Est. 1841 | YORK ST JOHN UNIVERSITY

Gornale, Shivanand S., Kamat, Priyanka, Siddalingappa, Rashmi, Li, Kefang and Wen Goh, Khang (2026) Automated Maternal–Fetal Health Analysis through Deep Neural Network Integration in Ultrasound Imaging. International Journal of Computer Information Systems and Industrial Management Applications, 18 (5s). pp. 1038-1068.

Downloaded from: https://ray.yorksj.ac.uk/id/eprint/15541/

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/10.70917/ijcisim-2026-2852

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

International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 1038-1068
https://doi.org/10.70917/ijcisim-2026-2852
Article

Automated Maternal–Fetal Health Analysis through Deep Neural Network Integration in Ultrasound Imaging

Shivanand S. Gornale[1], Priyanka Kamat[1*], Rashmi Siddalingappa[2], Kefang Li[3], and Khang Wen Goh[4]

[1]Department of Computer Science, School of Mathematics and Computing Sciences, Rani Channamma University, Belagavi, Karnataka, India;
[2]Department of Computer and Data Science, York St John University, United Kingdom.
[3]Faculty of Applied Sciences, Macao Polytechnic University, R. de Luís Gonzaga Gomes, Macao, China.
[4]Faculty of Data Science and Information Technology, INTI International University,71800 Nilai, Malaysia.
E-mails:[1]shivanand1971@rcub.ac.in,[1]priyankakamat@rcub.ac.in,[2]r.siddalingappa@yorksj.ac.uk,[3]kefengl@mpu.edu.mo, [4]khangwen.goh@newinti.edu.my.

Abstract: Ultrasonography is a widely used, non-invasive modality for monitoring fetal development and identifying growth abnormalities during pregnancy. This study proposes an attention-enhanced deep learning framework for automated fetal ultrasound image classification, addressing two tasks: (i) biometric parameter-based classification into four categories Head Circumference (HC), Femur Length (FL), Abdominal Circumference (AC), and Crown–Rump Length (CRL); and (ii) trimester-based classification into three stages using HC, FL, and the publicly available HC18 benchmark dataset. The framework employs a feature-level ensemble fusion strategy by integrating 1024-dimensional feature representations from four pretrained convolutional neural networks, ResNet50, VGG16, VGG19, and Xception, producing a unified 4096-dimensional fused vector to capture complementary and hierarchical features. Gradient-weighted Class Activation Mapping (Grad-CAM) is applied as an attention-enhanced preprocessing step using a strictly frozen ImageNet-pretrained VGG16, generating spatially weighted heatmap overlays that are superimposed onto the original fetal ultrasound images and used as enriched RGB inputs to the ensemble classifier, improving both classification performance and model interpretability. The entire pipeline, including Grad-CAM generation, dataset partitioning, data augmentation, and ensemble classification, operates automatically with no manual intervention at any stage. Data augmentation is applied exclusively to the training partition after dataset splitting, ensuring no data leakage into the validation or test sets. Experimental results demonstrate that the proposed fused ensemble model achieves 99.63% accuracy for biometric parameter-based classification, and 97.21%, 98.58%, and 92.72% for HC-based, FL-based, and HC18-based trimester classification, respectively, outperforming all individual baseline models. These findings confirm the effectiveness of combining Grad-CAM attention-enhanced inputs with deep ensemble feature fusion for robust, interpretable, and generalizable fetal ultrasound analysis in clinical decision support systems.

Keywords: Ultrasound Fetal Medical Image, Parameter-based Image Classification, Trimester-based Image Classification, Deep Learning Techniques, Feature Extraction.

1. Introduction

Ultrasound imaging is a cornerstone of obstetric care, offering non-invasive, real-time visualization of the fetus and maternal anatomy [1, 2]. It is routinely used to monitor fetal growth, estimate gestational age, measure critical biometric parameters, Head Circumference (HC), Femur Length (FL), Abdominal Circumference (AC), and Crown–Rump Length (CRL), evaluate fetal weight, and detect structural or developmental abnormalities. Pregnancy is divided into three trimesters: first (0–13 weeks), second (14–26 weeks), and third (27–40 weeks), during which fetal growth is closely tracked to support clinical decision-making and ensure maternal and neonatal safety [3]. Accurate and

Copyright: © 2026 by the authors

consistent measurement of these biometric parameters is clinically critical, as errors in parameter identification can lead to incorrect gestational age estimation, missed intrauterine growth restriction (IUGR), and inappropriate clinical management decisions [4, 5]. Manual ultrasound measurement is subject to inter-observer variability of up to 5–10% among trained sonographers [6, 7], which can translate directly into misdiagnosis or delayed clinical intervention, particularly in low-resource settings where experienced personnel may be unavailable.

Despite substantial reductions in global maternal mortality due to improved healthcare and increased prenatal care access, significant disparities persist in low-resource regions, where shortages of trained personnel, limited infrastructure, and restricted access to diagnostic tools contribute to disproportionately high maternal mortality rates [8, 9]. Automated fetal ultrasound analysis has the potential to reduce dependence on specialist expertise, support point-of-care screening, and provide quality control in large-scale prenatal programs. For an AI-assisted diagnostic tool to be clinically viable, it must achieve misclassification rates comparable to or lower than inter-observer variability among trained sonographers [6, 7]. Furthermore, clinical translation of such tools requires prospective validation and regulatory approval under established frameworks such as FDA 510(k) clearance or CE marking under the European Medical Device Regulation [10], aspects that remain important directions beyond the scope of the current study.

Recent advances in artificial intelligence (AI) and deep learning have demonstrated significant potential in medical image analysis, encompassing segmentation, classification, and anomaly detection [11, 12]. Convolutional neural networks (CNNs) pretrained on large-scale datasets have shown strong transfer learning capability for medical imaging tasks, enabling effective feature extraction even with limited domain-specific training data [13]. In fetal ultrasound, most prior studies have focused on single-parameter segmentation, particularly HC, or on fetal plane classification, providing limited insight into comprehensive multi-parameter fetal growth assessment across trimesters [14, 15]. Furthermore, existing approaches rarely integrate interpretability mechanisms with classification pipelines, limiting their transparency and clinical trustworthiness.

To address these limitations, this study proposes (i) an automated deep learning framework for parameter-based classification across HC, FL, AC, and CRL, and (ii) trimester-based classification using HC, FL, and the publicly available HC18 benchmark dataset [16, 17]. The system integrates feature-level ensemble fusion across four pretrained convolutional neural networks ResNet50, VGG16, VGG19, and Xception to capture complementary global and fine-grained feature representations [18, 19]. Gradient-weighted Class Activation Mapping (Grad-CAM) is applied as an attention-enhanced preprocessing step using a strictly frozen ImageNet-pretrained VGG16 [20, 21], generating spatia **l**y weighted heatmap overlays that are superimposed onto the original fetal ultrasound images and used as enriched inputs to the ensemble classifier [11, 22, 23]. Data augmentation is applied exclusively to the training partition after dataset splitting, ensuring no data leakage into the validation or test sets. The proposed framework is evaluated against the clinical inter-observer variability benchmark to assess its potential utility as a reliable and interpretable tool for automated fetal assessment in obstetric decision support systems. The main contributions are as follows:

1.1 Key Contributions

- Development of a custom dataset of 3,587 fetal ultrasound images covering HC, FL, AC, and CRL for multi-parameter analysis.
- Design of a unified deep learning framework for parameter-based and trimester-wise classification of fetal ultrasound images.
- Integration of multiple pretrained CNNs (ResNet50, VGG16, VGG19, Xception) with a feature fusion strategy for enhanced representation learning.
- Integration of Grad-CAM–driven attention-enhanced inputs along with a leakage-free evaluation protocol to enhance model interpretability.

2. Related Work

This section presents a comprehensive literature review of prior research conducted on fetal ultrasound image analysis, fetal biometry estimation, segmentation, and classification using deep learning techniques[24, 25].

Deep learning approaches, particularly CNN-based models, have been widely used in fetal ultrasound imaging for classification, detection, segmentation, and standard plane analysis. Baumgartner et al.[1] introduced SonoNet, a weakly supervised, real-time framework based on VGG16 for fetal standard-plane detection and localization,

demonstrating strong clinical applicability. Al-Razgan et al. [2]proposed an attention-guided CNN (AG-CNN) that improves feature representation and outperforms traditional architectures such as ResNet, DenseNet, and VGG. Ishikawa et al. [21]developed a CNN-based method for fetal part classification and position estimation with high recall across anatomical structures, supporting automated analysis. Further improvements have been achieved through hybrid and optimized deep learning frameworks. Rauf et al.[23] introduced a feature fusion-based model with optimization strategies, while Sivasubramanian et al.[26] proposed a lightweight attention-enhanced EfficientNet-based architecture, achieving high accuracy with reduced complexity. Transfer learning has also been widely adopted. Ghabri et al [27]. demonstrated that pre-trained CNNs such as ResNet, DenseNet, and MobileNet achieve strong performance on maternal–fetal datasets. In segmentation and biometric analysis, Gornale et al.[9, 17, 28] proposed a multi-stage framework that combines denoising, segmentation, and classification, achieving high accuracy in trimester-based analysis. They also developed deep learning pipelines for fetal biometric measurements, such as head circumference estimation. However, these methods are often limited to controlled datasets and single-parameter analysis, with limited robustness to complex clinical variability. Several recent studies have also explored hybrid and interpretable approaches. Rathika et al.[29] introduced feature selection–based classification using optimized GLCM features, while Harikumar et al.[30] incorporated LIME for interpretability in CNN-based fetal plane classification. Hasan et al.[31] proposed an ensemble deep learning framework for fetal brain plane classification, and Liang et al.[32] developed SPRNet with partial transfer learning, achieving high recognition accuracy. Despite these advances, most approaches remain CNN-centric and dependent on labelled data.

