimester data without accounting for the anatomical and contrast variability observed across gestational stages. For instance, first trimester (FT) images often show small fetal heads with low contrast and poorly defined boundaries, whereas second trimester (ST) and third trimester (TT) images display larger head sizes, brighter skull edges, and increased ossification. FL images similarly exhibit trimester-dependent variations in size, shape, and orientation, causing inconsistent performance when models trained on mixed data are applied to different gestational stages [16-20]. These differences are visually demonstrated in Figure *1*, where trimester-wise ultrasound samples show clear anatomical variability: (a) our created dataset containing images with HC and FL parameters and (b) the publicly available HC18 dataset containing images with HC parameters.

The originality of this study lies in its trimester-aware, multi-stage computational pipeline designed specifically for fetal biometric analysis. Unlike existing approaches, our framework incorporates a trimester-specific classification module, enabling gestational-stage-aware processing and allowing the model to adapt to trimester-dependent anatomical variability [21-24]. Additionally, a denoising autoencoder is integrated as a dedicated preprocessing stage to enhance ultrasound image quality while preserving fine anatomical details, an aspect rarely explored in trimester-specific fetal analysis. Furthermore, an ensemble of seven state-of-the-art architectures (U-Net, DeepLabV3+, DenseNet-U-Net, MFP-UNet, Attention U-Net, MobileNet-U-Net, and ResNet-U-Net) is employed not as a simple reuse of existing models but to leverage their complementary strengths, such as robust encoder-decoder design, attention mechanisms, and multi-path feature aggregation, to achieve consistent segmentation across diverse anatomical presentations [25-30]. By aligning computational modeling with fetal developmental stages, the proposed method significantly improves segmentation consistency and biometric measurement precision, supports early anomaly detection, and contributes to evidence-based clinical decision-making and improved prenatal care [31-34].

The main contributions of this study are summarized as follows:

• Creation of a new custom dataset including 1,426 HC and 1,404 FL ultrasonic images across all trimesters.

• Integration of fetal age as a crucial feature, which was absent in HC18, enabling precise trimester-based analysis.

• A new pipeline is designed, integrating a denoising autoencoder, an ensemble of segmentation models, and a trimester-specific classification module tailored for fetal biometric parameters.

• Extensive experiments classifying HC and FL parameters into FT, ST, and TT, providing detailed trimester-specific insights.

• Demonstration of significant improvements in classification accuracy compared to HC18, validating the new dataset's value in more accurate fetal measurements.

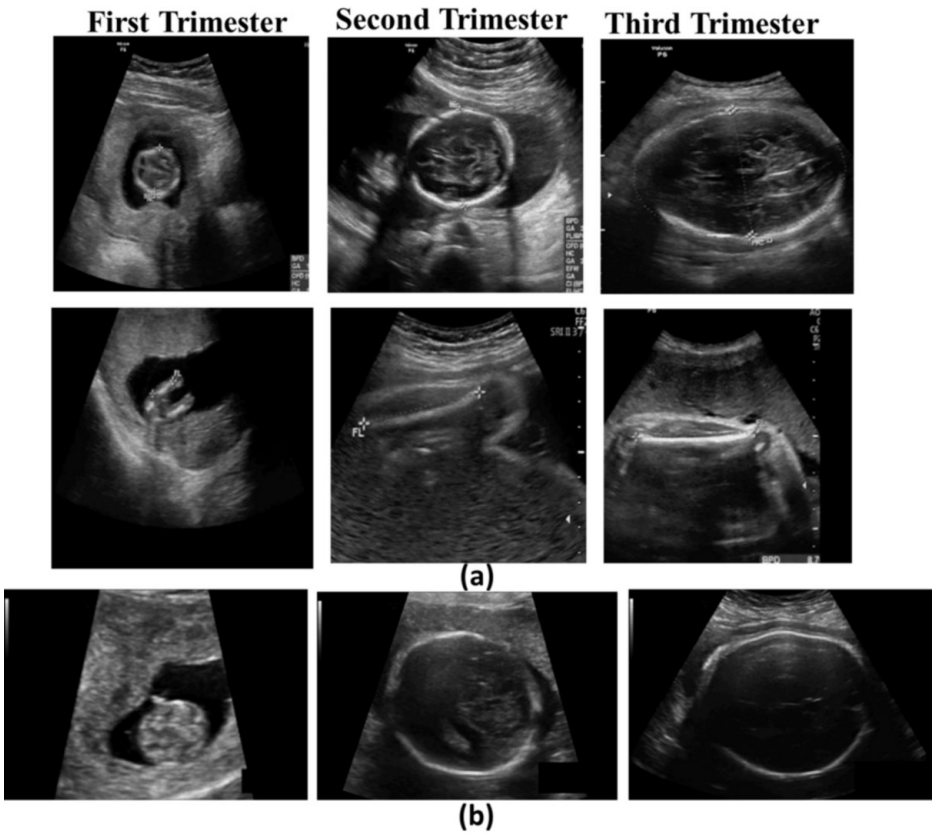

FIGURE 1: Trimester-wise ultrasound samples: (a) Created dataset containing images with HC and FL parameters. (b) The publicly available HC18 dataset containing images with HC parameters

FL, Femur Length; HC, Head Circumference

Related work

Mengistu et al. [3] have developed a deep learning model for detecting fetal head abnormalities from ultrasound images using data from Ethiopian healthcare facilities. Among various architectures, SegNet achieved the best performance, with 98% accuracy and a Dice coefficient of 0.97. The model accurately classified microcephaly, macrocephaly, and normal cases using WHO guidelines and showed strong agreement with expert measurements for BPD and HC.

Danish et al. [5] have introduced a dataset of 500 ultrasound scans from 4 to 10 weeks of gestation created for gestational sac segmentation. UNet, UNet++, DeepLabV3, and ResUNet models, each with a ResNet50 encoder, were trained and evaluated using 5-fold cross-validation. Among them, ResUNet outperformed others with a Dice score of 0.978 and intersection over union (IoU) of 0.946. A new biometry-based method was also proposed for automatic gestational age estimation, achieving a low mean absolute error of just 0.07 weeks compared to expert sonographers.

Chougule et al. [6] have proposed a method to measure fetal HC from 2D ultrasound images using a combination of biometry-based image processing and segmentation techniques. Models like SegNet, GCN, and HRNet were evaluated for semantic segmentation, with HRNet achieving the highest performance, yielding an average Dice score of 96%.

Halder et al. [7] have presented the Residual U-Net, demonstrating superior performance over traditional U-Net and Attention U-Net in fetal head segmentation, effectively addressing the vanishing gradient problem. It achieved a highest Dice coefficient of 97.17% and Jaccard Index of 94.51% on the validation set. The study also included fetal HC measurement and head position estimation.

Alzubaidi et al. [9] have introduced ETLM, a novel method that combines transfer learning with ensemble learning for fetal head segmentation and measurement in ultrasound images. After evaluating eight segmentation networks, their ensemble approach achieved an mean IoU (MIoU) is 98.53%.

Ashkani Chenarlogh et al. [10] have proposed the Fast U-Net, a lightweight architecture designed to reduce computational load in clinical settings. Evaluated on datasets for AC and HC segmentation, the model maintained a Dice coefficient of 97.45% to standard U-Net while significantly reducing processing time, making it suitable for real-time clinical deployment.

Dubey et al. [12] have proposed the DR-ASPnet model for fetal head (FH) segmentation and HC estimation in ultrasound images. To tackle issues like image blurring and pixel size variation, they employed pre-processing and data augmentation techniques. The model combines appearance-based and hierarchical density regression with a deep convolutional classifier to improve segmentation precision. It achieved a Dice coefficient of 98.86%.

Sobhaninia et al. [15] have proposed a multi-task deep learning model based on the Link-Net architecture for fetal head segmentation and circumference estimation in 2D ultrasound. Incorporating an Ellipse Tuner module, the model trained on 999 images demonstrated improved segmentation performance over single-task networks, producing smoother and more accurate elliptical outlines.

Zeng et al. [17] have introduced DAG V-Net, a deeply supervised attention-gated V-Net model for fetal head segmentation in ultrasound images. By integrating attention mechanisms and deep supervision, DAG V-Net outperformed traditional U-Net and V-Net models, achieving a DSC of 97.63%.

Nagabotu et al. [19] have developed an enhanced U-Net model that incorporates attention mechanisms and scale information to improve the segmentation of fetal head measurements from 2D ultrasound images [35-36]. The model demonstrated superior accuracy, with DSC values of 97.90% and MIoU values of 97.81%, in predicting key parameters, including head circumference, occipitofrontal diameter, and biparietal diameter, compared to existing approaches.

Rayed et al. [20] have comprehensively reviewed deep learning techniques in medical image segmentation, discussing commonly used preprocessing steps, datasets, and architectures. The review assesses the strengths and limitations of current methods, outlining ongoing challenges and providing valuable guidance for future research and innovation in the field.

Al-Razgan et al. [22] have developed the AG-CNN model, which uses adaptive feature extraction and attention mechanisms to enhance fetal anatomical plane detection. The model outperformed DenseNet169, ResNet50, and VGG16, achieving lower losses and higher accuracies on curated datasets. The AG-CNN model achieved an accuracy of 94%.

Wang et al. [24] have introduced the FT Decoder, an efficient fine-tuning strategy for U-Net, targeting improved fetal head segmentation in ultrasound images from low-resource settings. By training only the decoder stack, the approach reduces trainable parameters by 85.8%. The FT Decoder enhances the average DSC by 1.7% and 7.87% for high- and low-resource settings, respectively, demonstrating its effectiveness for resource-constrained scenarios.

Sivasubramanian et al. [26] have proposed a lightweight AI architecture incorporating CNNs and attention mechanisms to classify a large-scale fetal ultrasound dataset comprising 12,000 images. Utilizing EfficientNet backbones (B0 and V2B0) pre-trained on ImageNet1k and enhanced with attention modules, the model achieved impressive results: 96.25% Top-1 accuracy, 99.80% Top-2 accuracy, and an F1-score of 95.76%, outperforming conventional models while using 40 times fewer parameters. Grad-CAM-based explainability further enabled clinical interpretability, making the model suitable for real-time deployment on edge devices to support prenatal diagnostics.

Ghabri et al. [28] have developed deep learning models using InceptionResNetV2, InceptionNet, DenseNet, MobileNet, and ResNet50 to classify fetal ultrasound images. By applying image cropping, augmentation, and data cleaning, they enhanced the quality of a public dataset. Their transfer learning approach achieved remarkable performance, with 99.78% accuracy, 99.77% F1-score, and 99.78% AUC, demonstrating its potential for deployment in resource-constrained healthcare settings.

Hasan et al. [32] have introduced an ensemble deep transfer learning framework for automatic fetal brain plane classification, combining U-Net for segmentation with a majority voting ensemble of three pre-trained classifiers. The model achieved 97.68% accuracy on a blind test set, exhibiting superior performance and robustness compared to existing methods in the domain.

Fiorentino et al. [34] have reviewed advancements in deep learning techniques for fetal ultrasound image analysis, focusing on standard plane detection, anatomical structure analysis, and biometric estimation. Evaluating 145 studies since 2017, the review highlights the strengths and limitations of existing methods while addressing challenges related to dataset availability, clinical relevance, and evaluation metrics.

In the past decade, researchers have explored various methods for analysing fetal images and detecting

abnormalities using different image processing techniques. Image representation and description are essential early steps in this process. Many researchers have contributed valuable work in this area. A summary of recent studies is provided in Table 1.

Dataset	Parameters	Methodology	Results	Limitations
Created Own Dataset	HC	SegNet	DSC = 97%, Accuracy = 98%	Small dataset; needs more diverse data. The model does not replace clinical diagnosis.
Created Own Dataset	GS	ResUNet	DSC = 97.8%, Accuracy = 98%, MIoU = 94.6%	Limited generalizability due to single-center data and manual plane selection; expand data sources and add real-time image analysis.
HC18 Grand Challenge dataset	HC	SegNet, GCN, and HRNet	DSC = 96%	It depends on segmentation accuracy; it lacks real-time validation and large-scale clinical evaluation.
HC18 Grand Challenge dataset	HC	Ensemble Transfer Learning	MIoU = 98.53%	Limited feasibility of deploying US machines in all settings; resolution constraints may affect measurement accuracy.
HC18 Grand Challenge dataset	HC	DR-ASPnet	DSC = 98.86%	Image quality issues due to maternal factors, anatomical variability, and limited access to 3D imaging in low-resource settings.
Created Own Dataset	HC, FL, AC	U-Net	Accuracy = 99.86%	Limited sample size; lower segmentation accuracy.
HC18 Grand Challenge dataset	HC	Deeply supervised attention-gated (DAG) V-Net	DSC = 97.63%	Limited sample size; lower segmentation accuracy in the first trimester; higher measurement error in the third trimester.
HC18 Grand Challenge dataset	HC	U-Net	DSC = 97.90%, MIoU = 97.81%	Challenges include noise in ultrasound images, variations in fetal head development, and the presence of overlapping sutures and blurred boundaries.
Created Own Dataset	Multi-Organ	AG-CNN	Accuracy = 94%	Needs a more diverse dataset, real-time deployment validation, and collaboration with clinicians for clinical translation.
HC18 Grand Challenge dataset	HC	Mobilenet V2	PA = 97.77%, DSC = 96.28%, MIoU = 92.87%	Limited data access in low-resource settings and difficulty in optimizing fine-tuning strategies.
Created Own Dataset	HC, FL, AC	U-Net, Deeplabv3+	HC: MIoU = 93%, FL: MIoU = 89%, AC: MIoU = 61%	Limited training data; not yet integrated with plane detection; needs validation through expert vs. novice comparison.

TABLE 1: Summary of recent work and limitations

AC, Abdominal Circumference; DSC, Dice Similarity Coefficient; FL, Femur Length; GS, Gestational Sac; HC, Head Circumference; MIoU, Mean Intersection over Union; PA, Pixel Accuracy

Although there has been progress in fetal biometry, a significant research gap remains in the trimester-based classification of fetal HC and FL. Existing methods often fail to capture the variations in these parameters across different trimesters, limiting the accuracy of fetal growth assessments. Notably, there is a lack of specialized datasets specifically designed for trimester-wise analysis of HC and FL. To address this gap, the present study introduces a comprehensive, multi-stage computational pipeline incorporating a deep learning framework tailored for both segmentation and trimester-based classification. This approach involves the creation of a dataset that systematically categorizes ultrasound images by trimester, thereby enhancing the model's ability to predict HC and FL. The proposed methodology, detailed in the following sections, is designed to overcome existing limitations through robust deep learning strategies and efficient data handling techniques.

Materials And Methods

Proposed method

The proposed methodology aims to perform a comprehensive analysis of fetal biometric parameters,

specifically HC and FL, through a trimester-aware multi-stage deep learning pipeline designed for ultrasound image processing. This framework integrates both the publicly available HC18 benchmark dataset and a newly curated custom dataset (1,426 HC and 1,404 FL images), allowing for a more diverse and representative evaluation of fetal development across various gestational stages. The pipeline consists of three primary components: pre-processing, segmentation, and classification. The proposed model is shown in Figure 2.