Recently, more advanced paradigms such as foundation models, self-supervised learning, federated learning, and domain adaptation have emerged to address challenges like data scarcity, domain shift, and label noise [33–35]. However, these methods require large-scale data and high computational resources, limiting their clinical deployment. In contrast, CNN-based methods remain practical due to their efficiency and simplicity. Motivated by this, the proposed work adopts a fused CNN framework integrating multiple pretrained backbones with attention-enhanced inputs to improve feature representation, interpretability, and robustness, making it suitable for resource-constrained clinical environments[36, 37].

3. Materials and Methods

This section describes the proposed deep learning framework for fetal ultrasound image classification. The overall pipeline consists of six sequential stages: (i) dataset description, (ii) preprocessing, (iii) attention-enhanced input generation via Grad-CAM, (iv) dataset partitioning, (v) data augmentation, and (vi) Feature-Level Fusion-Based Ensemble Framework. Each stage is described in detail below.

3.1 Dataset Description

Fetal biometric parameters are critical for monitoring growth and detecting potential complications. Existing datasets lack focused coverage of these parameters. To address this, a custom dataset is created, comprising images for HC, FL, AC, and CRL, with HC and FL specifically used for trimester-based classification. This dataset supports research in fetal biometric analysis and enhances the training of deep learning models for accurate parameter- and trimester-specific predictions.

3.1.1 HC18 Dataset

The HC18 dataset contains 1,334 fetal head ultrasound images (800×540 pixels) collected from 551 pregnant women across the first, second, and third trimesters at Radboud University Medical Center, Netherlands. All images were acquired using Voluson E8 and Voluson 730 ultrasound systems, with pixel sizes ranging from 0.052 to 0.326 mm. The dataset is divided into 999 training and 335 testing samples, and HC measurements are provided only for the training set. Expert sonographers manually annotated the skull boundaries using ellipses, and the CMO Arnhem-Nijmegen Ethics Committee approved all procedures in accordance with the Declaration of Helsinki[16]. Despite its usefulness, HC18 has limitations: it includes only the HC parameter, trimester labels are not explicitly provided, and the complexity of fetal images makes trimester categorization challenging. To address these gaps, a larger custom dataset was developed, incorporating four biometric parameters and enabling trimester-wise classification for both HC and FL images.

Trimester classification is defined as follows: The HC18 dataset is a publicly available benchmark in which head circumference (HC) measurements are provided only for the training set, while the test set does not include HC annotations. Accordingly, this study utilizes all 999 training images for trimester classification. Trimester categories are defined as follows: fetuses up to 13 weeks are assigned to the first trimester, 14–26 weeks to the second trimester,

and 27–40 weeks to the third trimester. Gestational age is estimated by systematically mapping HC measurements to standardized World Health Organization (WHO) fetal growth charts, as summarized in Table 1. A rule-based approach is employed to convert HC values into gestational weeks and corresponding trimester labels. All samples were successfully labelled using this procedure, and no cases were excluded as "Abnormal," as all HC values fell within the valid WHO reference ranges. The final annotated dataset is compiled into a structured CSV file to support efficient analysis and model development. The resulting class distribution comprises 165 first-trimester, 693 second-trimester, and 141 third-trimester samples [3, 38].

Table 1. The World Health Organization (WHO) Fetal Growth Chart is Used to Estimate Gestational Age from Head Circumference (HC) Measurements.

Gestational Age (Weeks)	Head Circumference (mm) by percentile								
	2.5	5	10	25	50	75	90	95	97.5
14	86	88	91	95	100	104	107	110	112
15	97	99	102	106	111	115	119	122	124
16	108	111	114	118	123	128	132	134	137
17	120	123	126	130	135	140	144	147	149
18	132	135	138	143	148	153	157	160	162
19	145	147	150	155	161	166	170	173	175
20	157	159	163	168	173	179	183	186	188
21	169	172	175	180	186	191	196	199	201
22	181	184	187	193	198	204	209	212	214
23	193	196	199	205	210	216	221	224	227
24	204	207	211	216	222	228	233	236	239
25	215	218	222	227	233	239	245	248	251
26	225	228	232	238	244	250	256	259	262
27	234	238	242	248	254	261	267	270	273
28	243	247	251	257	264	270	277	280	283
29	251	256	260	266	273	280	286	290	293
30	259	264	268	274	281	288	295	299	302
31	266	271	275	282	289	296	303	307	311
32	273	278	282	289	296	304	311	315	318
33	279	284	289	295	303	311	318	322	326
31	285	290	295	302	309	317	324	328	332
35	291	296	300	307	315	323	330	335	338
36	296	301	306	313	321	329	336	340	344
37	302	306	311	318	326	334	341	345	349
38	307	311	315	324	332	339	347	350	354
39	313	316	320	329	337	344	352	355	359
40	319	321	325	334	342	350	357	360	363

doi:10.1371/journal.pmed.1002220.t007

3.1.2 Custom Dataset

The custom dataset makes a significant contribution to the field by adding more cases to improve the diversity and complexity of medical images. This provides a strong foundation for thorough evaluation. The dataset includes fetal age in weeks, and the ima ges are labelled according to the corresponding trimester based on this information. The dataset comprises a total of 3,587 images categorized into HC, FL, AC, and CRL. However, CRL images are limited to the first trimester, and AC images are absent for the third trimester; therefore, due to incomplete trimester-wise representation, these parameters were excluded from trimester classification, which was conducted using only HC and FL. There are 1,426 HC images, which include 148 from the FT, 380 from the ST, and 898 from the TT. Then, there are 1,409 FL images, with 116 from the FT, 395 from the ST, and 898 from the TT. Additionally, there are 553 CRL images and 199 AC images. Figure 1 presents representative samples of images categorized by their respective parameters.

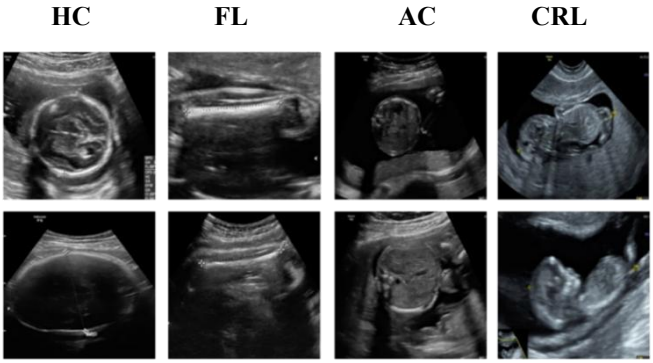


Figure 1. Samples of Fetal Images

The images were acquired using the VOLUSON P6 Ultrasound Machine at Metgud Hospital - Advanced Laparoscopy Centre and IVF in Belagavi, Karnataka, India. All images were acquired using a VOLUSON P6 ultrasound system from pregnant women aged 21–38 years. Only cases with normal fetal growth were included, while abnormal conditions, severe pregnancy complications, and multiple pregnancies were excluded to ensure consistency. The images were stored in JPG format with an original resolution of 640 × 480 pixels. To focus on clinically relevant regions, regions of interest (ROIs) corresponding to the fetal head (for HC) and femur (for FL) were extracted using a semi-manual cropping protocol guided by anatomical landmarks. The cropped images were scaled to a uniform resolution of 300 × 300 pixels to ensure consistent input dimensions for m odel training. Although resizing reduces spatial detail, care was taken to preserve the ROI such that essential anatomical features remain intact, and no significant loss of discriminative information was observed. All cropped samples were visually verified under expert supervision to maintain consistency. The dataset does not include predefined training and testing splits, allowing flexible experimental design. All identifying information was removed to ensure patient privacy, and data collection was conducted in accordance with established ethical guidelines and the PCNDT Act, with approval obtained from the District Hea lth and Family Welfare Office, Belagavi, Karnataka [9, 28] .

3.2 Proposed Methodology

The proposed methodology classifies key fetal biometric parameters Head Circumference (HC), Femur Length (FL), Abdominal Circumference (AC), and Crown–Rump Length (CRL) and performs trimester-based classification to enhance prenatal diagnostics. A strictly frozen ImageNet-pretrained VGG16 is used solely as a domain-agnostic attention generator to produce Grad-CAM attention-enhanced overlay images, which serve as enriched RGB inputs to the ensemble classifier. Four deep learning backbone architectures VGG16, VGG19, ResNet50, and Xception are trained on these attention-enhanced inputs, and their extracted feature representations are fused through a model fusion strategy to enhance classification accuracy and robustness[18, 19, 39].

3.2.1 Fetal Image Acquisition

Ultrasound images are collected using the VOLUSON P6 Ultrasound Machine, a state-of-the-art imaging system known for its high resolution and precision in visualizing internal anatomical structures. The acquired images are exported as JPGs and systematically organized into well-structured folders, ensuring efficient storage, easy access, and streamlined data management. The proposed methodology is illustrated in Figure 2.

3.2.2 Pre-Processing Techniques

Preprocessing of fetal ultrasound images enhances quality, reduces artifacts, and highlights key anatomical features for deep learning. Images are organized by biometric parameters

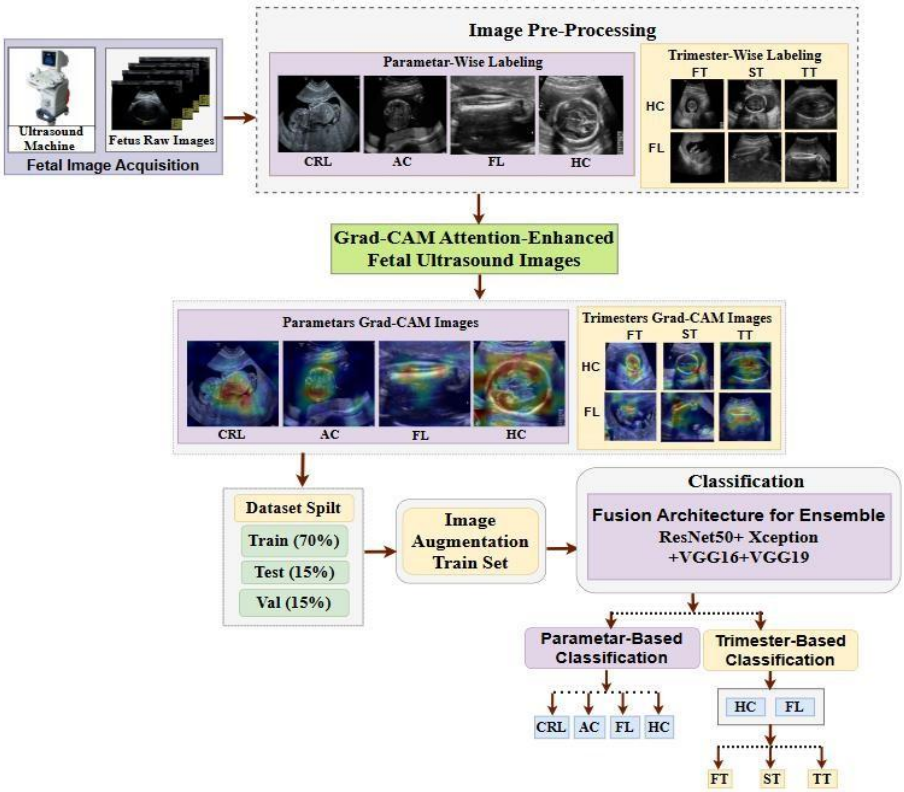


Figure 2. Block Diagram of Proposed Method

(HC, AC, FL, CRL) and trimester, then cropped to focus on the fetal head, removing irrelevant background. All images are scaled to a uniform resolution of 300 × 300 pixels and converted to grayscale to simplify input and improve contrast[17, 40]. Intensity normalization and denoising standardize pixel values and suppress artifacts. These steps produce a consistent, high-quality dataset that improves model training, robustness, and reproducibility. Figure 3 illustrates the workflow, including intensity histograms showing enhanced contrast and edge clarity **.**

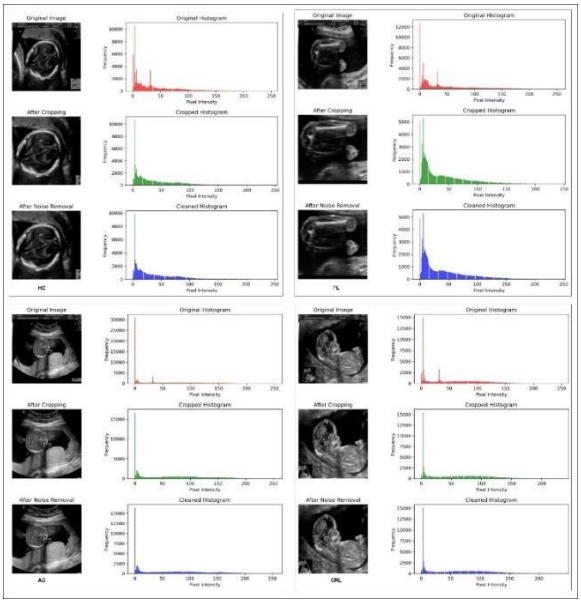


Figure 3. Preprocessing Representation of Fetal Biometric Parameters

3.3 *Attention-Enhanced Input Generation via Grad-CAM*

To improve the spatial discriminability of fetal ultrasound images, Grad-CAM is applied as a preprocessing step. This method highlights image regions that contribute most to model predictions [20, 21].

A pretrained VGG16 network (ImageNet) is used only as a domain-agnostic attention generator. The model weights are kept strictly frozen, and no fine-tuning is performed on any portion of the fetal biometry dataset. The softmax layer is removed to obtain raw logits. Each image is resized to 224 × 224 and normalized using ImageNet preprocessing. For every input image, a forward pass is performed, and the top predicted ImageNet class index is selected as the backpropagation target:

$$c = \mathbf{arg\,max}(\text{softmax}(y)) \quad (1)$$

It is important to note that this target class corresponds to an ImageNet category entirely unrelated to the fetal task labels (HC, FL, AC, CRL), confirming that no task-specific label information is involved in heatmap generation.

The gradient of the class score y^c is computed with respect to the feature maps of the last convolutional layer (block5_conv3, size $\mathbf{14} \times \mathbf{14} \times \mathbf{512}$). These gradients are averaged spatially to obtain channel-wise importance weights:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (2)$$

where A^k represents the activation at location (i, j) in feature map k, and Z is the total number of spatial locations.

The Grad-CAM heatmap is then computed as a weighted combination of feature maps followed by ReLU activation:

$$L^c_{\text{Grad-CAM}} = \mathbf{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (3)$$

The ReLU operation keeps only positive contributions and removes irrelevant regions. The resulting heatmap (size $\mathbf{14} \times \mathbf{14}$) is normalized to (0,1), scaled to (0,255), and colorized using a jet colormap. It is then resized to match the original image dimensions.

The final attention-enhanced image is obtained by overlaying the heatmap onto the original image with a fixed blending coefficient of α = 0.4:

$$I_{\text{enhanced}} = \alpha \cdot H + I_{\text{original}}, \alpha = \mathbf{0.4} \quad (4)$$

where H_{jet} is the colorized heatmap and I_{original} is the input image.

In this study, VGG16 is utilized exclusively for attention extraction. The network is not used for classification, and its weights remain fixed throughout. The attention-enhanced images produced by Equation (4) superimpose RGB overlays of identical spatial dimensions to the original input and serve as the sole inputs to the ensemble classifier. They are not used as binary masks, additional input channels, or independent feature streams. Since the heatmap generator relies on strictly frozen ImageNet weights with no exposure to the experimental dataset or its labels, applying Grad-CAM before dataset partitioning introduces no data leakage or circularity.

The overall Grad-CAM-based image generation process is illustrated in Figure 4.

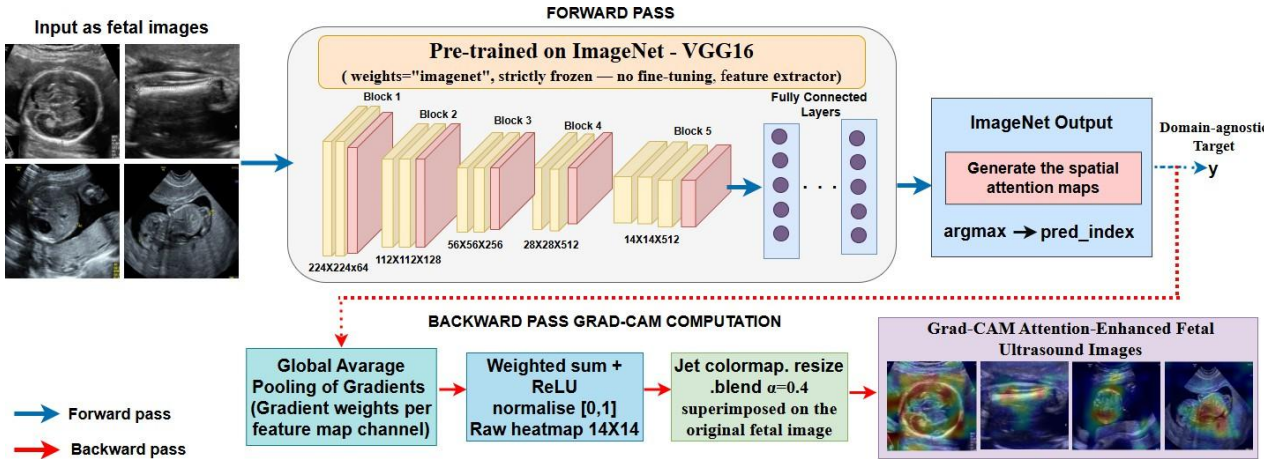


Figure 4. Grad-CAM Attention-Enhanced Input Generation Framework Using Frozen ImageNet-Pretrained VGG16

The Grad-CAM attention-enhanced images used as inputs to the ensemble classifier are shown in Figure 5, while Figure 6 presents the original fetal ultrasound images for each biometric parameter alongside their corresponding Grad-CAM heatmaps and pixel intensity histograms. For Head Circumference (HC), the original grayscale images are displayed on the left, and the Grad-CAM overlay images are shown on the right, highlighting the anatomical regions receiving the closest attention across the First, Second, and Third trimesters, with warmer colors indicating stronger activation. Similar visualizations are provided for Femur Length (FL), Abdominal Circumference (AC), and Crown–Rump Length (CRL), where the Grad-CAM maps consistently emphasize the anatomica **l**y relevant regions for each biometric parameter. The pixel intensity histograms further quantify the spatial distribution of attention within each image, confirming that the model concentrates on high-importance discriminative regions rather than background areas. Overall, these results demonstrate that the Grad-CAM attention-enhanced inputs enable the ensemble classifier to effectively utilize key anatomical features for both parameter-based and trimester-based classification, while providing interpretable and clinically meaningful insights into the model's decision-making process [20, 21].