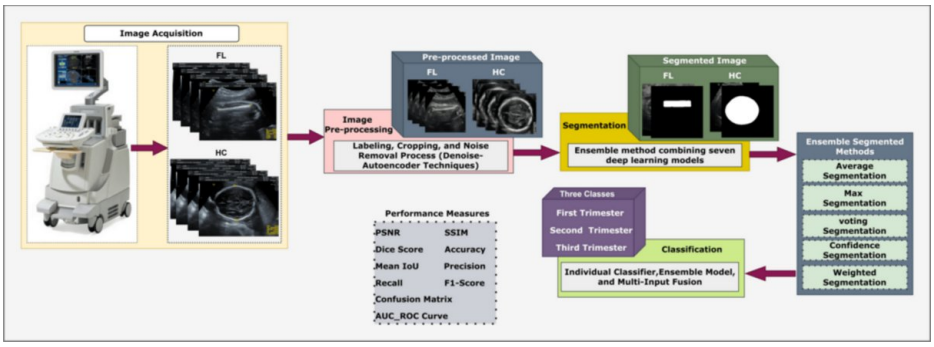

FIGURE 2: Block diagram of the proposed model

New Dataset Creation

In the present study, we have created a dataset comprising ultrasound images for HC and FL parameters. The dataset includes 1,426 HC images and 1,404 FL images, each categorized into FT, ST, and TT with guidance from medical professionals. The images were acquired using a VOLUSON P6 ultrasound machine at Metgud Hospital - Advanced Laparoscopy Centre and IVF, located in Belagavi, Karnataka, India. All images were captured in JPG format at an original resolution of 640 × 480 pixels and subsequently cropped to 300 × 300 pixels to focus on the region of interest. The dataset is provided without predefined training or testing splits, allowing researchers to partition the data according to their specific requirements. For annotation, an experienced sonographer manually labelled each image: the skull region is marked with an ellipse for HC measurement, and the femur region is enclosed within a rectangle for accurate FL assessment. Figure 3 shows sample ultrasound images from the dataset with important fetal biometric annotations. Image (a) displays the FL, and image (b) shows the HC. These examples help to show the clear structure and consistent labeling in the dataset.

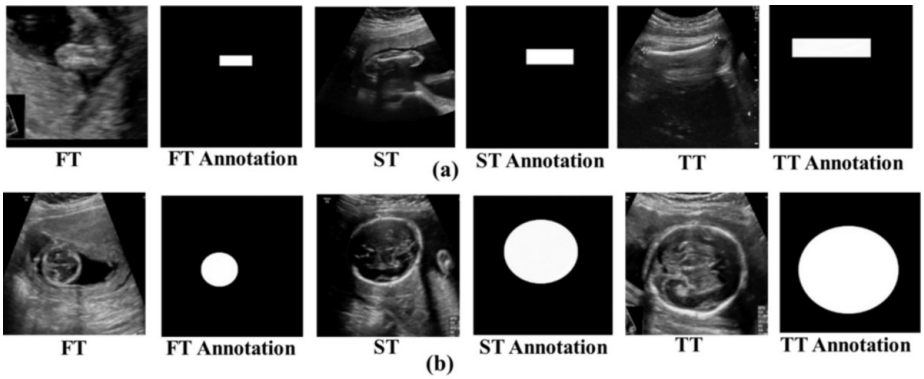

FIGURE 3: Sample ultrasound images from the created dataset illustrating annotated fetal biometric parameters: (a) Femur length and (b) head circumference

FT, First Trimester; ST, Second Trimester; TT, Third Trimester

The HC18 dataset presents several limitations that restrict trimester-specific fetal analysis. It does not include trimester labels for individual images and provides only HC measurements along with pixel size information [1]. Furthermore, the test set lacks HC values and segmentation annotations, making it difficult to perform detailed evaluations on unseen data. Although the dataset reports the total number of images per trimester, it does not specify the trimester category for each image. To address these limitations, a custom dataset is developed containing gestational age information for each fetus. This

addition enables accurate classification of images by trimester, allowing for more structured analysis of fetal development across different stages of pregnancy. The custom dataset also includes both HC and FL parameters, offering a larger number of samples than the HC18 dataset. These improvements enhance the statistical robustness of model training and validation. Table 2 presents the distribution of fetal ultrasound images by trimester for both the HC18 and the proposed custom dataset.

Datasets	First Trimester	Second Trimester	Third Trimester
Custom dataset-HC	148	380	898
Custom dataset-FL	112	395	897
HC18	165	693	141

TABLE 2: Trimester-wise distribution of fetal ultrasound images in the HC18 and custom datasets

HC, Head Circumference; FL, Femur Length

Limitations of Custom Dataset

This study aims to analyse fetal ultrasound images across different trimesters of pregnancy. A longitudinal dataset containing images of the same fetus in the FT, ST, and TT would be ideal for capturing continuous developmental changes. Such data could enhance model performance by providing consistent growth patterns and improving generalization. However, due to practical limitations, it is not possible to collect images from the same individuals across all trimesters. Data collection is constrained by limited patient availability, privacy concerns, and ethical regulations that restrict direct patient interaction at the host institution. As a result, the study relies solely on pre-recorded and ethically approved ultrasound data provided by hospital personnel.

Preprocessing Noise Removal

Ultrasound images used in this study were acquired using the VOLUSON P6 Ultrasound Machine, which provides high-resolution imaging suitable for visualizing fetal anatomical structures. The images were exported in JPG format and systematically organized into structured folders for efficient storage, retrieval, and further analysis. Before model training, a structured preprocessing pipeline is implemented to enhance image quality and highlight critical anatomical features [16]. This preprocessing is not merely for visual enhancement but directly supports downstream analysis by improving boundary clarity and structural detail, which are essential for accurate segmentation and biometric measurement. Initially, each image is inspected, and the region of interest is identified to focus on relevant fetal structures [31-34]. The images are then cropped to remove background artifacts, converted to grayscale to reduce computational complexity, and standardized to 300 × 300 pixels, ensuring uniformity and standardized input for the deep learning models [37-40].

As ultrasound images are likely to contain speckle noise, which reduces contrast and complicates the delineation of anatomical boundaries, a convolutional denoising autoencoder (CDAE) is employed for noise suppression. Unlike traditional denoising filters such as Gaussian or median filters, which often oversmooth and blur critical anatomical details, the CDAE effectively suppresses noise while preserving edges and fine structural information [41-43]. The CDAE follows an encoder-decoder architecture, consisting of two convolutional layers for encoding and two transposed convolutional layers for decoding. Rectified Linear Unit activations are used in all layers except the output layer, where a sigmoid function is applied to constrain pixel intensity values to the range (0,1). For training, grayscale images are synthetically corrupted with noise to simulate realistic multiplicative speckle noise. The dataset is divided into 90% training and 10% validation sets. The CDAE is trained using the Adam optimizer with a learning rate of 0.001 and mean squared error (MSE) as the loss function, with early stopping (patience = 10 epochs) to prevent overfitting. Denoising performance is quantitatively assessed using peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), both of which demonstrated significant improvements after denoising [44-49]. The trained CDAE is finally applied to the entire dataset, producing cleaner images with enhanced anatomical boundaries. This step is critical, as high-quality images substantially improve segmentation accuracy for delineating HC and FL regions. The block diagram of the CDAE architecture is presented in Figure 4.

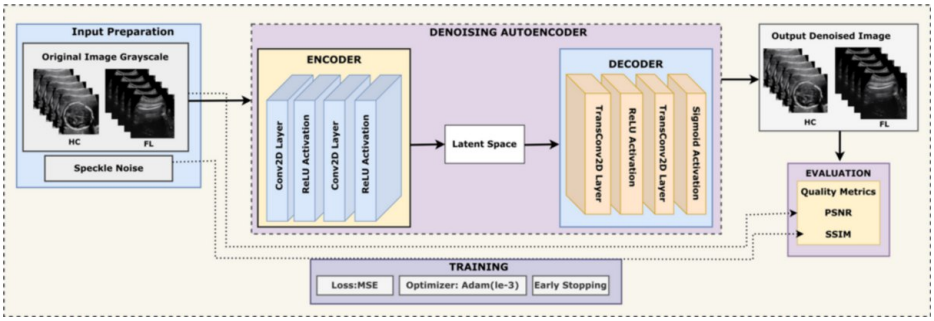

FIGURE 4: Block diagram of the convolutional denoising autoencoder used for speckle noise reduction

FL, Femur Length; HC, Head Circumference; MSE, Mean Squared Error; PSNR, Peak Signal-to-Noise Ratio; SSIM, Structural Similarity Index Measure

Figure 5 presents a comprehensive visual and statistical analysis of the proposed denoising method's performance on (a) the publicly available HC18 dataset and (b) a custom-created dataset with HC and (c) a custom-created dataset with FL parameters. The analysis demonstrates qualitative improvements through side-by-side image comparisons and quantitative validation via multiple metrics, including PSNR, SSIM, pixel correlation analysis, and intensity distribution statistics, providing complete evidence of effective noise reduction while preserving structural image details.

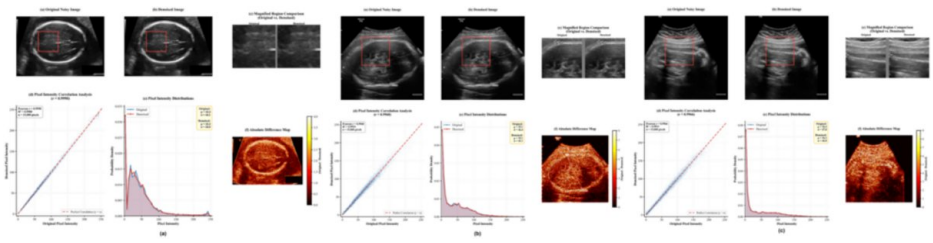

FIGURE 5: Comprehensive performance evaluation of the proposed denoising method across (a) the publicly available HC18 dataset with HC parameter, (b) the custom dataset with HC parameter, and (c) the custom dataset with FL parameter.

FL, Femur Length; HC, Head Circumference

Feature Extraction for Segmentation

Image segmentation in medical imaging is critical for delineating anatomical structures and identifying abnormalities, thereby supporting accurate diagnosis, treatment planning, and disease monitoring [34]. It serves as a fundamental component in automated medical image analysis pipelines. In this study, segmentation is performed to isolate HC and FL regions, enabling a quantitative understanding of anatomical variations across trimesters. This step directly influences biometric measurement accuracy and helps analyze how anatomical changes impact image analysis performance.

Denoised images and their corresponding ground truth masks are used as inputs to deep learning models for segmentation tasks. A total of seven architectures are evaluated: U-Net, featuring an encoder-decoder structure with skip connections to preserve spatial information; DeepLabV3+, integrating atrous spatial pyramid pooling for multi-scale feature extraction; DenseNet-U-Net, leveraging densely connected blocks for feature reuse; MFP-UNet, employing a multi-scale feature pyramid for detailed object segmentation; Attention U-Net, enhancing feature selection via attention mechanisms; MobileNet-U-Net, using depthwise separable convolutions for computational efficiency; and ResNet-U-Net, incorporating residual connections to facilitate deeper network training.

All models are implemented in PyTorch, with images resized to 256 × 256 pixels and normalized using ImageNet statistics. The dataset is split into 80% training and 20% validation sets. Training is conducted for 30 epochs using the Adam optimizer and a combined Dice-Binary Cross-Entropy (BCE) loss function. Segmentation performance is assessed using Dice, IoU, and pixel accuracy to evaluate the impact of the

proposed preprocessing and ensemble strategies on fetal image analysis [10,14]. Additional metrics such as precision, recall, and F1-score are also calculated. Automated checkpointing preserved the best-performing model based on the validation Dice coefficient.

To further improve segmentation accuracy, an ensemble learning approach is implemented to leverage the complementary strengths of all seven architectures. Five fusion strategies are explored: (i) average ensemble, combining predictions via arithmetic mean; (ii) weighted ensemble, assigning dynamic weights based on individual model performance; (iii) maximum ensemble, selecting the highest-confidence prediction per pixel; (iv) majority voting ensemble, applying threshold-based binary decisions; and (v) confidence-based ensemble, weighting predictions by model certainty scores [9,47-49]. The ensemble approach is validated on a separate 20% test split, and performance metrics are computed for each fusion strategy to identify the optimal combination for the segmentation task. Figure *6* illustrates the proposed ensemble deep learning framework for fetal image segmentation.

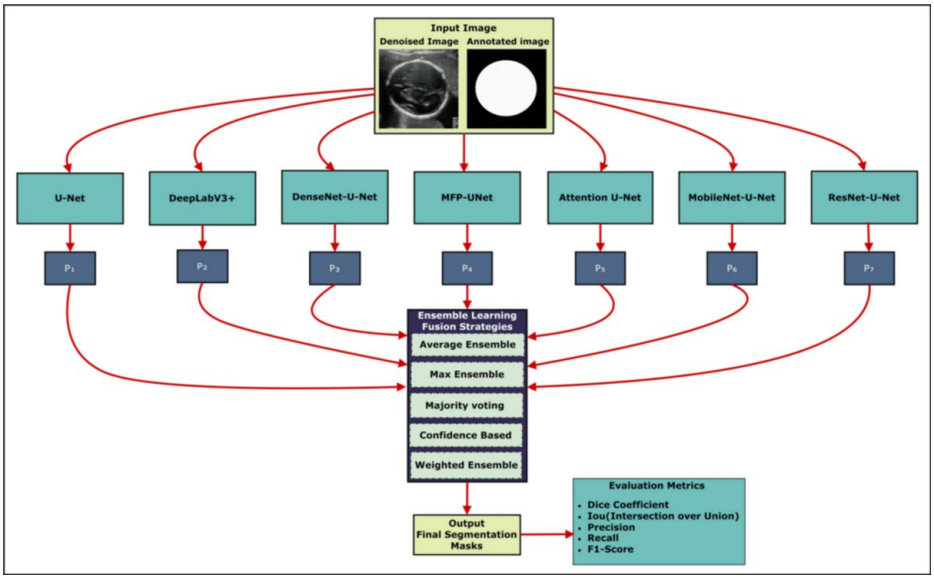

FIGURE 6: Ensemble deep learning framework for fetal image segmentation

Figure *7* presents the segmentation results of fetal biometric parameters using both the custom-created dataset and the publicly available HC18 dataset. Each subfigure illustrates the original ultrasound image, its corresponding annotated mask, and the outputs from seven different segmentation models: U-Net, DeepLabV3+, DenseNet-U-Net, MFP-U-Net, U-Net with Attention, MobileNet-U-Net, and ResNet-U-Net [9]. Subfigure (a) corresponds to HC segmentation from the custom dataset, subfigure (b) shows FL segmentation, and subfigure (c) displays HC segmentation from the HC18 dataset. The figure highlights the visual differences in segmentation quality across architectures, demonstrating the robustness and adaptability of each model across datasets and biometric targets.

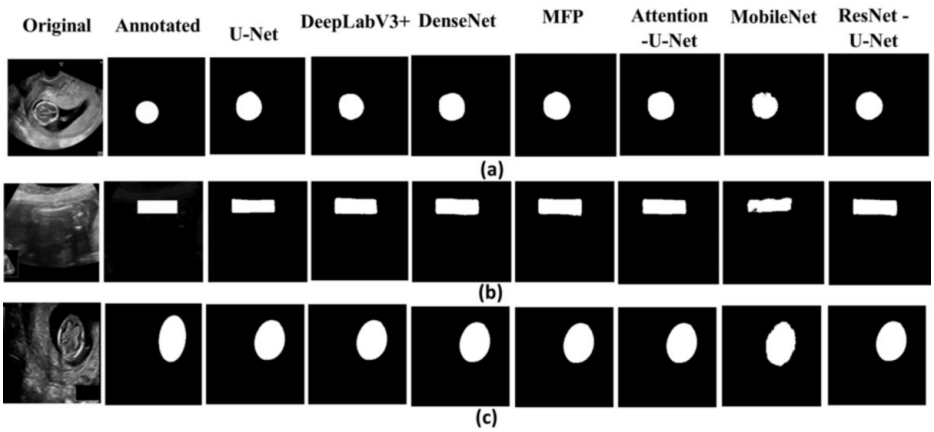

FIGURE 7: Segmentation results using seven deep learning models: (a) HC–custom dataset, (b) FL–custom dataset, and (c) HC–HC18 dataset, showing original image, ground truth annotation, and predicted mask.