Although VGG16 predicts ImageNet categories unrelated to fetal anatomy, the scientific validity of this approach is grounded in established transfer learning theory. Deep CNNs pretrained on large-scale datasets learn universal low-level features, edges, textures, and blob-like structures that transfer effectively across visual domains, including medical imaging [41, 42]. The ImageNet class index serves solely as the mathematical target for gradient backpropagation, revealing spatial feature importance independent of label semantics[20]. In fetal ultrasound images, the anatomical structure of interest is the most visually salient and texturally complex region, while background regions are acoustically homogeneous and texturally flat. The frozen VGG16 assigns the highest activation to the fetal anatomical structure as a consequence of visual saliency rather than semantic label knowledge, producing medically meaningful attention maps[43]. This is empirically confirmed by the Grad-CAM visualizations in Figure 5, which consistently highlight anatomically relevant fetal structures across all biometric parameter categories.

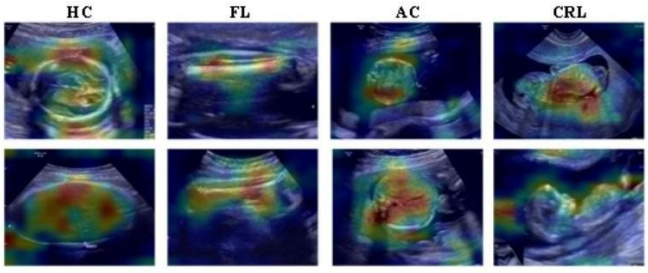


Figure 5. Sample Grad-CAM Attention-Enhanced Images

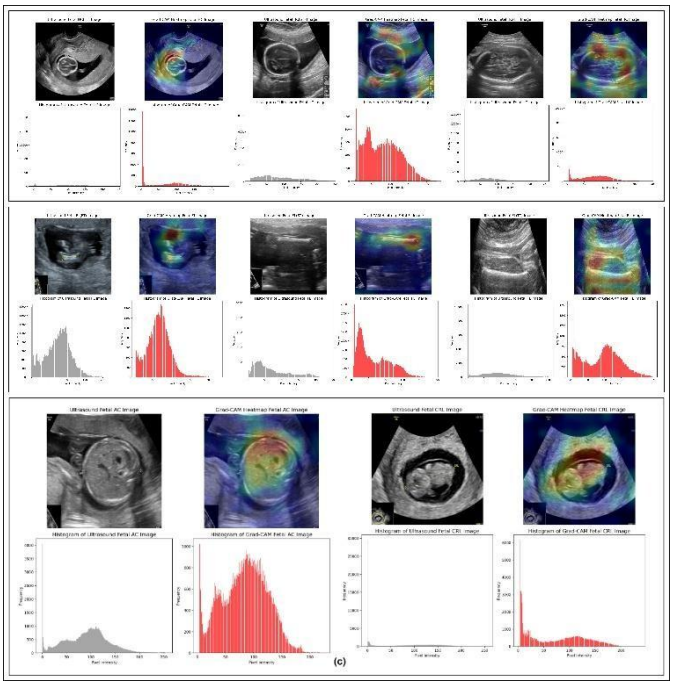


Figure 6. Representative Grad-CAM Attention-Enhanced Fetal Ultrasound Images with Corresponding Activation Heatmaps and Pixel Intensity Histograms for (a) Head Circumference (HC), (b) Femur Length (FL) Across First, Second, and Third Trimesters, and

(c) Abdominal Circumference (AC) and Crown–Rump Length (CRL).

3.4 Data Augmentation

The study implements a data augmentation pipeline to increase the diversity and size of the fetal ultrasound image training set. A total of 15 advanced augmentation techniques is applied, encompassing geometric, photometric, and noise-based

transformations to simulate real-world variations while preserving fetal anatomical structures. Geometric augmentations included rotation (±15°), scaling (0.9 -1.1×), translation (±10% of image dimensions), shear (±10°), and horizontal and vertical flipping. Photometric adjustments involved modifications in contrast, hue, and saturation, while Gaussian noise and blurring are used to replicate common imaging artifacts. All images are normalized to the range 0 -1 before being fed into deep learning models. These augmentation strategies are designed to enhance model generalization, prevent overfitting, and ensure methodological transparency and reproducibility in fetal medical image analysis[22]. The sample augmented images are depicted in Figure 7.

3.5 *Feature-Level Fusion-Based Ensemble Framework*

Feature extraction is performed using four pretrained convolutional neural networks: ResNet50, Xception, VGG16 , and VGG19, all initialized with ImageNet weights, selected for their demonstrated ability to learn rich and discriminative feature representations from medical imaging data [44, 45]. Each backbone processes the Grad-CAM attention-enhanced images as inputs and extracts deep feature representations, which are subsequently integrated through the proposed ensemble fusion strategy to enhance classification accuracy and robustness. The architectural details of all four backbone models are presented in Figure 8.

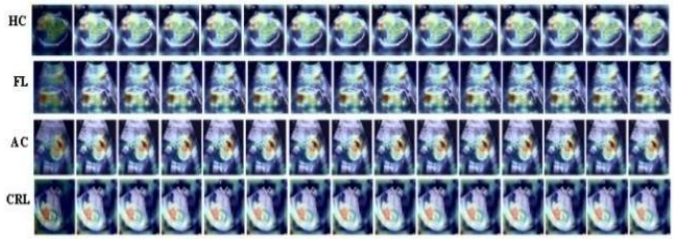


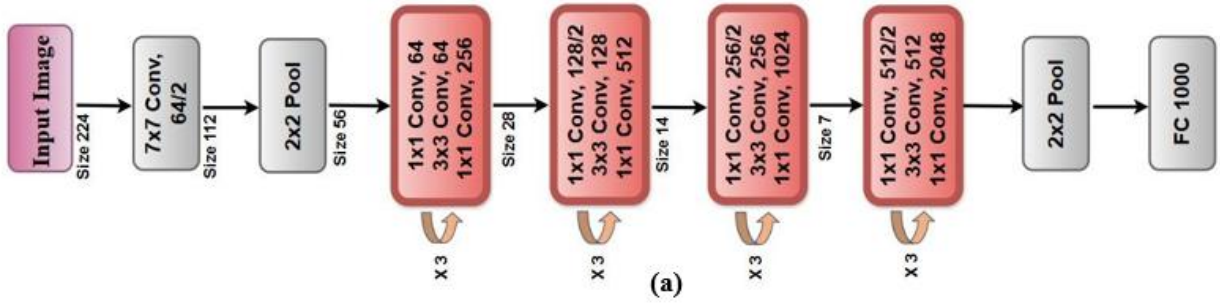
Figure 7. Samples of Augmented Grad-CAM images

3.5.1 **ResNet50:** ResNet50 is a 50-layer deep convolutional neural network that uses residual connections to address the vanishing gradient problem and enable efficient training of deep architectures. Each residual block learns a mapping defined as $y = F(x, W_i) + x$, where x is the input and $F(x, W_i)$ represents the residual function. This design improves gradient flow and model performance. In this study, a pre-trained ResNet50 model (ImageNet) is fine-tuned on fetal ultrasound images. Lower layers are frozen to retain general features, while higher la yers are trained for domain-specific feature extraction. A Global Max Pooling layer, followed by a fully connected layer (1024 neurons) and a softmax classifier, is used for final classification[18, 19, 46].

3.5.2 **Xception:** Xception (Extreme Inception) is a deep convolutional neural network that replaces standard convolutions with depth-wise separable convolutions to improve efficiency and feature representation. The operation is defined as $Y = P(D(X))$, where $D(X)$ represents depth-wise convolution and $P(\cdot)$denotes pointwise convolution. This factorization reduces parameters while enhancing learning capacity. In this study, a pre-trained Xception model (ImageNet) is fine-tuned on fetal ultrasound images. Early layers are frozen to retain general features, while deeper layers are trained to capture domain-specific patterns. A Global Average Pooling layer, followed by a fully connected layer (1024 neurons) and a softmax classifier, is used for final classification[39].

3.5.3 **VGG16 & VGG19:** VGG16 and VGG19 are deep

convolutional neural networks developed by the Visual Geometry Group (VGG) at Oxford, consisting of 16 and 19 weight layers, respectively. Both architectures use uniform 3 × 3convolutional filters and ReLU activations to effectively learn hierarchical image features. The convolutional operation is defined as $y = \sigma(W * x + b)$, where $x$is the input, $W$and $b$are learnable parameters, and $\sigma$is the activation function. In this study, pre-trained VGG16 and VGG19 models (ImageNet) are fine-tuned on fetal ultrasound images. Lower layers are frozen to retain general features, while deeper layers are retrained for domain-specific learning. A Global Max Pooling layer, followed by a fully connected layer (1024 neurons) and a softmax classifier, is used for final classification[18, 19].



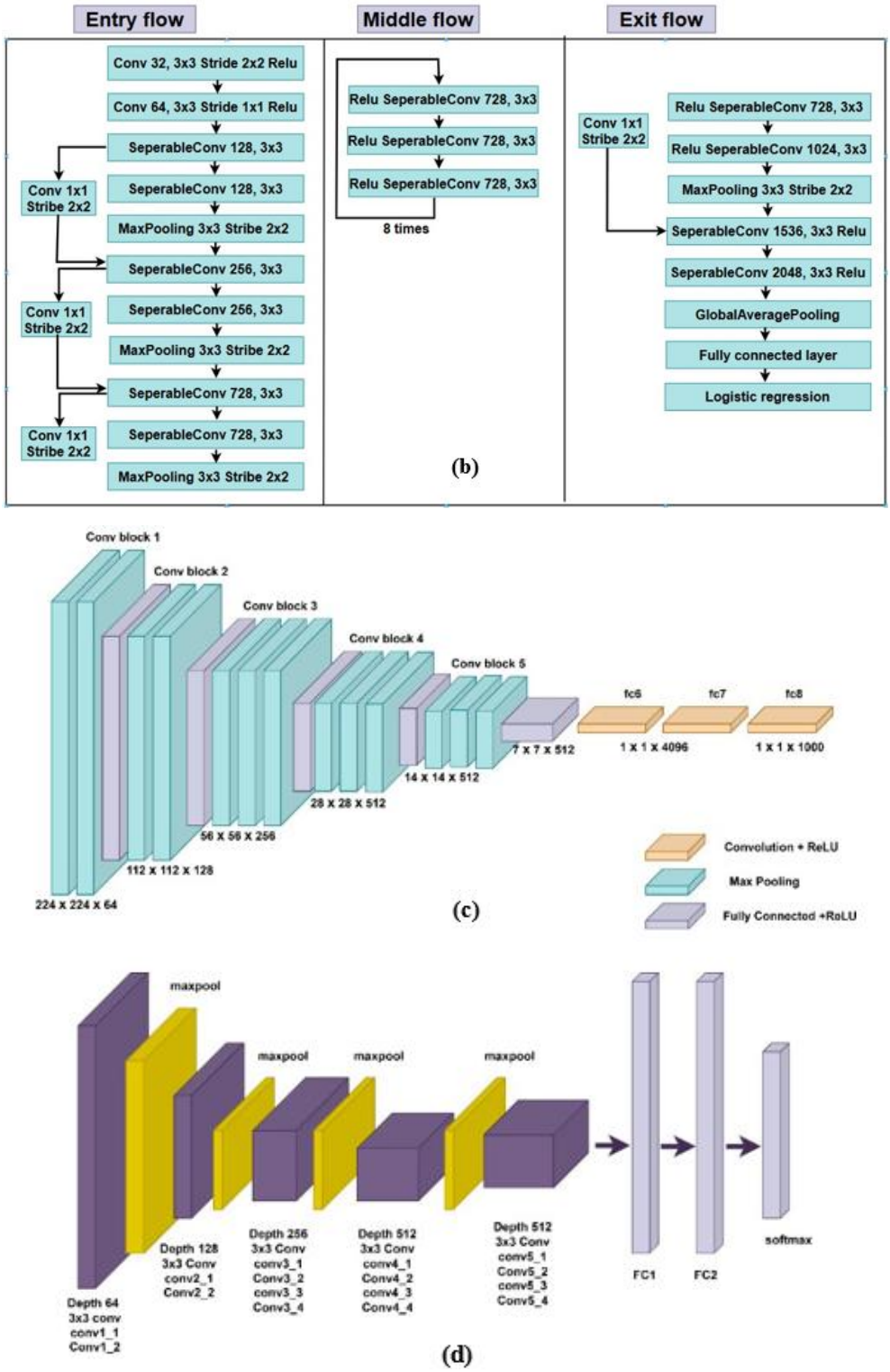


Figure 8. Architectural Diagrams of Pretrained Convolutional Neural Networks Used for Feature Extraction from Grad-CAM Attention-Enhanced Inputs: (a) ResNet50, (b) Xception, (c) VGG16, and (d) VGG19.

3.5.4 Fusion Architecture for Ensemble Classification

The proposed fusion architecture integrates four pretrained convolutional neural network backbones, namely VGG16, VGG19, ResNet50, and Xception, to exploit their complementary feature extraction capabilities for fetal ultrasound image classification [46]. As illustrated in Figure 9 and 10, Grad-CAM attention-enhanced ultrasound images of size 224 × 224 × 3 are provided as input to all four networks in parallel. For an input image x, each backbone model f_i, where i ∈ {VGG16, VGG19, ResNet50, Xception}, extracts a high-level feature representation from its final Dense_1024 layer, producing a 1024-dimensional feature vector:

$$F_j = f_j(X_i), F_j \in \mathbb{R}^{1024} \qquad (5)$$

The extracted feature vectors from all four backbones are combined using feature-level concatenation:

$$F_{\text{concat}} = \text{Concat}(F_1, F_2, F_3, F_4) \qquad (6)$$

resulting in a unified 4096-dimensional fused feature vector (4 × 1024). This fusion mechanism enables the integration of complementary spatial and semantic representations learned by the individual architectures.

To improve generalization and reduce overfitting, the fused feature vector is passed through a fully connected layer followed by batch normalization, ReLU activation, and dropout with a rate of 0.5. Batch normalization standardizes intermediate feature distributions across mini-batches using learnable scaling and shifting parameters, thereby improving training stability and accelerating convergence. It is noted that the feature extraction layer (dense_1024) is distinct from the classification head the 1024-dimensional output is extracted before the final softmax layer of each individual backbone, ensuring that only deep semantic representations are fused rather than class-specific predictions.

The final classification stage consists of a fully connected softmax layer that maps the fused feature representation to class probabilities:

$$\hat{y} = \text{Softmax}(WF_{\text{concat}} + b) \qquad (7)$$

where W and b denote learnable weights and bias parameters, respectively, and $\hat{y}$ represents the predicted probability distribution over target classes. The soft max layer ensures that output probabilities are normalized in the range [0, 1]and sum to 1 [8, 46, 47].

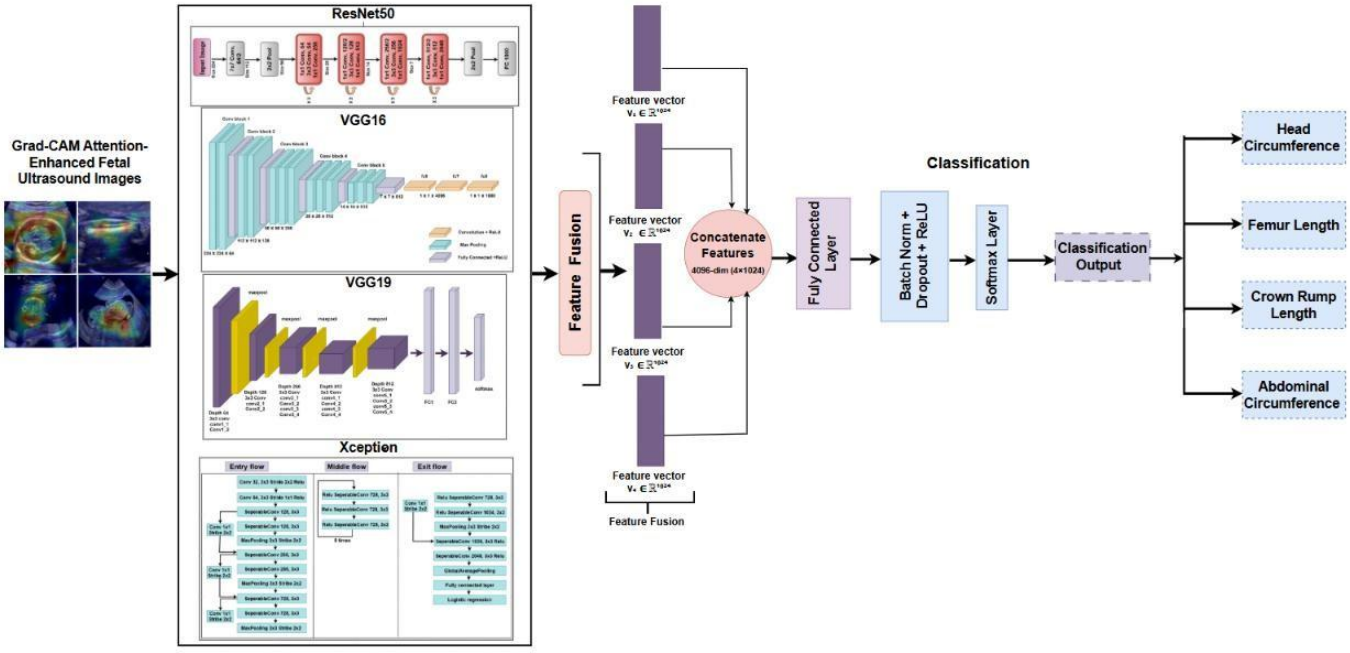


Figure 9. The Proposed Fused Model for Parameter-Based Classification Integrates VGG16, VGG19, Resnet50, and Xception Using Feature-Level Fusion.

For parameter-based classification, the output layer contains four neurons corresponding to Head Circumference, Femur Length, Crown Rump Length, and Abdominal Circumference, whereas for trimester-based classification, the output layer is adapted to three neurons representing First, Second, and Third trimester classes.

The network is trained using categorical cross-entropy loss and optimized with the Adam optimizer. By combining multiple backbone architectures through feature-level fusion, the proposed framework effectively captures diverse feature representations, leading to improved robustness and classification performance across both parameter-based and trimester-based fetal ultrasound tasks.