FL, Femur Length; HC, Head Circumference

Figure *8* displays qualitative segmentation results of fetal biometric parameters using ensemble fusion strategies applied to ultrasound images from both the custom-created dataset and the publicly available HC18 dataset. Each row shows the original fetal ultrasound image, the ground truth mask, and the outputs from five ensemble strategies: arithmetic averaging, maximum probability, confidence-based fusion, majority voting, and weighted averaging. The ensemble-based segmented masks are presented both as binary masks and as color overlays on the original ultrasound images to facilitate visual comparison with the annotated ground truth. Subfigures (a), (b), and (c) correspond to different biometric parameters and datasets: (a) HC segmentation from the custom dataset, (b) FL segmentation from the custom dataset, and (c) HC segmentation from the HC18 dataset. The ensemble outputs show improved anatomical alignment, reduced segmentation variability, and smoother contours compared to individual model predictions. These visual improvements demonstrate the strength of the ensemble approach in combining multiple model outputs to enhance segmentation robustness, especially in the presence of ultrasound artifacts or low-contrast regions. Overall, the figure highlights the effectiveness of ensemble learning in accurately delineating fetal biometric structures across diverse datasets and imaging conditions.

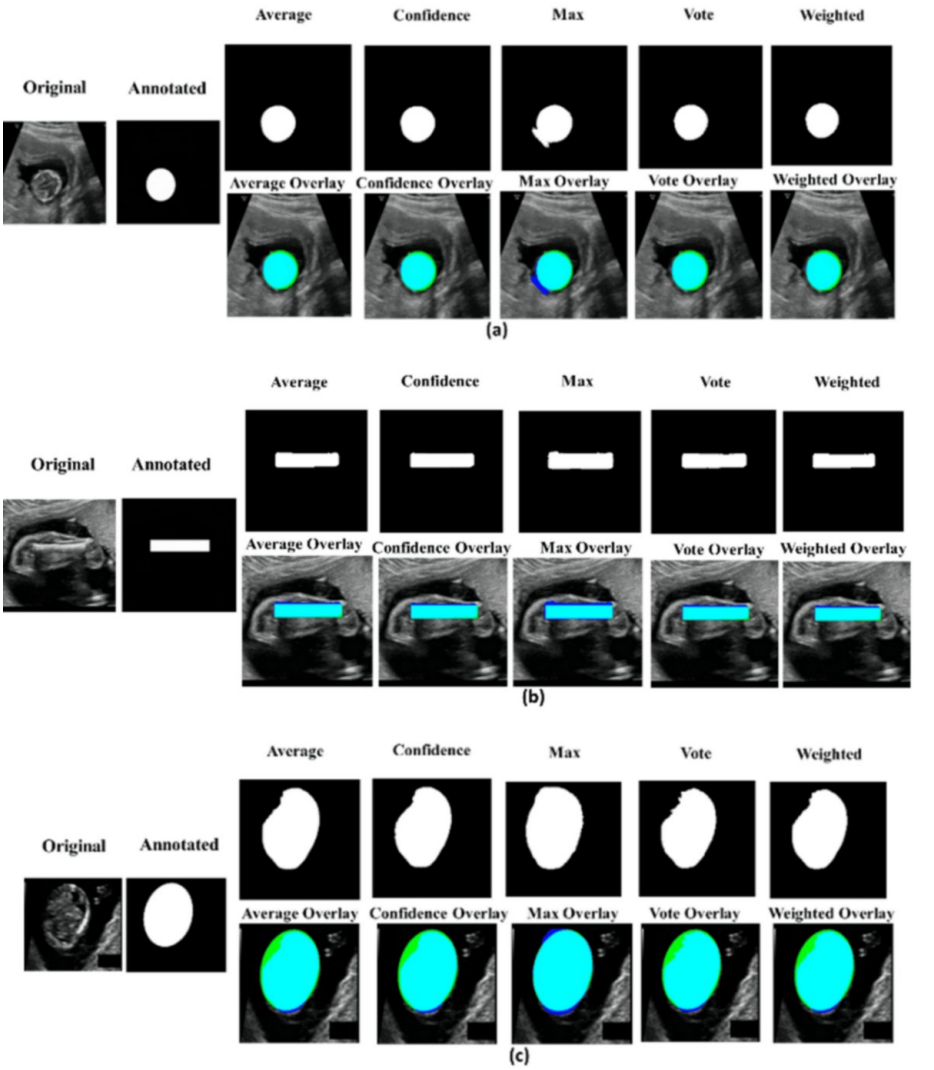

FIGURE 8: Ensemble segmentation results for fetal biometric parameters. Shown are original images, ground truth masks, and outputs from five ensemble methods for (a) the custom dataset HC parameter, (b) the custom dataset FL parameter, and (c) the publicly available HC18 dataset

FL, Femur Length; HC, Head Circumference

Classification

Image classification plays a pivotal role in trimester-specific fetal growth assessment, as accurate categorization depends on detecting subtle anatomical variations across different gestational stages. In this study, data augmentation techniques, including horizontal flipping and 90-degree clockwise rotation, were employed to improve model generalization. Each original ultrasound image was transformed into five variants (original, brightened, cropped, flipped, and rotated) [35-37]. Class labels are automatically extracted from filenames using regular expressions to identify the three trimester categories: FT, ST, and TT.

In this study, a multi-input classification strategy is proposed as the primary contribution because fetal growth assessment depends on the correlated nature of biometric parameters. HC and FL provide complementary information; therefore, combining them improves trimester classification consistency and reduces misclassification compared to single-parameter models. To enrich the feature space, the classification framework utilized segmentation-driven inputs, where five different segmentation strategies: maximum probability, averaging, confidence-based, voting, and weighted fusion are applied to generate multiple representations of each ultrasound image. These segmentation strategies emphasize different anatomical details, enabling the classification models to capture both fine-grained

morphological structures and broader contextual information. The overall classification pipeline is illustrated in Figure 9, which shows the stages of individual model training, multi-input feature fusion, and final prediction.

Individual Segmentation Models

As a baseline, five separate DenseNet121 models are trained, each processing images derived from one segmentation strategy. Transfer learning is applied using ImageNet pre-trained weights, where initially only the newly added classification layers are trained. The last 20 layers of the DenseNet121 base model are later unfrozen for fine-tuning to adapt the models to fetal ultrasound image characteristics. Each segmentation strategy contributed uniquely to feature learning. The maximum probability segmentation model produced sharp and deterministic anatomical boundaries, allowing the model to learn confident morphological features. The averaging-based model captured smooth transitions, making it sensitive to subtle tissue variations that are critical for trimester differentiation. The confidence-based model focused on high-certainty regions, improving robustness to noise and poor-quality scans. The voting-based model reduced individual model bias by emphasizing consensus anatomical structures, while the weighted segmentation model learned hierarchical feature relevance by assigning higher importance to more reliable predictions. These individual models provided valuable baseline comparisons for evaluating the proposed approach.

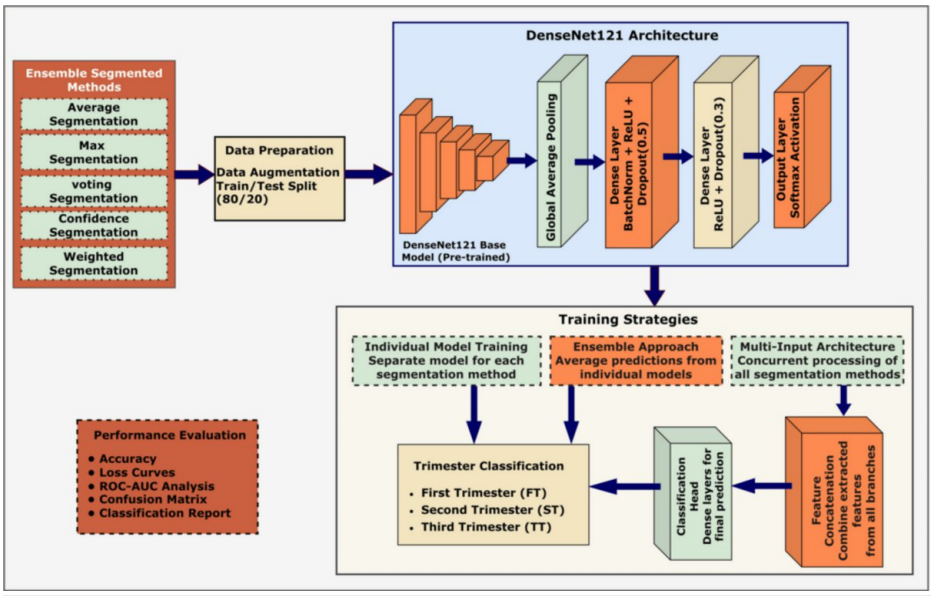

FIGURE 9: DenseNet121-based multi-segmentation input classification framework

Ensemble Learning Approach

To exploit the complementary strengths of the baseline models, a prediction-level ensemble learning approach is implemented. During testing, the softmax probability vectors produced by each model are combined using arithmetic averaging [9,39-41]. This fusion strategy balanced sharp morphological details from the maximum probability model, smooth tissue transitions from the averaging model, high-certainty features from the confidence-based model, stable consensus structures from the voting model, and optimized feature weighting from the weighted fusion model. Although the ensemble strategy enhanced prediction stability and reduced bias compared to individual models, it is limited because the fusion occurred only at the decision level, preventing the network from learning deeper inter-feature correlations [45-49]. This process can be described by Equation (1).

$$P_{\text{ensemble}}(c \mid x) = \frac{1}{N} \sum_{i=1}^{N} P_i(c \mid x) \quad (1)$$

where, $P_{\text{ensemble}}(c|x)$ represents the ensemble-predicted probability that input x belongs to class c, N is the number of models, and $P_i(c|x)$ is the probability assigned by the i^{th} model. The arithmetic mean ensures equal contribution from each model, enhancing prediction robustness.

Multi-Input Neural Network Architecture

To address this limitation, a multi-input DenseNet121 architecture is developed as the proposed classification framework and represents the core innovation of this study. Unlike the ensemble approach, which combines model predictions after they are generated, the multi-input network integrates complementary features during the feature extraction stage, allowing it to learn optimal correlations across different segmentation variants. The architecture consists of five parallel DenseNet121 branches, each independently processing a distinct segmentation variant. Features extracted from these branches are concatenated into a unified feature vector and passed through fully connected layers with Batch Normalization and Dropout for final trimester classification [46-49]. This integrated design allows the network to automatically determine the most informative segmentation-derived features, effectively combining sharp morphological indications, smooth contextual transitions, and reliable compromise information.

The multi-input architecture demonstrated superior performance compared to both single-input and ensemble strategies. As discussed in Experimental Results, the proposed approach significantly improved trimester classification accuracy and F1-score while reducing misclassification errors, confirming its effectiveness for trimester-specific fetal analysis [48-51]. Furthermore, feature visualization using Principal Component Analysis and t-distributed Stochastic Neighbor Embedding revealed well-separated clusters for the FT, ST, and TT classes across all datasets, further validating the discriminative capability and generalization power of the proposed architecture (Figure *10*).

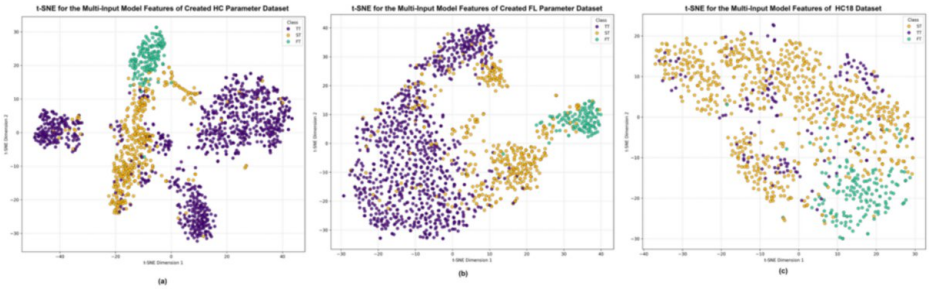

FIGURE 10: t-SNE visualizations of features from the multi-input DenseNet121 model on three datasets: (a) The custom dataset HC parameter. (b) The custom dataset FL parameter. (c) The publicly available HC18 dataset

FL, Femur Length; HC, Head Circumference; t-SNE, t-distributed Stochastic Neighbor Embedding

Dataset Splits Across Modeling Objectives

In this study, different dataset splits were applied across modeling objectives to ensure appropriate training and evaluation. For the noise reduction autoencoder, a 90:10 (training: validation) split was used to fine-tune the model and prevent overfitting. For the segmentation objective, an 80:20 (training: validation) split was employed, where the validation set guided model selection and supported ensembling of multiple models. For the classification objective, the dataset was divided into 80:20 (training: testing), with the testing set strictly reserved for final evaluation. Data augmentation techniques (brightness adjustment, cropping, flipping, and rotation) were applied to the training data to enhance generalization. In addition, controlled augmentations were applied to the testing data in the form of test-time augmentation to evaluate robustness under varying imaging conditions. Importantly, the testing data remained unseen during training and was used solely for reporting the final classification performance.

Results And Discussion

Dataset and evaluation

The HC18 publicly available dataset comprises 1,334 fetal head ultrasound images (800 × 540 pixels), including 999 for training and 335 for testing. Pixel sizes range from 0.052 to 0.326 mm. Training data include HC measurements and pixel sizes, while only pixel sizes are available for the test set. The images were collected from 551 pregnant women across all three trimesters at Radboud University Medical Center, Netherlands, using Voluson E8 and Voluson 730 ultrasound machines. Certified sonographers annotated the skull regions with ellipses, following ethical approval (CMO Arnhem-Nijmegen) and the Declaration of Helsinki [1]. Gestational age classification was performed based on HC values using

threshold ranges derived from standard fetal growth charts [2,7-9]. Instances with HC values outside reference ranges are labelled as “Abnormal” or “<8 weeks.” Trimesters are defined as follows: the first (up to 13 weeks), the second (14-26 weeks), and the third (27-40 weeks). All annotations and labels are compiled in a structured CSV file for further analysis [1-3].

Implementation details

The training configurations and hyperparameters adopted for the denoising autoencoder, ensemble segmentation, and classification tasks are summarized in Table 3. These include input image sizes, learning rates, optimizers, batch sizes, and evaluation metrics tailored to the requirements of each task. For the denoising autoencoder, evaluation emphasizes structural quality using PSNR and SSIM, whereas classification performance is measured with accuracy, precision, recall, F1-score, and AUC [31-34]. The ensemble segmentation models further integrate multiple fusion strategies, such as averaging, voting, and confidence-based methods, to enhance prediction robustness and reliability.