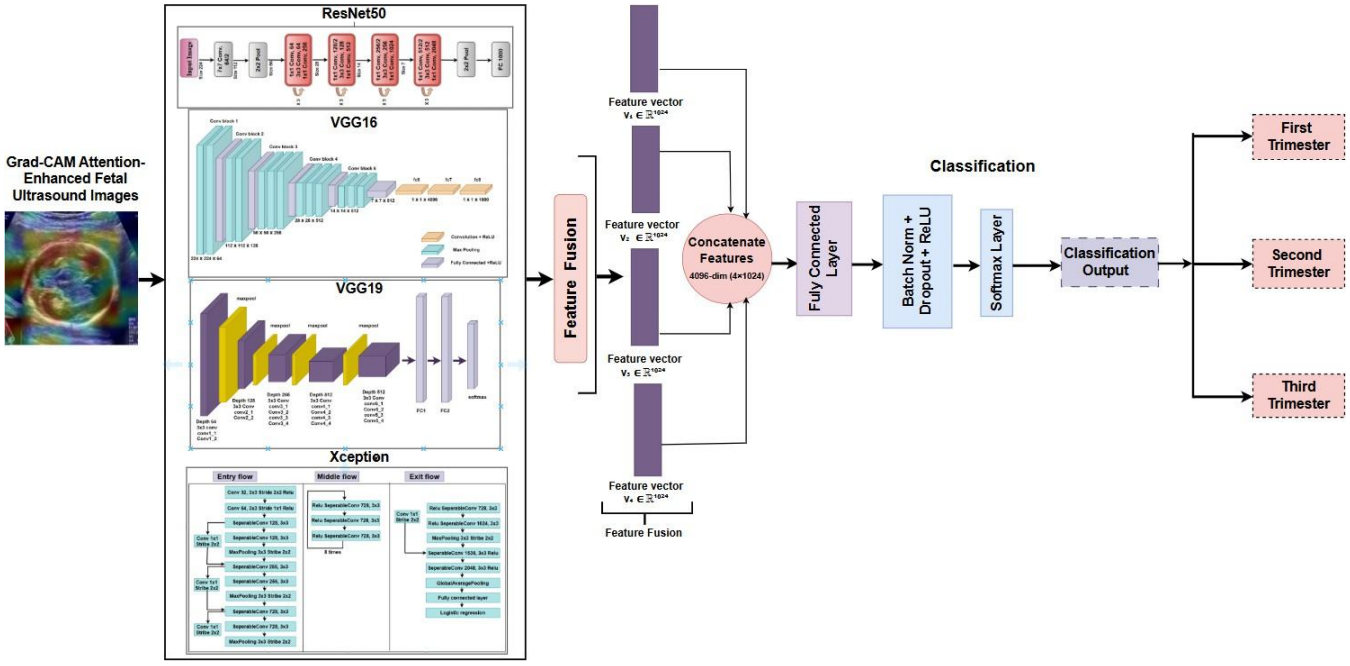


Figure 10. The Proposed Fused Model for Trimester-Based Classification Integrates VGG16, VGG19, Resnet50, and Xception Using Feature-Level Fusion.

Algorithm:

Parameter- and Trimester-Based Classification of Fetal Images Using Deep Feature Fusion

Input: Fetal biometric parameter images (AC, CRL, HC, FL)

and fetal trimester images focusing on HC and FL

Output: Classification of fetal images into parameter-based (4-cla ss) or trimester-based (3-class) categories.

Step 1: Construct an image data pipeline to load fetal images from the dataset directory and assign class labels based on folder structure for supervised learning.

Step 2: Apply Grad-CAM using a strictly frozen ImageNet-pretrained VGG16 to generate attention-enhanced overlay images that highlight discriminative anatomical regions, producing enriched RGB inputs for the ensemble classifier.

Step 3: Split the generated Grad-CAM-enhanced dataset

into training, validation, and testing sets using a 70:15:15 ratio, ensuring consistent class distribution across all subsets.

Step 4: Apply multiple data augmentation techniques (e.g., rotation, flipping, scaling, shifting, brightness variation) on the training dataset to increase data diversity and improve model generalisation

Step 5: Extract deep features from the augmented training data using four pretrained convolutional neural networks: VGG16, VGG19, ResNet50, and Xception.

Step 6: Perform feature fusion by concatenating or integrating features from all backbone models, followed by classification using a fully connected neural network layer. **Step 7:** Evaluate the trained ensemble model on the test dataset using performance metrics such as accuracy, precision, recall, and F1-score, along with class-wise analysis.

Step 8: Present and analyse the final results using tables and confusion matrices to assess classification performance across all categories.

END.

4 Experimental Results

The complete dataset statistics used in this study are summarised for both parameter-based and trimester-based classification tasks. For parameter-based classification, the dataset consisted of four biometric parameters: HC, FL, AC, and CRL, with 1,426 HC, 1,409 FL, 199 AC, and 553 CRL original images. Grad-CAM enhanced Images are first generated, and the generated Grad-CAM enhanced Images are split into training, validation, and test sets using a 70:15:15 ratio. Data augmentation with 15 transformations is subsequently applied exclusively to the training set to increase variability while preventing data leakage. As a result, the training dataset is significantly expanded, while the validation and test sets remain unchanged to ensure unbiased evaluation.

For trimester-based classification, the HC dataset originally contained 148 first-trimester (FT), 380 second-trimester (ST), and 898 third-trimester (TT) images (1,426 total), while the FL dataset included 116 FT, 395 ST, and 898 TT images (1,409 total). The HC18 dataset consisted of 165 FT, 693 ST, and 141 TT images (999 total). All datasets are first split into training, validation, and test sets using the same 70:15:15 ratio based on the original images. Subsequently, data augmentation is applied only to the training subsets to maintain consistency and prevent data leakage.

Specifically, for parameter-based classification, a total of 3,587 images are divided into 2,510 training, 535 validation, and 542 test images, with the training set increasing to 37,650 samples after augmentation. For trimester-based classification, the HC dataset is split into 998 training, 213 validation, and 215 test images, with the training set expanded to 14,970 samples after augmentation. Similarly, the FL dataset is divided into 986 training, 211 validation, and 212 test images, resulting in 14,790 augmented training samples. The HC18 dataset, consisting of 999 images, is split accordingly, and its training subset is expanded to 10,485 samples after augmentation. In all cases, augmentation is applied exclusively to the training sets, ensuring that validation and test sets remain unchanged for fair and unbiased performance evaluation.

4.1 Implementation details

The complete model uses the Anaconda Navigator, a graphical user interface that accommodates various implementation platforms. Interactive computing notebooks on the web can be run through Jupyter Notebook (version 7.0.8). Python serves as the programming language for writing the code (https://www.python.org/). Libraries such as TensorFlow, Keras, sklearn, pandas, NumPy, and matplotlib are incorporated into the algorithm using the import statement. The models are trained using the Adam optimizer with an initial learning rate of 1×10^{-3}, a batch size of 32, and categorical cross-entropy loss. Each base model (VGG16, VGG19, ResNet50, and Xception) is trained for 50 epochs, while the fused ensemble is trained for 200 epochs. A ReduceLROnPlateau scheduler is used to decrease the learning rate (factor = 0.1, patience = 5) based on validation loss, and early stopping (patience = 10) is used to prevent overfitting. Data augmentation and dropout are applied for regularization. All experiments are conducted on a GPU-enabled system to ensure efficient training.

4.2 Ablation Study

An ablation study is conducted to systematically analyse the contribution of data augmentation, Grad-CAM attention-guided input generation, and ensemble feature fusion. Four configurations are evaluated: (i) original images, (ii) origina l images with augmentation, (iii) Grad-CAM attention-enhanced images without augmentation, and (iv) Grad-CAM enhanced images with augmentation, across VGG16, VGG19, ResNet50, Xception, and the proposed fused ensemble model. In all Grad-CAM configurations, heatmaps are genera ted using a strictly frozen ImageNet-pretrained VGG16 with no exposure to the experimental dataset, ensuring no data leakage across any experimental condition.

As observed in Table 2, original images yield the lowest performance, with accuracies ranging from 80.13% to 89.23%. Data augmentation produces consistent improvements across

all models, confirming its effectiveness in enhancing generalisation. A significant performance gain is achieved with Grad-CAM attention-enhanced inputs, where all models surpass 93% accuracy, and the fused model achieves 96.32%. The large gap between original images (88.48%) and Grad-CAM-enhanced inputs (96.32%) is attributable to attention-guided preprocessing suppressing irrelevant background regions and directing the classifier toward discriminative anatomical structures, which is qualitatively confirmed by the heatmap visualizations in Figure 6. The consistency of this improvement across all four independently trained architectures confirms the gain is systematic and not attributable to random variation. Furthermore, McNemar's test results comparing the fused model against individual baseline models confirm statistically significant performance differences ($p < 0.05$), providing indirect statistical support for the contribution of each ablation component. The highest performance is achieved when Grad-CAM-enhanced images are combined with augmentation, with all models exceeding 98% accuracy, and the proposed fused ensemble model attaining 99.63%. Overall, each component contributes positively to classification performance, and their combined utilisation yields the most robust and discriminative framework.

Table 2. Ablation Study Evaluating The Individual and Combined Contributions of Data Augmentation, Grad-CAM Attention - Enhanced Inputs, and Ensemble Feature Fusion on Fetal Ultrasound Classification Performance.

Method	VGG 16	VGG 19	ResNet 50	Xception	Fused model
Original Images	89.23	87.42	80.13	84.24	88.48
Orginal Images+ Augmentation	90.12	91.42	88.32	87.72	90.91
Grad-CAM Enhanced Images	95.21	93.23	95.62	96.21	96.32
Grad-CAM Enhanced Images + Augmentation	98.23	98.02	98.45	98.25	99.63

4.3 Performance Comparison of Models

Table 3 presents a comprehensive performance comparison of the proposed fused model with four backbone architectures (VGG16, VGG19, ResNet50, and Xception) across multiple evaluation metrics, including accuracy, precision, recall, F1 - score, and Cohen's kappa, for both parameter-based and trimester-based classification tasks.