Hyperparameter	Value/Setting	Description
Hyperparameter Settings for Denoising Autoencoder		
Input Image Size	300 × 300	Resized grayscale input dimensions
Batch Size	8	Number of samples per training batch
Learning Rate	1×10^{-3}	Initial learning rate for the Adam optimizer
Optimizer	Adam	Optimization algorithm used for training
Max Epochs	150	Maximum number of training epochs
Early Stopping Patience	10	Stop training if validation loss does not improve
Evaluation Metrics	PSNR, SSIM	Used to assess image quality after denoising
Hyperparameter Settings for the Ensemble Segmentation Approach		
Image Size	256 × 256	Input size for resizing images and masks
Batch Size	4	Number of samples processed per batch
Model Paths	Multiple .pth files	Pretrained weights used for ensemble (UNet, ResNetUNet, etc.)
Ensemble Methods	['average', 'weighted', 'max', 'vote', 'confidence']	Fusion strategies to combine predictions from multiple models
Save All Predictions	True	Saves predictions for all images in both validation and test sets
Hyperparameter Settings for Classification approach		
Input Image Size	224 × 224 × 3	Standard input size for DenseNet121
Optimizer	Adam	Adaptive learning rate optimization algorithm
Fine-Tuning Learning Rate	1×10^{-5}	Lower rate to update base layers
Loss Function	Categorical Cross-Entropy	For multi-class classification
Batch Size (single model)	16	Number of samples per training step
Batch Size (multi-input)	8	Reduced due to higher memory requirements
Epochs (single model)	30	Number of complete passes through the training data
Epochs (multi-input)	40	More training is required for complex architecture
Dropout Rate	0.5 (1st layer), 0.3 (2nd layer)	Prevents overfitting by randomly deactivating neurons
Dense Layers	512 → 256 units	Fully connected layers for final classification
Learning Rate Scheduler	ReduceLROnPlateau	Reduces learning rate when validation loss plateaus
Early Stopping Patience	10 epochs	Stops training if no improvement is seen in validation loss
Activation Function	Softmax	For multi-class probability output
Evaluation Metrics	Accuracy, Precision, Recall, F1, AUC	Comprehensive performance measurement

TABLE 3: Hyperparameter settings for denoising autoencoder, ensemble segmentation, and classification approaches

AUC, Area Under the Curve; PSNR, Peak Signal-to-Noise Ratio; SSIM, Structural Similarity Index Measure

The choice of hyperparameters, including input image size, batch size, and learning rates, is tailored to the requirements of each task. For the denoising autoencoder, a larger input size of 300 × 300 is selected to retain fine-grained structural details critical for preserving anatomical information during noise reduction. In the segmentation task, images are resized to 256 × 256 to strike a balance between capturing

spatial resolution and ensuring feasible training with multiple ensemble models under GPU memory constraints. For the classification task, an input size of 224 × 224 × 3 is adopted to ensure compatibility with standard pretrained convolutional neural network (CNN) backbones (e.g., DenseNet121), enabling effective transfer learning. Differences in batch size, epochs, and dropout rates across tasks reflect the varying complexity and computational needs of each pipeline. Furthermore, the evaluation metrics are also task-specific: PSNR and SSIM were used to assess the perceptual quality of denoised images, Dice and IoU measured spatial overlap for segmentation, and accuracy, precision, recall, F1, and AUC captured discriminative ability for classification. Overall, hyperparameter and metric choices are carefully optimized to balance performance, generalization, and computational efficiency for their respective objectives.

Quantitative Performance Measures

Speckle Noise Modelling: Speckle noise is simulated by the following multiplicative model and is defined by Equation (2):

$$I_{\text{noise}}(x,y) = I_{\text{clean}}(x,y) + I_{\text{clean}}(x,y) \cdot N(x,y) \quad (2)$$

where, $I_{\text{noise}}(x,y)$ is the noisy fetal ultrasound image at pixel coordinates (*x*,*y*), $I_{\text{clean}}(x,y)$ is the corresponding clean image, and *N*(*x*,*y*) represents multiplicative noise at that pixel location. This formulation simulates realistic ultrasound artifacts by adding noise proportional to pixel intensity, aiding in model robustness evaluation during fetal image segmentation and classification experiments.

Loss Function: The MSE between the reconstructed image and the clean ground truth image $\tilde{x}$ as the objective function is represented by Equation (3):

$$L_{\text{MSE}} = \frac{1}{n}\sum_{i=0}^{n}(x_i - \tilde{x}_i)^2 \quad (3)$$

where, L_{MSE} denotes the mean squared error loss, n is the total number of pixels, x_i represents the true pixel intensity, and $\tilde{x}$ is the predicted pixel intensity at the ith position. This loss function penalizes large deviations between predicted and ground truth images, making it suitable for evaluating reconstruction quality in fetal ultrasound image segmentation and enhancement tasks.

PSNR is the standard image quality metric and is defined by Equation (4):

$$\text{PSNR} = 20 \cdot \log_{10}\left(\frac{\text{MAX}_I}{\sqrt{\text{MSE}}}\right) \quad (4)$$

where, PSNR measures the quality of a reconstructed or denoised fetal ultrasound image. MAX $_I$ denotes the maximum possible pixel intensity value in the image, and MSE is the mean squared error between the ground truth and reconstructed images. Higher PSNR values indicate better image quality, which is essential for preserving diagnostic details in fetal imaging.

SSIM is the standard image quality metric and is defined by Equation (5):

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

where, SSIM quantifies the perceptual similarity between two fetal ultrasound images x and y. Here, μ_x and μ_y are the mean intensities, and σ_y^2 are the variances, and σ_{xy} is the covariance between the images. C_1 and C_2 are small constants that stabilize the division. Higher SSIM values indicate better structural preservation, which is critical for maintaining anatomical details in medical imaging.

Hybrid Loss Function: The hybrid loss function that merges BCE loss with Dice loss is given by Equation (6):

$$L_{\text{Hybrid}} = L_{\text{BCE}} + L_{\text{Dice}} \quad (6)$$

where, L_{BCE} is the Binary Cross-Entropy loss defined by Equation (7):

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \quad (7)$$

where, L_{BCE} is the Binary Cross-Entropy loss, *N* denotes the total number of pixels, y_i is the ground truth label for the *i*th pixel (1 for foreground, 0 for background), and $\hat{y}_i$ is the predicted probability of that pixel belonging to the foreground. This loss function measures the divergence between predicted probabilities and actual binary labels, making it well-suited for fetal ultrasound segmentation tasks where accurate boundary delineation is critical.

L_{Dice} is the Dice loss, derived from the Dice coefficient, defined by Equation (8)

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \quad (8)$$

where, L_{Dice} is the Dice loss, *N* is the total number of pixels, y_i is the ground truth binary label for the *i*th pixel, and $\hat{y}_i$ is the predicted probability for that pixel. The constant ϵ is a small smoothing term to avoid division by zero. Dice loss is widely used in medical image segmentation, including fetal ultrasound, as it directly optimizes spatial overlap between predicted and ground truth regions.

Dice Coefficient: The Dice coefficient is a measure of overlap between two samples. It ranges from 0 (no overlap) to 1 (perfect overlap) and is defined by Equation (9):

$$\text{Dice Coefficient} = \frac{2 \times (A \cap B)}{A + B} \quad (9)$$

where, A is the set of pixels belonging to the predicted segmentation mask, and B is the set of pixels in the ground truth mask. A∩B represents the intersection between the two sets. The Dice coefficient measures the spatial overlap between prediction and ground truth, with values ranging from 0 (no overlap) to 1 (perfect overlap). In fetal ultrasound segmentation, a higher Dice coefficient indicates more accurate delineation of anatomical structures.

Mean Intersection over Union: IoU evaluates the extent of overlap between the actual and predicted areas, and it is determined using Equation (10):

$$\text{IoU} = \frac{A \cap B}{A \cup B} \quad (10)$$

where, *A* is the set of pixels in the predicted segmentation mask, *B* is the set of pixels in the ground truth mask, *A*∩*B* represents their intersection, and *A*∪*B* denotes their union. IoU, also known as the Jaccard index, quantifies the proportion of overlapping area between prediction and ground truth. In fetal ultrasound segmentation, a higher IoU score reflects greater accuracy in identifying and delineating relevant anatomical regions.

Performance Metrics

Precision: Precision measures the proportion of true positives out of the total predicted positives and is given by Equation (11):

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

where, TP (true positives) represents the number of fetal ultrasound images correctly classified as belonging to a specific gestational category, and FP (false positives) represents the number of images incorrectly classified as belonging to that category. Precision reflects the proportion of correctly identified cases among all cases predicted as positive, which is critical in reducing false alarms in medical diagnostics.

Recall (R): Recall measures the proportion of true positives out of the actual positives and is given by Equation (12):

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

where, TP (true positives) denotes the number of fetal ultrasound images correctly classified into the target gestational category, and FN (false negatives) represents the number of images that belong to the category but were misclassified. Recall measures the model's ability to correctly identify all relevant cases, which is crucial in minimizing missed diagnoses in medical imaging.

F1-Score (F1): F1-Score is the harmonic mean of Precision and Recall, which provides a balance between the two metrics, and is given by Equation (13):

$$F_1\ \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

where, Precision is the proportion of correctly identified positive fetal ultrasound images among all predicted positives, and Recall is the proportion of correctly identified positives among all actual positives. The F_1 Score is the harmonic mean of Precision and Recall, providing a balanced measure of accuracy that is particularly valuable in medical imaging when both false positives and false negatives must be minimized.

Ensemble Prediction Methods

Average: The average ensemble method computes the final prediction by taking the mean of the outputs from all *N* models, treating each model equally, and is given by Equation (14):

$$\hat{Y}_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} \hat{Y}_i \quad (14)$$

where, $\hat{Y}_{\text{avg}}$ represents the average predicted output across an ensemble of *N* models, and $\hat{Y}_i$ denotes the prediction from the *i*th model for a given fetal ultrasound image. This averaging approach reduces prediction variance and enhances robustness in both classification and segmentation tasks for fetal imaging.

Weighted: In the weighted ensemble method, each model's prediction is multiplied by a weight w_i, typically based on its validation performance, and the final output is the weighted sum of all models. It is determined using Equation (15):

$$\hat{Y}_{\text{weighted}} = \sum_{i=1}^{N} w_i \hat{Y}_i \quad \text{with} \quad \sum_{i=1}^{N} w_i = 1 \quad (15)$$

where, $\hat{Y}_{\text{weighted}}$ is the weighted ensemble prediction for a given fetal ultrasound image, w_i denotes the weight assigned to the *i*th model's prediction $\hat{Y}_i$, and *N* is the total number of models. The weights w_i sum to 1, ensuring proportional contribution from each model. This method allows higher-performing models to influence the final decision more strongly, improving accuracy in fetal image classification and segmentation.

Max: The max ensemble selects the highest prediction score among all models for each pixel, emphasizing the most confident prediction, and is determined using Equation (16):

$$\hat{Y}_{\text{max}} = \max_{i=1}^{N} \hat{Y}_i \quad (16)$$

where, $\hat{Y}_{\text{max}}$ represents the final ensemble prediction obtained by selecting the maximum predicted probability among *N* models for a given fetal ultrasound image. This approach prioritizes the most confident prediction, which can be beneficial in clinical decision-making where high-certainty classifications are preferred.

Majority Voting: Majority voting assigns the final label for each pixel based on the most frequently predicted class among all models, and is given by Equation (17):

$$\hat{Y}_{\text{vote}} = \text{mode}\left(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_N\right) \quad (17)$$

where, $\hat{Y}_{\text{vote}}$ denotes the ensemble prediction obtained through majority voting among *N* model

outputs $\left(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_N\right)$ for a given fetal ultrasound image. This method selects the class label predicted by the majority of models, reducing the influence of outlier predictions and improving overall classification reliability in medical imaging.

Confidence-based: The confidence-based ensemble dynamically weights model predictions based on softmax-normalized confidence scores s_i, giving more influence to more reliable models, and is given by Equation (18):

$$\hat{Y}_{\text{conf}} = \sum_{i=1}^{N} c_i \hat{Y}_i \quad \text{where} \quad c_i = \frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}} \qquad (18)$$

where, $\hat{Y}_{\text{conf}}$ is the confidence-weighted ensemble prediction for a given fetal ultrasound image, c_i represents the normalized confidence score of the i^{th} model, and $\hat{Y}_i$ is the corresponding prediction. The confidence score c_i is computed using a softmax function applied to the raw confidence values s_i, ensuring that all c_i sum to 1. This method assigns greater influence to models with higher prediction confidence, potentially improving diagnostic accuracy in fetal image classification and segmentation.

The mathematical formulations underlying loss functions and evaluation metrics used across the models are detailed in Equations (1)-(18). These equations define the models that learn from data, and performance is quantitatively assessed throughout the training and validation phases [36,37,40-43].

Experimental setup

The segmentation experiments were conducted on a local workstation equipped with an Intel [CPU model] processor, 16 GB RAM, and an NVIDIA GeForce RTX 2050 GPU, operating on Windows 11 with CUDA version 12.8 and NVIDIA driver version 571.96. Classification experiments were performed on Google Colab utilizing an NVIDIA A100 GPU with 40 GB VRAM. Model development and execution were carried out using Anaconda Navigator as the integrated environment, with interactive computing facilitated through Jupyter Notebook (version 7.0.8). All algorithms were implemented in Python, employing libraries such as TensorFlow, Keras, scikit-learn, pandas, NumPy, and matplotlib.

Ablation study

To systematically assess the effect of ensemble complexity and fusion strategies on segmentation and classification performance, we conducted an extensive ablation study across three datasets. For segmentation, the ensemble size is varied from three to seven models, employing five fusion techniques: averaging, weighted averaging, maximum selection, voting, and confidence-based fusion. For classification, we evaluated the influence of data augmentation, model architectures (CNN, DenseNet121, ConvNext-Base, MobileNetV2), and ensemble strategies using seven evaluation methods. These included five individual fusion approaches, their ensemble combination, and a multi-input fusion method [31-34,46-49]. This analysis offers key insights into the trade-offs between model diversity, ensemble size, and performance, facilitating optimal configuration for clinical deployment.

Segmentation Performance Analysis

The comprehensive ablation study on the FL parameter dataset demonstrates progressive performance improvements with increasing ensemble complexity, where the three-model ensemble (U-Net+MFP-UNet+Attention U-Net) establishes a strong baseline with Dice coefficient of 91.40%, MIoU of 84.96%, and an accuracy of 98.85%. The systematic addition of models shows consistent enhancement, with the 4-model ensemble (adding DeepLabV3+) achieving a Dice coefficient of 91.57% and MIoU of 85.75%, while the 5-model ensemble (incorporating MobileNet-U-Net and ResNet-U-Net) maintains similar performance at 91.53% Dice coefficient and 85.18% MIoU. The six-model ensemble demonstrates marginal improvements with Dice coefficient of 91.66% and MIoU of 85.40%, ultimately culminating in the seven-model ensemble achieving optimal performance with Dice coefficient of 92.84%, MIoU of 86.71%, and maintaining high accuracy of 98.92%. Across all ensemble configurations, the averaging and weighted averaging strategies consistently outperform the maximum selection approach, while confidence-based and voting strategies demonstrate competitive performance, indicating robust ensemble behavior across different fusion methodologies, as detailed in Table *4*.

Ensemble Segmentation Method	Dice Coefficient	MIoU	Accuracy	Precision	Recall	Specificity	F1-Score
U-Net+MFP-UNet+Attention U-Net							
Average	91.40	84.96	98.85	93.13	91.84	99.44	91.40
Weighted	91.40	84.96	98.85	93.14	91.83	99.45	91.40
Max	90.43	85.29	98.66	88.63	94.59	99.00	90.43
Vote	91.33	84.84	98.84	92.95	91.88	99.43	91.33
Confidence	91.42	85.00	98.85	93.25	91.77	99.46	91.42
U-Net+DeepLabV3pluse+DenseNet-U-Net+MFP-UNet							
Average	91.57	85.75	98.92	93.08	92.79	99.43	91.57
Weighted	91.57	85.75	98.92	93.08	92.79	99.43	91.57
Max	90.16	82.84	98.60	87.30	95.53	98.86	90.16
Vote	91.70	85.46	98.90	94.05	91.51	99.52	91.70
Confidence	91.57	85.75	98.92	93.08	92.79	99.43	91.57
U-Net+MFP-UNet+Attention U-Net+MobileNet-U-Net+ResNet-U-Net							
Average	91.53	85.18	98.87	93.63	91.61	99.49	91.53
Weighted	91.53	85.19	98.87	93.63	91.61	99.49	91.53
Max	90.37	81.53	98.35	84.87	95.45	98.59	90.37
Vote	91.48	85.10	98.86	93.48	91.66	99.48	91.48
Confidence	91.53	85.19	98.87	93.63	91.61	99.49	91.53
U-Net+DeepLabV3pluse+DenseNet-U-Net+MFP-UNet+Attention U-Net+ResNet-U-Net							
Average	91.66	85.40	98.89	93.58	91.91	99.48	91.66
Weighted	82.66	85.40	98.89	93.58	91.91	99.48	82.66
Max	89.54	81.16	98.30	84.11	95.91	98.50	89.54
Vote	91.51	85.14	98.87	94.21	91.01	99.54	91.51
Confidence	91.67	85.42	98.89	93.65	91.86	99.49	91.67
U-Net+DeepLabV3pluse+DenseNet-U-Net+MFP-UNet+Attention U-Net+MobileNet-U-Net+ResNet-U-Net							
Average	92.84	86.71	98.92	93.66	92.17	99.49	92.84
Weighted	92.84	86.71	98.92	93.66	92.17	99.49	92.84
Max	89.43	80.99	98.28	83.78	96.01	98.46	89.43
Vote	92.80	86.63	98.91	93.51	92.23	99.47	92.80
Confidence	92.84	86.71	98.92	93.73	92.10	99.49	92.84

TABLE 4: Ensemble segmentation performance on the femur length parameter dataset with varying model combinations and fusion strategies

MIoU, Mean Intersection over Union

The HC parameter dataset exhibits superior performance characteristics compared to the FL parameter dataset, with the three-model ensemble achieving substantially higher baseline metrics, including Dice coefficient of 96.41%, an MIoU of 93.08%, and an accuracy of 98.07%. The progressive ensemble expansion demonstrates consistent improvements, with the four-model ensemble reaching a Dice coefficient of 96.65% and MIoU of 93.53%, while the five-model ensemble maintains competitive performance at 96.38% Dice coefficient and 93.02% Mean IoU. The six-model ensemble shows slight enhancement with

Dice coefficient of 96.43% and MIoU of 93.12%, ultimately achieving peak performance with the seven-model ensemble at a Dice coefficient of 97.81%, an MIoU of 94.27%, and an accuracy of 98.11%. The dataset demonstrates excellent precision scores exceeding 95% across all ensemble configurations, with particularly notable specificity values exceeding 98.5%, indicating superior true negative identification capabilities. The ensemble strategy analysis reveals consistent behaviour with averaging approaches outperforming maximum selection, while confidence-based strategies demonstrate the highest precision values, suggesting effective uncertainty quantification for this dataset, as detailed in Table *5*.

Ensemble Segmentation Method	Dice Coefficient	MIoU	Accuracy	Precision	Recall	Specificity	F1-Score
U-Net+MFP-UNet+Attention U-Net							
Average	96.41	93.08	98.07	95.94	96.91	98.46	96.41
Weighted	96.41	93.08	98.07	95.93	96.91	98.46	96.41
Max	96.18	92.65	97.92	94.66	97.76	97.94	96.18
Vote	96.39	93.04	98.06	95.90	96.89	98.45	96.39
Confidence	96.41	93.08	98.07	95.94	96.90	98.46	96.41
U-Net+DeepLabV3pluse + DenseNetU-Net+MFP-UNet							
Average	96.65	93.53	98.15	96.48	96.48	98.84	96.65
Weighted	96.65	93.53	98.15	96.48	96.48	98.84	96.65
Max	96.27	92.84	97.92	94.56	98.06	97.84	96.27
Vote	96.56	93.36	98.11	96.82	96.31	98.77	96.56
Confidence	96.66	93.54	98.16	96.50	96.84	98.64	96.66
U-Net+MFP-UNet+Attention U-Net+ MobileNet-U-Net +ResNet-U-Net							
Average	96.38	93.02	98.06	96.19	96.58	98.56	96.38
Weighted	96.38	93.02	98.06	96.19	96.58	98.56	96.38
Max	95.55	91.51	97.57	93.17	98.09	97.35	95.55
Vote	96.37	93.01	98.06	96.15	96.61	98.55	96.37
Confidence	96.36	92.28	98.05	96.20	96.54	98.57	96.36
U-Net+DeepLabV3pluse+ DenseNet-U-Net+ MFP-UNet+ Attention U-Net+ResNet-U-Net							
Average	96.43	93.12	98.09	96.23	96.61	98.59	96.43
Weighted	96.43	93.12	98.09	96.23	96.61	98.59	96.43
Max	95.50	91.42	97.54	92.89	98.29	97.22	95.50
Vote	96.37	93.10	98.08	96.27	96.58	98.60	96.37
Confidence	96.42	93.10	98.08	96.27	96.58	98.60	96.42
U-Net+DeepLabV3pluse+ DenseNet-U-Net+ MFP-UNet+ Attention U-Net+MobileNet-U-Net+ ResNet-U-Net							
Average	97.81	94.27	98.11	97.32	96.72	98.60	97.81
Weighted	97.81	94.27	98.11	97.32	96.73	98.60	97.81
Max	96.45	91.34	97.49	92.66	98.46	97.09	96.45
Vote	97.79	93.24	98.10	96.29	96.72	98.59	97.79
Confidence	97.80	94.26	98.11	97.32	96.70	98.60	97.80

TABLE 5: Ensemble segmentation performance on the HC parameter dataset with varying model combinations and fusion strategies.

MIoU, Mean Intersection over Union

The HC18 public dataset demonstrates the highest overall performance across all evaluation metrics, with the three-model ensemble establishing an exceptionally strong baseline of 97.82% Dice coefficient, 95.73% MIoU, and 98.71% accuracy. The ensemble expansion maintains consistently high performance, with the four-model ensemble achieving 97.91% Dice coefficient and 95.91% MIoU, while the five-model ensemble demonstrates 97.78% Dice coefficient and 95.66% MIoU. The six-model ensemble shows competitive performance at 97.80% Dice coefficient and 97.70% MIoU, ultimately reaching peak

performance with the seven-model ensemble at 98.74% Dice coefficient, 95.79% MIoU, and 98.73% accuracy. This dataset exhibits outstanding specificity scores exceeding 99% across all ensemble configurations, demonstrating exceptional true negative identification capabilities crucial for clinical applications. The precision and recall metrics consistently exceed 97% and 96% respectively, indicating balanced sensitivity and specificity trade-offs. The ensemble strategy analysis reveals that averaging and confidence-based approaches achieve nearly identical performance, suggesting well-calibrated uncertainty estimates, while the maximum strategy shows the most significant performance degradation compared to other datasets, emphasizing the importance of appropriate ensemble fusion strategies for high-performing baseline models, as detailed in Table *6*.

Ensemble Segmentation Method	Dice Coefficient	MIoU	Accuracy	Precision	Recall	Specificity	F1-Score
U-Net+MFP-UNet+Attention U-Net							
Average	97.82	95.73	98.71	97.92	97.73	99.09	97.82
Weighted	97.82	95.73	98.71	97.92	97.73	99.09	97.82
Max	97.57	95.27	98.55	96.59	98.59	98.50	97.57
Vote	97.79	95.68	98.69	97.88	97.70	99.08	97.79
Confidence	97.82	95.74	98.71	97.93	97.72	99.10	97.82
U-Net+DeepLabV3pluse+DenseNet-U-Net+MFP-UNet							
Average	97.91	95.91	98.76	98.06	97.77	99.15	97.91
Weighted	97.91	95.91	98.76	98.06	97.77	99.15	97.91
Max	97.41	94.97	98.44	95.97	98.92	98.20	97.41
Vote	97.82	95.74	98.72	98.41	97.25	99.31	97.82
Confidence	97.91	95.91	98.76	98.05	97.78	99.15	97.91
U-Net+MFP-UNet+Attention U-Net+MobileNet-U-Net+ResNet-U-Net							
Average	97.78	95.66	98.69	97.91	97.66	99.09	97.78
Weighted	97.78	95.66	98.69	97.91	97.66	99.09	97.78
Max	96.35	93.01	97.83	93.92	98.96	97.33	96.35
Vote	97.76	95.62	98.68	97.87	97.66	99.07	97.76
Confidence	97.77	95.54	98.68	97.90	97.65	99.09	97.77
U-Net+DeepLabV3pluse+DenseNet-U-Net+MFP-UNet+Attention U-Net+ResNet-U-Net							
Average	97.80	97.70	98.70	97.88	97.73	99.08	97.80
Weighted	97.80	97.70	98.70	97.88	97.73	99.08	97.80
Max	96.22	92.77	97.75	93.56	99.10	97.15	96.22
Vote	97.79	95.69	98.69	97.87	97.72	99.07	97.79
Confidence	97.76	95.63	98.68	98.14	97.40	99.19	97.76
U-Net+DeepLabV3pluse+DenseNet-U-Net+MFP-UNet+Attention U-Net+MobileNet-U-Net+ResNet-U-Net							
Average	98.74	95.79	98.73	97.88	97.82	99.09	98.74
Weighted	98.75	95.79	98.73	97.88	97.82	99.09	98.75
Max	96.07	92.51	97.68	93.19	99.21	97.02	96.07
Vote	97.83	95.75	98.71	97.86	97.80	99.08	97.83
Confidence	98.74	95.78	98.72	97.87	97.82	99.09	98.74

TABLE 6: Ensemble segmentation performance on the HC18 public dataset with varying model combinations and fusion strategies

MIoU, Mean Intersection over Union

Classification Performance Analysis

This work conducted a comprehensive ablation study examining the impact of data augmentation, model architectures, and ensemble methods on classification accuracy across three datasets. Four architectures (CNN, DenseNet121, ConvNext-Base, MobileNetV2) are evaluated with and without data augmentation using seven evaluation approaches: five individual methods (Max, Average, Confidence, Vote, Weighted),

their Ensemble combination, and a multi-input fusion approach.

Table 7 presents the results, which demonstrate that data augmentation significantly improves performance for most architectures, with DenseNet121 showing the most substantial gains on the HC Parameter dataset (from 84.96% to 90.04% maximum accuracy) and the FL Parameter dataset (from 85.40% to 88.60%). The multi-input approach consistently achieved the highest accuracy across all configurations, with DenseNet121 + augmentation + multi-input reaching peak performance of 92.50% on the HC Parameter dataset, 90.60% on the FL Parameter dataset, and 83.68% on the HC18 dataset. The Ensemble method also demonstrated strong performance, typically outperforming individual methods but falling short of multi-input results. While CNN architectures showed variable responses to augmentation (86.68% to 84.96% on HC Parameter and 81.85% to 86.95% on FL Parameter), advanced architectures like ConvNext-Base maintained consistent performance improvements with augmentation (88.23% to 90.15% on HC Parameter and 86.23% to 87.71% on FL Parameter). The ablation study confirms that the combination of DenseNet121 architecture, data augmentation, and multi-input approach provides optimal accuracy performance, validating our architectural choices and demonstrating the effectiveness of our proposed methodology.

Model	Max	Average	Confidence	Vote	Weighted	Ensemble	Multi-Input
Results for Custom Dataset HC Parameter							
CNN (without augmentation)	86.68	86.56	85.98	86.79	86.44	85.63	84.63
CNN (with augmentation)	84.96	85.66	84.96	86.01	85.31	86.36	88.46
DenseNet121 (without augmentation)	84.96	86.36	89.16	87.41	88.81	87.41	**91.95**
DenseNet121 (with augmentation)	90.04	90.95	91.37	90.53	90.53	90.15	**92.50**
ConvNext-Base (without augmentation)	88.23	87.41	88.18	87.31	86.23	82.20	88.89
ConvNext-Base (with augmentation)	90.15	90.74	90.80	91.09	90.88	86.40	91.32
MobileNetV2 (without augmentation)	83.57	88.81	87.76	86.36	88.33	88.46	89.12
MobileNetV2 (with augmentation)	90.01	90.12	91.23	91.22	90.12	87.71	90.14
Results for Custom Dataset FL Parameter							
CNN (without augmentation)	81.85	79.71	83.27	79.71	81.49	80.07	85.05
CNN (with augmentation)	86.95	87.07	86.83	86.62	86.83	85.76	87.40
DenseNet121 (without augmentation)	85.40	81.85	83.98	83.62	82.56	83.98	**87.18**
DenseNet121 (with augmentation)	88.60	88.39	88.96	88.03	88.11	85.26	**90.60**
ConvNext-Base (without augmentation)	86.23	86.89	86.81	86.18	87.23	84.32	87.89
ConvNext-Base (with augmentation)	87.71	87.09	88.13	87.88	88.01	85.01	89.30
MobileNetV2 (without augmentation)	81.85	82.21	78.65	82.92	79.72	84.34	86.23
MobileNetV2 (with augmentation)	89.17	89.25	89.60	89.74	89.28	85.83	88.19
Results for HC18 Dataset							
CNN (without augmentation)	74.00	72.50	72.50	74.50	73.00	71.00	73.50
CNN (with augmentation)	75.33	73.83	74.66	74.50	74.50	74.66	75.33
DenseNet121 (without augmentation)	73.00	72.00	70.50	73.00	69.00	73.00	**79.00**
DenseNet121 (with augmentation)	81.28	79.78	79.98	78.98	79.28	75.57	**83.68**
ConvNext-Base (without augmentation)	76.31	77.28	80.19	78.23	78.01	72.31	79.33
ConvNext-Base (with augmentation)	80.01	80.18	81.91	79.14	78.18	74.53	81.14
MobileNetV2 (without augmentation)	80.93	81.21	81.31	79.28	79.18	75.13	81.93
MobileNetV2 (with augmentation)	75.32	76.13	80.11	77.38	78.31	73.33	79.88

TABLE 7: Accuracy (%) performance comparison of five individual methods, their ensemble, and the multi-input fusion strategy across model architectures on the HC–custom dataset, FL–custom dataset, and HC–HC18 dataset.

FL, Femur Length; HC, Head Circumference

Experimental results

This section presents the experimental results corresponding to the three fundamental components of the proposed framework: noise removal, segmentation, and classification. A comprehensive evaluation is conducted in the subsequent paragraphs, highlighting the performance of the denoising methodology, the effectiveness of the segmentation algorithm, and the classification accuracy achieved on the processed fetal biometry image datasets.

Experiments on Noise Removal

The performance of the proposed denoising autoencoder is quantitatively evaluated using two widely recognized image quality metrics: PSNR and average SSIM. These metrics assess the effectiveness of the denoising process in preserving image details and structural information. The evaluation is conducted on two types of datasets: (a) the publicly available HC18 dataset and (b) a custom dataset consisting of ultrasound images for HC and FL parameters. As summarized in Table 8, the proposed model achieves consistently high PSNR and average SSIM values across all datasets, with the HC18 dataset demonstrating the best performance (PSNR = 39.43 dB, average SSIM = 0.9791). These results validate the robustness of the denoising method in improving image quality while maintaining structural consistency, which is crucial for subsequent segmentation and classification tasks.