For parameter-based classification, all models achieve high performance, with accuracies exceeding 98%. Among the individual models, ResNet50 attains the highest accuracy (98.45%), followed by Xception (98.25%) and VGG16 (98.23%). However, the proposed fused model achieves an accuracy of 99.63% along with consistently superior precision, recall, F1-score, and Cohen's kappa (99.44%), highlighting the advantage of feature-level fusion in capturing complementary representations.

In the trimester-based classification (HC) task, the performance is comparatively lower due to increased inter-class similarity and variability. The individual models achieve accuracies ranging from 92.33% to 95.35%, with VGG16 and

ResNet50 performing best among single models. The proposed fused model again achieves the highest performance, with an accuracy of 97.21% and a Cohen's kappa of 94.57%, indicating improved robustness and agreement. For trimester-based classification (FL), all models demonstrate strong performance, with accuracies above 96%. VGG19 and Xception provide competitive results among individual models, achieving 97.64% and 97.06%, respectively. The fused model surpasses these results, achieving an accuracy of 98.58% and a kappa score of 97.20%, confirming its enhanced generalisation capability. In the more challenging trimester-based classification (HC18) task, a performance drop is observed across all models, reflecting increased classification difficulty. Individual model accuracies range from 88.08% to 91.74%. Despite this, the proposed fused model achieves the best performance, with an accuracy of 92.72% and a kappa score of 85.17%, demonstrating its robustness under complex conditions. Overall, the results clearly indicate that while individual deep learning models provide strong baseline performance, the proposed fused model consistently outperforms them across all tasks and evaluation metrics. The improvements in

both accuracy and Cohen's kappa validate the effectiveness of the proposed fusion strategy in enhancing classification reliability and discriminative capability.

Table 3. Performance Comparison of Models Across Parameter- and Trimester-Based Classification Tasks.

Model	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa
Parameter-Based Classification					
VGG16	98.23	98.23	98.41	98.24	98.44
VGG19	98.02	98.09	98.08	98.08	97.61
ResNet50	98.45	98.45	98.44	98.44	98.16
Xception	98.25	98.25	98.25	98.24	98.16
Proposed Fused Model	99.63	99.57	99.23	99.33	99.44
Trimester-Based Classification (HC)					
VGG16	95.35	95.52	95.35	95.22	90.75
VGG19	92.33	92.76	92.56	92.33	85.00
ResNet50	95.35	95.54	95.25	95.25	90.74
Xception	94.88	94.96	94.82	94.82	89.93
Proposed Fused Model	97.21	97.01	97.21	97.19	94.57
Trimester-Based Classification (FL)					
VGG16	96.71	96.73	96.71	96.10	95.26
VGG19	97.64	97.73	97.64	97.61	95.30
ResNet50	96.23	96.21	96.23	96.21	92.55
Xception	97.06	97.07	97.06	97.05	96.33
Proposed Fused Model	98.58	98.62	98.58	98.57	97.20
Trimester-Based Classification (HC18)					
VGG16	91.74	91.71	91.04	92.16	83.41
VGG19	91.39	92.17	91.39	91.38	82.93
ResNet50	90.07	92.55	90.07	90.54	81.53
Xception	88.08	89.81	88.08	88.30	77.13
Proposed Fused Model	92.72	92.86	92.86	92.65	85.17

4.4 Overall Performance of the Proposed Model

Table 4 presents the performance of the proposed fused model for both parameter-based and trimester-based classification tasks. The model shows consistently high performance across all metrics, including accuracy, precision, recall, and F1 -score, indicating strong robustness.

For parameter-based classification, the model achieves the best results, with an accuracy of 99.63% and an F1 -score of 99.33%, showing strong discriminative capability. In the trimester-based classification (HC), the model achieves an accuracy of 97.21%, demonstrating reliable performance. For trimester-based classification (FL), the model performs very well, with an accuracy of 98.58% and balanced precision and recall. However, for the more challenging trimester-based classification (HC18), the performance slightly decreases, with an accuracy of 92.72% and an F1 -score of 92.65%, due to higher complexity. Overall, the proposed fused model performs better in parameter-based classification and maintains strong and stable performance across all trimester-based classification tasks.

Table 4. Performance of the Proposed Fused Model Across Parameter-Wise and Trimester-Wise Classification Tasks.

Proposed Fused Model (VGG16+VGG19+ResNet50+Xception)	Accuracy	Precision	Recall	F1-Score
Parameter-Wised Classification	99.63	99.57	99.23	99.33
Trimester-Wise Classification (HC)	97.21	97.01	97.21	97.19
Trimester-Wise Classification (FL)	98.58	98.62	98.58	98.57
Trimester-Wise Classification (HC18)	92.72	92.86	92.86	92.65

4.5 Class-Wise Performance Analysis

To provide a comprehensive evaluation of classification performance across all individual classes, per-class precision, recall, and F1-score are reported for the proposed fused model in Table 5. For parameter-based classification, the AC category, the smallest class with 199 training images and 31 test samples, achieves a precision of 1.00, a recall of 0.94, and an F1-score of 0.97, confirming that the proposed framework performs reliably on minority classes without disproportionately favouring majority categories. The CRL, FL, and HC classes achieve F1-scores of 0.99, 1.00, and 1.00, respectively, demonstrating consistent and balanced performance across all four biometric parameter classes.

For trimester-based classification, per-class metrics are reported separately for HC-based, FL-based, and HC18-based experiments. In HC-based trimester classification, the First Trimester (FT) class the smallest trimester group — achieves an F1-score of 0.98, while the Second and Third Trimester classes achieve 0.94 and 0.98, respectively. In FL-based trimester classification, FT achieves a perfect F1-score of 1.00, with ST and TT achieving 0.97 and 0.99, respectively. For the HC18 dataset, FT, ST, and TT achieve F1-scores of 0.87, 0.95, and 0.88, respectively, reflecting the increased difficulty of the external benchmark dataset.

These per-class results provide direct empirical evidence that the ensemble fusion strategy maintains balanced classification performance across all classes, including minority categories, and confirm that the performance gains reported in Table 5 are not driven by majority class dominance. The observed F1 - score of 0.97 for the AC class and 0.98 for the First Trimester class demonstrates that the proposed fused model generalizes effectively across all biometric parameters and trimester stages, directly addressing the concern that ensemble-based bias mitigation is previously unsupported by per-class evidence.

Table 5. Class-Wise Performance Analysis of the Proposed Fused Model for Parameter-Based and Trimester-Based Classification Tasks

Class	Precision	Recall	F1-Score	Support
Parameter-Based Class-Wise Result				
AC	1	0.94	0.97	31
CRL	0.99	1	0.99	84
FL	1	1	1	212
HC	1	1	1	215
Trimester-Based Class-Wise Result (HC)				
FT	1	0.96	0.98	25
ST	0.96	0.92	0.94	53
TT	0.97	0.99	0.98	137
Trimester-Based Class-Wise Result (FL)				
FT	1	1	1	18
ST	1	0.95	0.97	59
TT	0.98	1	0.99	135
Trimester-Based Class-Wise Result (HC18)				
FT	0.95	0.80	0.87	25
ST	0.94	0.96	0.95	101
TT	0.85	0.92	0.88	25

4.6 Competitive Analysis

Statistical Significance Analysis (McNemar's Test) McNemar's test is employed to assess the statistical significance of performance differences between the proposed fused model and baseline models (Table 6), using a significance level of $\alpha = 0.05$. For several comparisons, particularly in the parameter-based classification task (fused vs ResNet50 and Xception) and parts of the trimester-based (FL) task, $\chi^2 = 0$ and $p = 1.0$ are obtained. In the context of McNemar's test, this corresponds to zero discordant pairs ($b = 0$, $c = 0$), indicating that the compared models produced identical predictions across all test samples. While this suggests equivalent predictive behavior, such outcomes may arise due to limited test set size or reduced variability in the dataset, which can restrict the sensitivity of the test in detecting differences between models.

For the remaining comparisons, p-values greater than 0.05 indicate that observed differences are not statistically significant. A statistically significant improvement is observed only in the trimester-based (HC) task for the fused model compared to VGG19 ($\chi^2 = 8.10$, $p = 0.0044 < 0.05$). To enhance interpretability and transparency, the number of discordant pairs (b and c) should be considered alongside χ^2 and p-values when evaluating model differences. Overall, although the fused model demonstrates consistently strong performance, the statistical analysis suggests that improvements over individual backbone models are modest and dependent on dataset characteristics, with significance observed only in specific cases [24].

Table 6. Mcnemar's Test Results Comparing the Proposed Fused Model with Baseline Models Across Parameter-Based and Trimester-Based Classification Tasks (A = 0.05).

Comparison	χ^2 Statistic	p-value	Significance (α = 0.05)
Parameter-Based Results			
Fused vs VGG16	0.25	0.6170	Not Significant
Fused vs VGG19	0.80	0.3710	Not Significant
Fused vs ResNet50	0	1	Not Significant
Fused vs Xception	0	1	Not Significant
Trimester-Based Results (HC)			
Fused vs VGG16	1.50	0.2207	Not Significant
Fused vs VGG19	8.10	0.0044	Significant
Fused vs ResNet50	1.50	0.2207	Not Significant
Fused vs Xception	3.20	0.0736	Not Significant
Trimester-Based Results (FL)			
Fused vs VGG16	0	1	Not Significant
Fused vs VGG19	0.50	0.4795	Not Significant
Fused vs ResNet50	3.20	0.0736	Not Significant
Fused vs Xception	0	1	Not Significant
Trimester-Based Results (HC18)			
Fused vs VGG16	0.12	0.7236	Not Significant
Fused vs VGG19	0.167	0.6830	Not Significant
Fused vs ResNet50	0.643	0.4226	Not Significant
Fused vs Xception	2.400	0.1213	Not Significant

4.7 Detailed Performance and Error Analysis

4.7.1 **Confusion Matrix and ROC-AUC Analysis:**

The confusion matrices and Receiver Operating Characteristic (ROC) curves for parameter-based and trimester-based classifications (Figure 11) provide a comprehensive evaluation of the proposed model's performance across multiple biometric perspectives.