Dataset	PSNR (dB)	SSIM
HC	35.67	0.9619
FL	36.43	0.9622
HC18	39.43	0.9791

TABLE 8: Quantitative evaluation of denoising performance using PSNR and SSIM

FL, Femur Length; HC, Head Circumference; PSNR, Peak Signal-to-Noise Ratio; SSIM, Structural Similarity Index Measure

The high PSNR and SSIM values indicate that the denoising autoencoder preserves structural information effectively while minimizing noise, making it well-suited for enhancing fetal medical imaging data.

Experiments on Segmentation

This Study evaluates the effectiveness of various deep learning-based segmentation architectures, using a comprehensive set of performance metrics, including Dice Coefficient, Mean Intersection over Union, Precision, Recall, and F1-Score. The models are evaluated on three datasets: (a) the created dataset with HC parameter, (b) the created dataset with FL parameter, and (c) the HC18 public dataset. The results are presented in Tables 9, 10, and 11, respectively.

As shown in Table 9, the DenseNet-U-Net architecture achieved the highest performance with a Dice Coefficient of 96.99% and MIoU of 93.50%, indicating superior segmentation accuracy and overlap with the ground truth masks. While other models such as U-Net, DeeplabV3+, and MFP-UNet also performed well with Dice scores above 96%, the DenseNet-based approach consistently outperformed them in both precision and IoU. The MobileNet-U-Net exhibited relatively lower performance, which may be attributed to its lightweight nature, leading to reduced representation capacity.

Segmentation Method	Dice Coefficient	MIOU	Precision	Recall	F1-Score
U-Net	96.17	92.65	95.29	97.10	96.17
DeeplabV3+	96.09	92.51	95.27	96.95	96.09
DenseNet-U-Net	96.99	93.50	96.84	96.38	96.99
MFP-UNet	96.09	92.50	95.84	96.38	96.09
U-Net-Attention	96.05	92.42	95.76	96.37	96.05
MobileNet-U-Net	94.18	89.08	94.51	93.93	94.18
ResNet-U-Net	95.98	92.31	96.42	95.58	95.98

TABLE 9: Segmentation performance metrics for the custom dataset head circumference parameter

MIoU, Mean Intersection over Union

Table 10 reports the segmentation metrics for the FL parameter. Again, the DenseNet-U-Net model demonstrated the best overall performance with a Dice Coefficient of 92.40% and MIoU of 85.55%.

DeeplabV3+ and ResNet-U-Net also yielded competitive results with F1-scores above 91%. The performance of all models on the FL dataset is slightly lower compared to the HC dataset, potentially due to greater anatomical variability or lower contrast in FL images. The MobileNet-U-Net showed the lowest performance across all metrics, suggesting its limitations in capturing finer structural details in the FL images.

Segmentation Method	Dice Coefficient	MIOU	Precision	Recall	F1-Score
U-Net	91.64	84.64	93.72	89.80	91.64
DeeplabV3+	92.11	85.43	92.35	92.00	92.11
DenseNet-U-Net	92.40	85.55	92.26	92.70	92.40
MFP-UNet	91.33	84.10	89.81	93.01	91.33
U-Net-Attention	91.62	84.65	92.91	90.53	91.62
MobileNet-U-Net	87.13	77.42	90.01	84.80	87.13
ResNet-U-Net	91.81	84.95	92.88	90.93	91.82

TABLE 10: Segmentation performance metrics for the custom dataset femur length parameter

MIoU, Mean Intersection over Union

In Table *11*, segmentation performance on the HC18 dataset is presented. All models achieved high accuracy, with DenseNet-U-Net and DeeplabV3+ leading the performance. DenseNet-U-Net achieved a Dice Coefficient of 97.76%, Precision of 97.22%, and the highest Recall of 98.39%, highlighting its robustness and consistency in segmenting fetal head structures. DeeplabV3+ also performed competitively with a Dice of 97.52% and a high IoU. Compared to the created datasets, all models performed slightly better on HC18, possibly due to higher image quality and standardized annotations.

Segmentation Method	Dice Coefficient	MIOU	Precision	Recall	F1-Score
U-Net	96.95	94.14	97.02	96.93	96.95
DeeplabV3+	97.52	95.18	98.04	97.03	97.52
DenseNet-U-Net	97.76	94.80	97.22	98.39	97.76
MFP-UNet	97.08	94.38	97.10	97.13	97.08
U-Net-Attention	97.02	94.28	96.75	97.35	97.02
MobileNet-U-Net	94.62	90.01	94.44	94.98	94.62
ResNet-U-Net	96.95	94.14	97.02	96.93	96.95

TABLE 11: Segmentation performance metrics for the HC18 dataset

MIoU, Mean Intersection over Union

To further enhance segmentation accuracy and robustness, multiple ensemble strategies are employed, including Average, Weighted, Max, Vote, and Confidence-based fusion. These methods combined predictions from multiple individual models to generate a final segmentation output. The evaluation is conducted on three datasets: (i) the created dataset with HC parameter, (ii) the created dataset with FL parameter, and (iii) the HC18 public dataset. The performance is assessed using seven evaluation metrics: Dice Coefficient, MIoU, Accuracy, Precision, Recall, Specificity, and F1-Score. The results are summarized in Table *12*.

For the HC parameter, the Average, Weighted, and Confidence-based ensemble strategies achieved the best performance with a Dice Coefficient of 97.81%, MIoU of 94.27%, and Accuracy of 98.11%. These methods showed consistent precision (97.32%) and high specificity (98.60%), indicating excellent discrimination between foreground and background. The Max-based ensemble, while achieving a high

Recall (98.46%), showed slightly lower precision (92.66%), leading to a comparatively reduced Dice score of 96.45%.

For the FL parameter, the Average, Weighted, and Confidence-based ensembles again yielded identical and superior results, with a Dice Coefficient of 92.84% and MIoU of 86.71%. These methods maintained high Accuracy (98.92%), Specificity (99.49%), and well-balanced Precision (93.66-93.73%) and Recall (92.10-92.17%). The Max strategy, while achieving the highest Recall (96.01%), demonstrated reduced overall performance due to its lower precision (83.78%) and Dice score (89.43%).

On the HC18 dataset, the Weighted and Average ensemble methods delivered the best performance, achieving a Dice Coefficient of 98.74-98.75%, an MIoU of 95.79%, and an Accuracy of 98.73%. These methods also showed strong performance across all other metrics, including Precision (97.88%), Recall (97.82%), and Specificity (99.09%), indicating a high degree of reliability. Although the Max ensemble achieved a high Recall (99.21%), its lower Dice score (96.07%) and Precision (93.19%) suggest over-segmentation tendencies.

Ensemble Segmentation Method	Dice Coefficient	MIOU	Accuracy	Precision	Recall	Specificity	F1-Score
Results for Created Dataset HC Parameter							
Average	97.81	94.27	98.11	97.32	96.72	98.60	97.81
Weighted	97.81	94.27	98.11	97.32	96.73	98.60	97.81
Max	96.45	91.34	97.49	92.66	98.46	97.09	96.45
Vote	97.79	93.24	98.10	96.29	96.72	98.59	97.79
Confidence	97.80	94.26	98.11	97.32	96.70	98.60	97.80
Results for Created Dataset FL Parameter							
Average	92.84	86.71	98.92	93.66	92.17	99.49	92.84
Weighted	92.84	86.71	98.92	93.66	92.17	99.49	92.84
Max	89.43	80.99	98.28	83.78	96.01	98.46	89.43
Vote	92.80	86.63	98.91	93.51	92.23	99.47	92.80
Confidence	92.84	86.71	98.92	93.73	92.10	99.49	92.84
Results for HC18 Dataset							
Average	98.74	95.79	98.73	97.88	97.82	99.09	98.74
Weighted	98.75	95.79	98.73	97.88	97.82	99.09	98.75
Max	96.07	92.51	97.68	93.19	99.21	97.02	96.07
Vote	97.83	95.75	98.71	97.86	97.80	99.08	97.83
Confidence	98.74	95.78	98.72	97.87	97.82	99.09	98.74

TABLE 12: Ensemble segmentation results for the custom dataset (HC and FL parameters) and the publicly available HC18 dataset

FL, Femur Length; HC, Head Circumference; MIoU, Mean Intersection over Union

Experiments on Classification

Table *13* shows the test accuracy (%) of the DenseNet121 model on three datasets: the custom HC parameter dataset consists of 1,426 fetal ultrasound images, stratified by gestational age into three trimesters: FT = 148 images, ST = 380 images, and TT = 898 images. The dataset is initially divided into 1,140 training images and 286 testing images, following an 80:20 split ratio. To improve model generalization and reduce the risk of overfitting, five distinct data augmentation techniques are employed. This process increased the total number of training samples to 7,130 images (i.e., each original image augmented five times). After augmentation, the trimester-wise distribution is updated to 740 FT, 1,900 ST, and 4,490 TT images. The augmented dataset is then partitioned into 5,704 training images and

1,426 testing images, preserving the original train-test distribution ratio.

In the case of the custom FL parameter dataset, a total of 1,404 fetal ultrasound images are collected and stratified by gestational age into three trimesters: FT: 112 images, ST: 395 images, and TT: 897 images. The dataset is initially divided following an 80:20 split, resulting in 1,123 training images and 281 testing images. To enhance model generalization and reduce the potential for overfitting, five distinct data augmentation techniques are applied, expanding the dataset to 7,020 images (i.e., each original image is augmented fivefold). Post-augmentation, the trimester-wise image distribution increased to 560 FT, 1,975 ST, and 4,485 TT images. The augmented dataset is then partitioned into 5,616 training images and 1,404 testing images, maintaining the original train-test split ratio.

The HC18 benchmark dataset consists of 999 fetal ultrasound images, with class-wise labelling based on estimated gestational age derived from HC measurements. Classification into trimesters is performed using threshold values obtained from standard fetal growth charts, resulting in 165 FT, 693 ST, and 141 TT images. An 80:20 train-test split is applied, yielding 799 training images and 200 testing images. To improve model generalization and reduce overfitting, five distinct data augmentation techniques are employed, resulting in an expanded dataset of 4,995 images (i.e., each original image augmented five times). After augmentation, the trimester-wise distribution increased to 825 FT, 3,465 ST, and 705 TT images. The augmented dataset is then partitioned into 3,996 training images and 999 testing images, maintaining the original train-test split ratio.

The results are obtained using different inference strategies, including Max, Average, Confidence, Vote, Weighted, ensemble, and the proposed Multi-Input method. The proposed Multi-Input method achieved the highest accuracy, with 92.50% for the HC parameter and 90.60% for the FL parameter on the custom dataset, and 83.68% on the HC18 dataset. These values are higher than those of all other methods. The results show that the Multi-Input method combines complementary features effectively, improving the accuracy and reliability of fetal biometric parameter estimation.

Model	Custom Dataset		HC18 Dataset
	HC Parameter	FL Parameter	
Average	90.04	88.60	81.28
Weighted	90.95	88.39	79.78
Max	91.37	88.96	79.98
Vote	90.53	88.11	78.98
Confidence	90.53	85.26	79.28
Ensemble	90.15	85.26	75.57
Proposed method Multi-Input	92.50	90.60	83.68

TABLE 13: Test accuracy (%) of DenseNet121 model on the custom HC, FL, and HC18 datasets

FL, Femur Length; HC, Head Circumference

The multi-input model demonstrated superior classification performance across all three datasets. Table *14* summarizes the class-wise evaluation metrics, including precision, recall, F1-score, and accuracy, for the custom HC parameter dataset, the custom FL parameter dataset, and the HC18 benchmark dataset. For the HC dataset, the model achieved high F1-scores of 96%, 86%, and 95% for the FT, ST, and TT, respectively. Similarly, in the FL dataset, strong performance is observed with F1-scores of 91% for FT, 82% for ST, and 94% for TT. The HC18 dataset, despite being more challenging, showed reasonable performance with F1-scores of 81%, 89%, and 64% for FT, ST, and TT classes, respectively. These results highlight the robustness of the multi-input strategy.

Class	Precision	Recall	F1-Score	Accuracy
Results for Custom Dataset HC Parameter				
FT	95	96	96	95.95
ST	84	89	86	89.21
TT	96	93	95	93.32
Results for Custom Dataset FL Parameter				
FT	85	98	91	98.21
ST	89	76	82	76.15
TT	92	96	94	96.00
Results for HC18 Dataset				
FT	74	88	81	88.48
ST	90	87	89	86.89
TT	65	62	64	62.41

TABLE 14: Class-wise performance metrics (%) of the multi-input DenseNet121 model on the custom HC, FL, and HC18 datasets

FL, Femur Length; FT, First Trimester; HC, Head Circumference; ST, Second Trimester; TT, Third Trimester

Figure *11* (a), (b), and (c) illustrate the evaluation results of the proposed model on the custom HC dataset, FL dataset, and HC18 benchmark dataset, respectively. Each subfigure presents three components: the confusion matrix, the ROC-AUC curve, and the training history. These visualizations collectively provide insights into the model's classification performance, discriminative capability, and training stability across the three datasets.

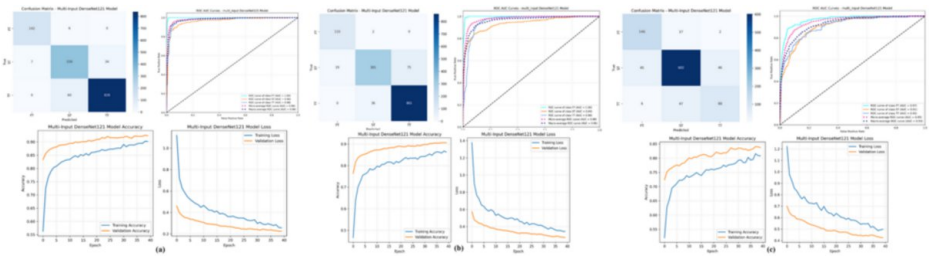

FIGURE 11: Confusion matrix, AUC-ROC curve, and training history of the DenseNet121 model on (a) the custom dataset HC parameter, (b) the custom dataset FL parameter, and (c) the publicly available HC18 dataset

AUC-ROC, Area Under the Receiver Operating Characteristic Curve; HC, Head Circumference; FL, Femur Length

Cross-validation results

This study introduces a cross-validation-based deep learning framework for trimester-wise fetal ultrasound image classification using DenseNet121 and its variants. A 5-fold stratified cross-validation approach with class-balanced folds is employed to ensure robust performance assessment across different data partitions. The experimental evaluation demonstrates stable and reliable classification performance across all folds. Table *15* summarizes the accuracy results: the custom dataset with HC parameters achieved consistently high performance, with accuracy values ranging from 88.57% to 91.65% and a mean accuracy of 89.80% ± 1.18%. When applying FL parameters, the model achieved moderate but consistent accuracy, ranging from 85.75% to 87.89% with a mean accuracy of 86.75% ± 0.92%. In contrast, the HC18 benchmark dataset yielded lower accuracy, between 76.58% and 79.58%, with a mean accuracy of 78.44% ±

1.16%.

These findings confirm that the custom dataset configurations, particularly with HC parameters, outperform the standard HC18 dataset. The consistent performance across all folds highlights the robustness and clinical applicability of the proposed methodology for automated fetal trimester classification.

Dataset	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD
Custom Dataset HC Parameter	91.65	89.69	89.90	90.18	88.57	89.80 ± 1.18
Custom Dataset FL Parameter	86.18	85.75	86.11	87.89	87.82	86.75 ± 0.92
HC18 Dataset	78.48	79.58	78.58	76.58	78.98	78.44 ± 1.16

TABLE 15: Classification accuracy (%) results using 5-fold cross-validation

HC, Head Circumference; FL, Femur Length

Comparative analysis

This comparative analysis examines recent studies employing various deep learning architectures for medical image segmentation. Table *16* presents a comparative summary of recent approaches for automatic segmentation of fetal HC using the HC18 Grand Challenge dataset. Chougule et al. [6] employed SegNet, GCN, and HRNet, achieving a Dice Similarity Coefficient (DSC) of 96%. Halder et al. [7] utilized U-Net and Attention U-Net models and reported a DSC of 97.17% and a Jaccard Coefficient of 94.51%. Alzubaidi et al. [9] applied an ensemble transfer learning framework, attaining a high MIoU of 98.53%. Ashkani Chenarlogh et al. [10] explored multiple U-Net variants, including MFP U-Net and dilated U-Net, and achieved a DSC of 97.45% and Jaccard Coefficient of 95%. More recent work by Dubey et al. [12] demonstrated the DR-ASPNet model, reporting a DSC of 98.86%, while Zeng et al. [17] proposed a DAG V-Net architecture and achieved a DSC of 97.63%. Conventional U-Net approaches by Nagabotu and Namburu [19], Fiorentino et al. [41], and Li et al. [31] yielded DSCs of 97.90%, 97.30%, and 97.26% respectively, with corresponding MIoU scores ranging between 96.46% and 97.81%. Wang et al. [24] applied a lightweight MobileNetV2-based model, obtaining a DSC of 96.28% and MIoU of 92.87%. In comparison, the proposed ensemble-based approach achieves superior performance with a DSC of 98.74% and an accuracy of 98.74% on the HC18 dataset, surpassing most existing methods. Furthermore, the method is validated on a newly created clinical dataset, attaining a DSC and accuracy of 97.81% for HC segmentation and 92.84% for FL segmentation. These results demonstrate the robustness and generalizability of the proposed framework across different fetal biometric parameters and datasets.

Authors	Dataset	Parameters	Methodology	Results
Chougule et al. [6]	HC18 Grand Challenge dataset	HC	SegNet, GCN, and HRNet	DSC=96%
Halder et al. [7]	HC18 Grand Challenge dataset	HC	UNet, Attention UNet	DSC=97.17%, Jaccard Co=94.51%
Alzubaidi et al. [9]	HC18 Grand Challenge dataset	HC	Ensemble Transfer Learning	MIoU=98.53%
Ashkani Chenarlogh et al. [10]	HC18 Grand Challenge dataset	HC	U-Net, MFP U-Net, dilated U-Net, Attention U-Net.	HC18: DSC=97.45%, Jaccard Co=95%
Dubey et al. [12]	HC18 Grand Challenge dataset	HC	DR-ASPnet	DSC=98.86%
Zeng et al. [17]	HC18 Grand Challenge dataset	HC	Deeply supervised attention-gated (DAG) V-Net	DSC=97.63%
Nagabotu and Namburu [19]	HC18 Grand Challenge dataset	HC	U-Net	DSC=97.90%, MIoU=97.81%
Fiorentino et al. [41]	HC18 Grand Challenge dataset	HC	U-Net	DSC=97.30%
Li et al. [39]	HC18 Grand Challenge dataset	HC	U-Net, SaPNeT	DSC=97.26%, MIoU=96.46%
Wang et al. [24]	HC18 Grand Challenge dataset	HC	Mobilenet V2	PA=97.77%, DSC=96.28%, MIoU=92.87%
Proposed Work	**HC18 Grand Challenge dataset**	**HC**	**Ensemble approach**	**DSC=98.74%, Accuracy=98.74%**
	Created Dataset	**HC**		**DSC=97.81%, Accuracy=97.81%**
		FL		**DSC=92.84%, Accuracy=92.84%**

TABLE 16: Comparative performance of deep learning methods for fetal biometric segmentation

DSC, Dice Similarity Coefficient; FL, Femur Length; HC, Head Circumference; MIoU, Mean Intersection over Union

Table *17* presents a comparative evaluation of recent deep learning models applied to fetal biometric classification tasks, including HC, FL, abdominal circumference, and multi-organ recognition. Gornale et al. [12] employed a U-Net architecture on a custom dataset, achieving the highest classification accuracy of 99.86%. Oghli et al. [40] proposed MFP-U-Net, obtaining 95.56% accuracy on both the HC18 and a custom dataset. Al-Razgan et al. [22] reported 94% accuracy using an attention-gated CNN for multi-organ classification. Sivasubramanian et al. [26] achieved 96.25% accuracy with EfficientNetV2B0 combined with a multilayer perceptron. Ghabri et al. [28] demonstrated high performance (99.97%) using DenseNet169, while Hasan Aowlad Hossain [32] attained 97.68% using an ensemble model on the FETAL_PLANES_DB. In comparison, the proposed trimester-based classification framework utilizing DenseNet121 achieved 86.68% accuracy on the HC18 dataset and improved results on a custom dataset (92.50% for HC and 90.60% for FL), indicating its potential for stage-wise fetal biometric analysis. Unlike existing studies that focus on general biometric classification, our work uniquely addresses trimester-based classification, highlighting its novelty and clinical relevance.

Authors	Dataset	Parameters	Methodology	Results
Dubey et al. [12]	Created Own Dataset	HC, FL, AC	U-Net	Accuracy=99.86%
Oghli et al. [40]	HC18 Grand Challenge dataset and Created Dataset	HC, FL, AC	MFP-U-Net	Accuracy=95.56%
Al-Razgan et al. [22]	Created Own Dataset	Multi-Organ	AG-CNN	Accuracy=94%
Sivasubramanian et al. [26]	Created Own Dataset	Multi-Organ	EfficientNetV2B0 + MLP	Accuracy=96.25%
Ghabri et al. [28]	Created Own Dataset	Multi-Organ	DenseNet169	Accuracy=99.97
Hasan and Aowlad Hossain [32]	FETAL_PLANES_DB	Fetal Planes	Ensemble Model	Accuracy=97.68%
Proposed Work (Trimester-based Classification)	HC18 Grand Challenge dataset	HC	DenseNet121 (Multi-Input Classification)	Accuracy=86.68%
	Created Dataset	HC		Accuracy=92.50%
		FL		Accuracy=90.60%

TABLE 17: Comparative performance of deep learning methods for fetal biometric classification

AC, Abdominal Circumference; AG-CNN, Attention-Guided Convolutional Neural Network; FL, Femur Length; HC, Head Circumference; MFP-U-Net, Multi-Feature Pyramid U-Net; MLP, Multilayer Perceptron

Statistical significance analysis using paired t-test

The study is to determine whether the observed classification outcomes are statistically significant; paired t-tests are conducted on the experimental results. This test is particularly appropriate for evaluating the differences between two related samples, such as predicted versus actual values [48]. The paired t-test examines whether the mean difference between these paired observations is significantly different from zero, thereby validating the performance of the classification model [49]. Table *18* presents the results of the paired t-test applied to both the custom dataset comprising HC and FL parameters, and the publicly available HC18 dataset. These results provide evidence supporting the statistical significance of the classification outcomes achieved by the DenseNet121-based Multi-Input Classification model.

Hypothesis in a Paired t-Test

Null Hypothesis (H_0): The mean difference μ_d between the paired groups is zero, implying no significant difference.

$$H_0 : \mu_d = 0$$

Alternative Hypothesis (H_1): The mean difference μ_d is not zero, implying a significant difference.

$$H_1 : \mu_d \neq 0$$

The statistic t is defined as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (19)$$

where, the differences between the paired observations are:

$$d_i = |\text{Actual}_i - \text{Predicted}_i| \quad (20)$$

The mean difference is:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad (21)$$

the standard deviation of the differences:

$$S_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2} \quad (22)$$

The t-statistic compares the mean $\bar{d}$ of the differences to the variability in the S_d differences.

The degrees of freedom (df) for a paired t-test is:

$$df = n - 1$$

where,

d_i: absolute difference for sample i.

Actual$_i$: ground-truth measurement (manual annotation).

Predicted$_i$: automated model output.

$\bar{d}$: mean of the absolute differences across all n samples.

s_d: standard deviation of the absolute differences.

SE: standard error of the mean difference, defined as $s_d/\sqrt{n}$.

t: test statistic used to determine whether the mean absolute difference is significantly different from zero.

$df = n - 1$: degrees of freedom for the t-distribution.

Dataset	Class	Actual Value	Predicted Value	Difference (d_i)	Where,
Custom Dataset HC Parameter	FT	148	142	6	n = 3, df = 2, d =35.67, S_d = 27.39, $t_{calculated}$ = 2.25, t_{table} = 4.303 $t_{calculated} < t_{table}$, H_0 is accepted
	ST	380	339	41	
	TT	898	838	60	
Custom Dataset FL Parameter	FT	112	110	2	n = 3, df = 2, d = 44, S_d = 46.52, $t_{calculated}$ = 1.64, t_{table} = 4.303, $t_{calculated} < t_{table}$, H_0 is accepted
	ST	395	301	94	
	TT	897	861	36	
HC18 Dataset	FT	165	146	19	n = 3, df = 2, d = 54.34, S_d = 36.01, $t_{calculated}$ = 2.61, t_{table} = 4.303, $t_{calculated} < t_{table}$, H_0 is accepted
	ST	693	602	91	
	TT	141	88	53	

TABLE 18: Paired t-test results for actual and predicted values obtained by the multi-input DenseNet121 model on the custom (HC, FL parameters) and HC18 datasets

FL, Femur Length; FT, First Trimester; HC, Head Circumference; ST, Second Trimester; TT, Third Trimester

Table *18* presents the results of the paired t-test performed at a 5% significance level (α = 0.05) using a two-tailed test with 2 degrees of freedom [48]. The critical t-value from the standard t-distribution table for this degree of freedom is t_{table} = 4.303. The calculated t-values for the comparisons between actual and predicted values are as follows: 2.25 for the HC parameter in the custom dataset, 1.64 for the FL parameter in the custom dataset, and 2.61 for the HC18 dataset. In all three cases, the calculated t-values are less than the critical value (4.303), meaning that the null hypothesis (H_0) is accepted. This indicates that there is no statistically significant difference between the actual and predicted values in any of the datasets. Therefore, the differences observed can be considered minor and are likely due to random variation rather

than a consistent error or bias in the model's predictions. These findings support the reliability of the model in predicting fetal parameters across both custom and standard datasets.

In addition to the paired t-test (Table *18*), we conducted an additional statistical analysis to enhance the robustness of our findings. Specifically, the Wilcoxon signed-rank test was employed as a non-parametric alternative [52].

Dataset & Parameter	Class	Paired t-test (t, p)	Wilcoxon (W, p)
Custom HC Parameter	FT, ST, TT	t = 2.25, p > 0.05	W = 0.0, p = 0.25
Custom FL Parameter	FT, ST, TT	t = 1.64, p > 0.05	W = 0.0, p = 0.25
HC18 Dataset	FT, ST, TT	t = 2.61, p > 0.05	W = 2.0, p = 0.75

TABLE 19: Comparison of paired t-test and Wilcoxon signed-rank test results for actual and predicted values across trimesters (FT, ST, and TT)

FL, Femur Length; FT, First Trimester; HC, Head Circumference; ST, Second Trimester; TT, Third Trimester

Table *19* presents the results of statistical tests applied to compare the custom parameters (HC, FL) and the HC18 dataset across trimesters (FT, ST, TT). Both the paired t-test and the Wilcoxon signed-rank test were conducted to assess consistency between actual and predicted class distributions. The paired t-test results (HC: $t = 2.25$, $p > 0.05$; FL: $t = 1.64$, $p > 0.05$; HC18: $t = 2.61$, $p > 0.05$) indicate that no statistically significant differences were observed. Similarly, the Wilcoxon signed-rank test (HC: $W = 0.0$, $p = 0.25$; FL: $W = 0.0$, $p = 0.25$; HC18: $W = 2.0$, $p = 0.75$) confirmed the absence of significant variation across the datasets. Although an important point is that using paired t-tests on class counts may not be statistically optimal compared to performance-based metrics, our analysis aimed to provide an additional supportive check for distribution-level consistency. The non-parametric Wilcoxon test was included to strengthen the reliability of the results, particularly given the small sample sizes. Importantly, the primary validation of our models is based on performance metrics, including the Dice coefficient, IoU, pixel accuracy, precision, recall, and F1-score, which provide clinically relevant insights into segmentation accuracy. Thus, while the statistical comparison of counts is supplementary, the core findings and conclusions of the study rely on robust performance-based evaluations.

Discussion

This study presents a comprehensive evaluation of deep learning methods for fetal biometric segmentation and trimester-wise classification, benchmarked against recent state-of-the-art approaches. For HC segmentation on the HC18 dataset, earlier studies have reported DSC scores ranging from 96% to 98.86% using architectures such as SegNet, U-Net variants, ensemble learning frameworks, and DR-ASPNet [12]. The proposed ensemble-based segmentation framework achieved a DSC and accuracy of 98.74%, surpassing most existing methods [9]. Furthermore, the approach demonstrated strong generalizability on a newly created dataset, achieving a DSC of 97.81% for HC segmentation and an accuracy of 92.84% for FL segmentation. In fetal biometric classification, prior works have achieved accuracies of up to 99.97% with models such as XceptionNet, MFP-U-Net, EfficientNet, and DenseNet [28]. The proposed trimester-based classification framework, built on DenseNet121, achieved 86.68% accuracy on the HC18 dataset and higher results on the custom dataset (92.50% for HC and 90.60% for FL). Notably, the classification framework integrates segmentation-driven inputs, where five complementary segmentation strategies, maximum probability, averaging, confidence-based, voting, and weighted fusion, are applied to produce multiple representations of each ultrasound image [46-48]. Each strategy emphasizes different anatomical cues, allowing the classification models to capture both fine-grained morphological structures and broader contextual patterns. This multi-representation approach enhances feature diversity and contributes to improved prediction reliability. The complete classification pipeline, illustrated in Figure *9*, outlines the stages of individual model training, multi-input feature fusion, and final prediction. Five-fold stratified cross-validation confirmed the stability of the proposed approach, with mean accuracies of 89.80% for HC and 86.75% for FL on the custom dataset, and 78.44% on the HC18 dataset. Paired t-test analysis at a 5% significance level revealed no statistically significant differences between actual and predicted values, indicating that performance variations are due to random factors rather than systematic bias. Overall, the results demonstrate the effectiveness of the proposed segmentation. The classification pipeline is accurate, robust, and clinically applicable across diverse datasets and fetal biometric parameters.

Limitations

This study has certain limitations. While the custom dataset is sizable, it is single-centered, and longitudinal images from the same fetus across trimesters were not available, limiting the ability to perform longitudinal analysis and reducing generalizability across diverse populations. Future work will therefore focus on validating the framework on larger, multi-center datasets with broader demographic and acquisition variability. Moreover, although the proposed framework demonstrated strong technical performance, the evaluation primarily relied on accuracy metrics rather than direct validation against real-world diagnostic outcomes. Collaborative clinical studies with obstetricians and sonographers will be essential to assess its clinical impact and integration into routine practice. Finally, the ensemble and multi-input pipeline, though effective in enhancing segmentation accuracy and robustness, introduces additional computational complexity compared to single-model approaches, which may hinder real-time or edge deployment. To address this, future efforts will investigate optimization strategies such as model compression, knowledge distillation, quantization, and pruning to improve inference efficiency without compromising accuracy.

Conclusions

This study presents a comprehensive multi-stage computational pipeline for enhancing ultrasound fetal image quality and interpretability. The proposed framework successfully addresses key challenges in prenatal imaging through three integrated components: (1) denoising autoencoders for speckle noise reduction and artifact mitigation, (2) an ensemble segmentation approach combining seven state-of-the-art deep learning architectures, and (3) trimester-specific biometric classification for fetal HC and FL measurements. The ensemble segmentation strategy, incorporating U-Net, DeepLabV3+, DenseNet-U-Net, MFP-UNet, Attention U-Net, MobileNet-U-Net, and ResNet-U-Net architectures, demonstrates superior performance by leveraging complementary model strengths, including advanced encoder-decoder structures, multi-scale feature extraction capabilities, efficient feature reuse mechanisms, sophisticated attention mechanisms, and optimized computational efficiency. The integration of trimester-specific classification enables more precise, developmentally aware biometric assessments, contributing to improved fetal growth monitoring and anomaly detection. Experimental results demonstrate significant improvements in image quality metrics and segmentation accuracy compared to individual model approaches, validating the effectiveness of the proposed ensemble methodology. The framework shows particular promise for enhancing diagnostic confidence in challenging imaging scenarios commonly encountered in clinical practice. The proposed framework represents a significant advancement toward automated, reliable prenatal ultrasound analysis, which has the potential to significantly improve clinical workflow efficiency and diagnostic accuracy in obstetric care.

Future research directions include (1) real-time system deployment for live ultrasound examination support, (2) dataset expansion to encompass diverse demographic populations and varying imaging conditions to enhance model generalizability, and (3) rigorous clinical validation through collaborative studies with healthcare professionals and clinical trustworthiness.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Priyanka C. Kamat, Shivanand S. Gornale, Prakash S. Hiremath , Rashmi Siddalingappa

Acquisition, analysis, or interpretation of data: Priyanka C. Kamat, Shivanand S. Gornale

Drafting of the manuscript: Priyanka C. Kamat

Critical review of the manuscript for important intellectual content: Shivanand S. Gornale, Prakash S. Hiremath , Rashmi Siddalingappa

Supervision: Shivanand S. Gornale

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there

are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The authors sincerely thank KSTEPS, the Department of Science and Technology (DST), and the Government of Karnataka for their financial support and the Ph.D. fellowship that made this research possible. The authors are also deeply grateful to Dr. Vanita B. Metgud and Dr. Satwik B. Metgud from Metgud Hospital – Advanced Laparoscopy Centre and IVF, Belagavi, Karnataka, India, for providing the ultrasound fetal images and valuable insights into fetal development, which greatly supported this study. The source code used for this research is publicly available at https://github.com/priyakamat/Fetal-Image-Analysis.git

References

1. van den Heuvel TL, de Bruijn D, de Korte CL, van Ginneken B: Automated measurement of fetal head circumference using 2D ultrasound images. PLoS ONE. 2018, 13:e0200412. 10.1371/journal.pone.0200412
2. Kiserud T, Piaggio G, Carroli G, et al.: The World Health Organization fetal growth charts: a multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight. PLoS Medicine. 2017, 14:e1002220. 10.1371/journal.pmed.1002220
3. Mengistu AK, Assaye BT, Flatie AB, Mossie Z: Detecting microcephaly and macrocephaly from ultrasound images using artificial intelligence. BMC Medical Imaging. 2025, 25:183. 10.1186/s12880-025-01709-x
4. Xiao X, Zhang J, Shao Y, Liu J, Shi K, He C, Kong D: Deep learning-based medical ultrasound image and video segmentation methods: overview, frontiers, and challenges. Sensors. 2025, 25:2361. 10.3390/s25082361
5. Danish HM, Suhail Z, Farooq F: Deep learning-based automation for segmentation and biometric measurement of the gestational sac in ultrasound images. Frontiers in Pediatrics. 2024, 12:1453302. 10.3389/fped.2024.1453302
6. Chougule A, Roy P, Wagh MG, Goyal D: Artificial intelligence-based estimation of fetal head circumference with biparietal and occipitofrontal diameters using two-dimensional ultrasound images. Journal of Engineering and Science in Medical Diagnostics and Therapy. 2025, 8:021103. 10.1115/1.4065799
7. Halder L, Mohanto S, Islam M, Jawad MT: Fetal head segmentation and head circumference measurement using residual U-Net. 2nd International Conference on Information and Communication Technology (ICICT). 2024, 145-149. 10.1109/ICICT64387.2024.10839692
8. Gopikrishna K, Niranjan NR, Maurya S, Krishnan VU, Surendran S: Automated classification and size estimation of fetal ventriculomegaly from MRI images: a comparative study of deep learning segmentation approaches. Procedia Computer Science. 2024, 233:743-752. 10.1016/j.procs.2024.03.263
9. Alzubaidi M, Agus M, Shah U, Makhlouf M, Alyafei K, Househ M: Ensemble transfer learning for fetal head analysis: from segmentation to gestational age and weight prediction. Diagnostics. 2022, 12:2229. 10.3390/diagnostics12092229
10. Ashkani Chenarlogh V, Ghelich Oghli M, Shabanzadeh A, et al.: Fast and accurate U-net model for fetal ultrasound image segmentation. Ultrasonic Imaging. 2022, 44:25-38. 10.1177/01617346211069882
11. Yousefpour Shahrivar R, Karami F, Karami E: Enhancing fetal anomaly detection in ultrasonography images: a review of machine learning-based approaches. Biomimetics. 2023, 8:519. 10.3390/biomimetics8070519
12. Dubey G, Srivastava S, Jayswal AK, Saraswat M, Singh P, Memoria M: Fetal ultrasound segmentation and measurements using appearance and shape prior based density regression with deep CNN and robust ellipse fitting. Journal of Imaging Informatics in Medicine. 2024, 37:247-267. 10.1007/s10278-023-00908-8
13. Gornale S, Kamat P, Siddalingappa R, Kumar S: Deep learning techniques for a comprehensive analysis of fetal biometric parameters across trimesters. Transactions on Machine Learning and Artificial Intelligence. 2024, 12:18-45. 10.14738/tecs.123.16985
14. Kamboj A, Bhalla P, Kumar Sunkaria R: MFA-FHS: multiscale feature based self attention network for fetal head segmentation. Multimedia Tools and Applications. 2025, 1-23. 10.1007/s11042-025-20824-z
15. Sobhaninia Z, Rafiei S, Emami A, et al.: Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2019, 6545-6548. 10.1109/embc.2019.8856981
16. Gornale SS, Patravali PU, Hiremath PS: Osteoarthritis detection in knee radiographic images using multiresolution wavelet filters. Recent Trends in Image Processing and Pattern Recognition. Santosh KC, Gawali B (ed): Springer, Singapore; 2020. 1381:36-49. 10.1007/978-981-16-0493-5_4
17. Zeng Y, Tsui PH, Wu W, Zhou Z, Wu S: Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated V-Net. Journal of Digital Imaging. 2021, 34:134-148. 10.1007/s10278-020-00410-5
18. Jader RF, Kareem SW, Awla HQ: Ensemble deep learning technique for detecting MRI brain tumor. Applied Computational Intelligence and Soft Computing. 2024, 2024:6615468. 10.1155/2024/6615468
19. Nagabotu V, Namburu A: Precise segmentation of fetal head in ultrasound images using improved U-Net model. ETRI Journal. 2023, 46:526-537. 10.4218/etrij.2023-0057
20. Rayed ME, Islam SMS, Niha SI, Jim JR, Kabir MM, Mridha MF: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked. 2024, 47:101504. 10.1016/j.imu.2024.101504
21. Tagnamas J, Ramadan H, Yahyaouy A, Tairi H: Multi-task approach based on combined CNN-transformer for efficient segmentation and classification of breast tumors in ultrasound images. Visual Computing for Industry Biomedicine and Art. 2024, 7:2. 10.1186/s42492-024-00155-w
22. Al-Razgan M, Ali YA, Awwad EM: Enhancing fetal medical image analysis through attention-guided convolution: a comparative study with established models. Journal of Disability Research. 2024, 3:10.57197/jdr-2024-0005
23. Kumar Y, Shrivastav S, Garg K, Modi N, Wiltos K, Woźniak M, Ijaz MF: Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification. Scientific Reports. 2024, 14:25006.

10.1038/s41598-024-75876-2
24. Wang F, Silvestre G, Curran KM : Segmenting fetal head with efficient fine-tuning strategies in low-resource settings: an empirical study with U-Net. arXiv. 2024, 10.48550/arXiv.2407.20086
25. Yang Y, Yang P, Zhang B: Automatic segmentation in fetal ultrasound images based on improved U-net. Journal of Physics: Conference Series. 2020, 1693:012183. 10.1088/1742-6596/1693/1/012183
26. Sivasubramanian A, Sasidharan D, Sowmya V, Ravi V: Efficient feature extraction using light-weight CNN attention-based deep learning architectures for ultrasound fetal plane classification. Physical and Engineering Sciences in Medicine. 2025, 10.1007/s13246-025-01566-6
27. Yeung M, Sala E, Schönlieb CB, Rundo L: Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized Medical Imaging and Graphics. 2022, 95:102026. 10.1016/j.compmedimag.2021.102026
28. Ghabri H, Alqahtani MS, Ben Othman S, et al.: Transfer learning for accurate fetal organ classification from ultrasound images: a potential tool for maternal healthcare providers. Scientific Reports. 2023, 13:17904. 10.1038/s41598-023-44689-0
29. Ma R, Zhang Y, Zhang B, Fang L, Huang D, Qi L: Learning Attention in the Frequency Domain for Flexible Real Photograph Denoising. IEEE Transactions on Image Processing. 2024, 33:3707-3721. 10.1109/tip.2024.3404253
30. Ma R, Li S, Zhang B, Fang L, Li Z: Flexible and generalized real photograph denoising exploiting dual meta attention. IEEE Transactions on Cybernetics. 2023, 53:6395-6407. 10.1109/tcyb.2022.3170472
31. Salomon LJ, Alfirevic Z, Berghella V, et al.: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics and Gynecology. 2011, 37:116-126. 10.1002/uog.8831
32. Hasan MN, Aowlad Hossain ABM: Fetal brain planes classification using deep ensemble transfer learning from U-Net segmented fetal neurosonography images. International Journal of Image Graphics and Signal Processing. 2024, 16:74-86. 10.5815/ijigsp.2024.04.06
33. Bano S, Dromey B, Vasconcelos F, Napolitano R, David AL, Peebles DM, Stoyanov D: AutoFB: automating fetal biometry estimation from standard ultrasound planes. Medical Image Computing and Computer Assisted Intervention - MICCAI 2021. de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, Essert C (ed): Springer, Cham; 2021. 12907:228-238. 10.1007/978-3-030-87234-2_22
34. Fiorentino MC, Villani FP, Di Cosmo M, Frontoni E, Moccia S: A review on deep-learning algorithms for fetal ultrasound-image analysis. Medical Image Analysis. 2022, 83:102629. 10.1016/j.media.2022.102629
35. Ma R, Zhang B, Zhou Y, Li Z, Lei F: PID controller-guided attention neural network learning for fast and effective real photographs denoising. IEEE Transactions on Neural Networks and Learning Systems. 2022, 33:3010-3023. 10.1109/TNNLS.2020.3048031
36. Ma R, Li S, Zhang B, Hu H: Meta PID attention network for flexible and efficient real-world noisy image denoising. IEEE Transactions on Image Processing. 2022, 31:2053-2066. 10.1109/tip.2022.3150294
37. Huang A, Jiang L, Zhang J, Wang Q: Attention-VGG16-UNet: a novel deep learning approach for automatic segmentation of the median nerve in ultrasound images. Quantitative Imaging in Medicine and Surgery. 2022, 12:3138-3150. 10.21037/qims-21-1074
38. Bag BC, Maity HK, Koley C: Unet Mobilenetv2: Medical image segmentation using deep neural network (DNN). Journal of Mechanics of Continua and Mathematical Sciences. 2023, 18:21-29. 10.26782/jmcms.2023.01.00002
39. Li P, Zhao H, Liu P, Cao F: Automated measurement network for accurate segmentation and parameter modification in fetal head ultrasound images. Medical & Biological Engineering & Computing. 2020, 58:2879-2892. 10.1007/s11517-020-02242-5
40. Oghli MG, Shabanzadeh A, Moradi S, et al.: Automatic fetal biometry prediction using a novel deep convolutional network architecture. Physica Medica. 2021, 88:127-137. 10.1016/j.ejmp.2021.06.020
41. Fiorentino MC, Moccia S, Capparuccini M, Giamberini S, Frontoni E: A regression framework to head-circumference delineation from US fetal images. Computer Methods and Programs in Biomedicine. 2021, 198:105771. 10.1016/j.cmpb.2020.105771
42. Oliveira-Saraiva D, Mendes J, Leote J, Gonzalez FA, Garcia N, Ferreira HA, Matela N: Make it less complex: autoencoder for speckle noise removal—application to breast and lung ultrasound. Journal of Imaging. 2023, 9:217. 10.3390/jimaging9100217
43. Wang Z, Liu M, Cheng X, et al.: Self-adaption and texture generation: A hybrid loss function for low-dose CT denoising. Journal of Applied Clinical Medical Physics. 2023, 24:e14113. 10.1002/acm2.14113
44. Shi L, Di W, Liu J: Ultrasound image denoising autoencoder model based on lightweight attention mechanism. Quantitative Imaging in Medicine and Surgery. 2024, 14:3557-3571. 10.21037/qims-23-1654
45. Paulauskaite-Taraseviciene A, Siaulys J, Jankauskas A, Jakuskaite G: A robust blood vessel segmentation technique for angiographic images employing multi-scale filtering approach. Journal of Clinical Medicine. 2025, 14:354. 10.3390/jcm14020354
46. Goceri E: Polyp segmentation using a hybrid vision transformer and a hybrid loss function. Journal of Imaging Informatics in Medicine. 2024, 37:851-863. 10.1007/s10278-023-00954-2
47. Panç K, Sekmen S: Multi-CNN deep feature fusion and stacking ensemble classifier for breast ultrasound lesion classification. Forbes Tıp Dergisi. 2025, 6:147-155. 10.4274/forbes.galenos.2025.02360
48. Podda AS, Balia R, Barra S, Carta S, Neri M, Guerriero S, Piano L: Multi-scale deep learning ensemble for segmentation of endometriotic lesions. Neural Computing and Applications. 2024, 36:14895-14908. 10.1007/s00521-024-09828-2
49. Guo Y, Chu T, Li Q, et al.: Diagnosis of major depressive disorder based on individualized brain functional and structural connectivity. Journal of Magnetic Resonance Imaging. 2024, 61:1712-1725. 10.1002/jmri.29617
50. Saifullah S, Dreżewski R, Yudhana A, Suryotomo AP: Automatic brain tumor segmentation: advancing U-Net with ResNet50 encoder for precise medical image analysis. IEEE Access. 2025, 13:43473-43489. 10.1109/access.2025.3547430
51. Burgos-Artizzu XP, 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. 2020, 10:10200. 10.1038/s41598-020-67076-5
52. Sugimori H, Shimizu K, Makita H, Suzuki M, Konno S: A comparative evaluation of computed tomography images for the classification of spirometric severity of the chronic obstructive pulmonary disease with deep learning. Diagnostics. 2021, 11:929. 10.3390/diagnostics11060929