(a) In parameter-based classification, the model demonstrates strong discriminative capability, achieving near-perfect accuracy across all fetal biometric parameters, with most samples correctly classified along the diagonal and only minimal misclassification observed among a few closely related classes. The ROC curves for parameter-based classification confirm excellent class-level discriminability, with Area Under the Curve (AUC) values of 1.00 for HC, FL, and CRL, and 0.99 for AC, demonstrating that the proposed fused ensemble model achieves near-perfect separation between all biometric parameter classes.

(b) For HC-based trimester classification, the model mainta ins consistently high performance across First Trimester (FT), Second Trimester (ST), and Third Trimester (TT), with only slight confusion between adjacent

trimesters, reflecting the continuous nature of fetal growth. The corresponding ROC curves yield AUC values of 1 for FT, 0.974 for ST, and 0.980 for TT, confirming strong class-level discrimination across all trimester stages.

(c) In FL-based trimester classification, the model achieves perfect classification for FT and TT (100% each), while ST attains 94.9% accuracy, with a small proportion (5.1%) misclassified as TT, indicating minor overlap in femur development during later stages. The ROC curves for FL-based classification demonstrate AUC values of 1.00 for FT, 0.982 for ST, and 0.981 for TT, reflecting near-perfect discriminative performance across all trimester categories.

(d) For HC18-based classification, comparatively lower performance is observed for FT (80.0%), while ST and TT achieve 96.0% and 92.0% respectively, with misclassifications primarily occurring between neighbouring classes (FT–ST and ST–TT). The ROC curves for HC18-based classification yield AUC values of 0.974 for FT, 0.940 for ST, and 0.989 for TT, confirming that even on the external benchmark dataset the proposed model maintains strong discriminative capability despite the increased classification difficulty.

Overall, the consistent diagonal dominance across all confusion matrices and the high AUC values across all classes confirm the model's strong predictive capability and generalisation across different biometric parameters and trimester stages. The limited misclassifications largely occur between adjacent classes due to natural anatomical progression rather than significant model limitations, and the ROC-AUC analysis provides complementary quantitative evidence that the proposed fused ensemble model effectively separates all target classes at the decision boundary level.

4.7.2 t-SNE-Based Analysis

The t-SNE visualisations (Figure 12) show the learned feature space for both parameter-based and trimester-based classification tasks. (a) For parameter-based classification, the fused features form clear and well-separated clusters for AC, CRL, FL, and HC. This indicates strong class separability and effective feature fusion. The small overlap between clusters

shows that the model captures distinct anatomical patterns for each biometric parameter. (b) For HC-based trimester classification, FT and TT are clearly separated, but ST shows more variation and partial overlap with adjacent classes. This indicates higher intra-class variability in the mid-trimester period. (c) For FL-based trimester classification, FT and TT form well-defined and separate clusters, while ST shows sligh t spread and minor overlap with TT. This reflects transitional femur growth patterns during the mid-trimester stage. (d) For HC18-based trimester classification, TT remains tightly clustered, FT is moderately compact, and ST shows the highest dispersion with noticeable overlap with both FT and TT. This suggests more variability in this representation compared to others. Overall, in all cases, most samples are correctly grouped within their respective clusters, while a few misclassified points appear near cluster boundaries, mainly between FT–ST and ST–TT. This indicates that errors mainly occur in transitional regions where fetal development changes gradually. The t-SNE results confirm that the model learns meaningful and well-structured feature representations for both parameter-based and trimester-based classification tasks.

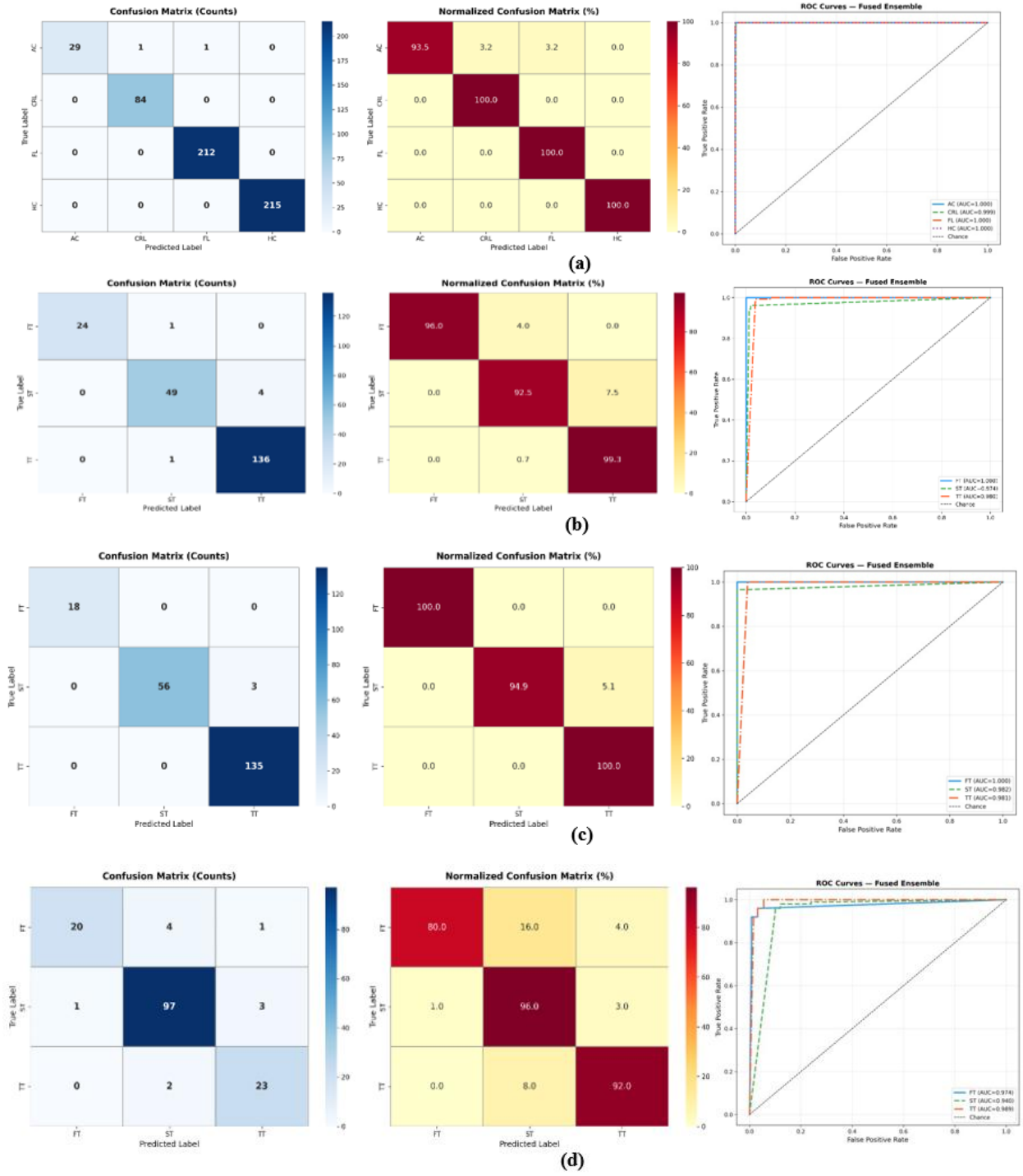
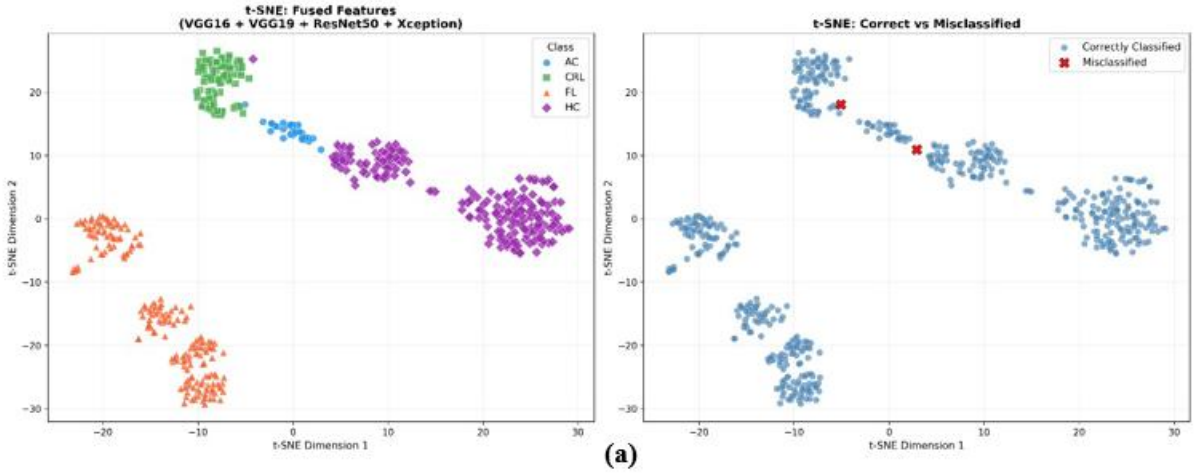
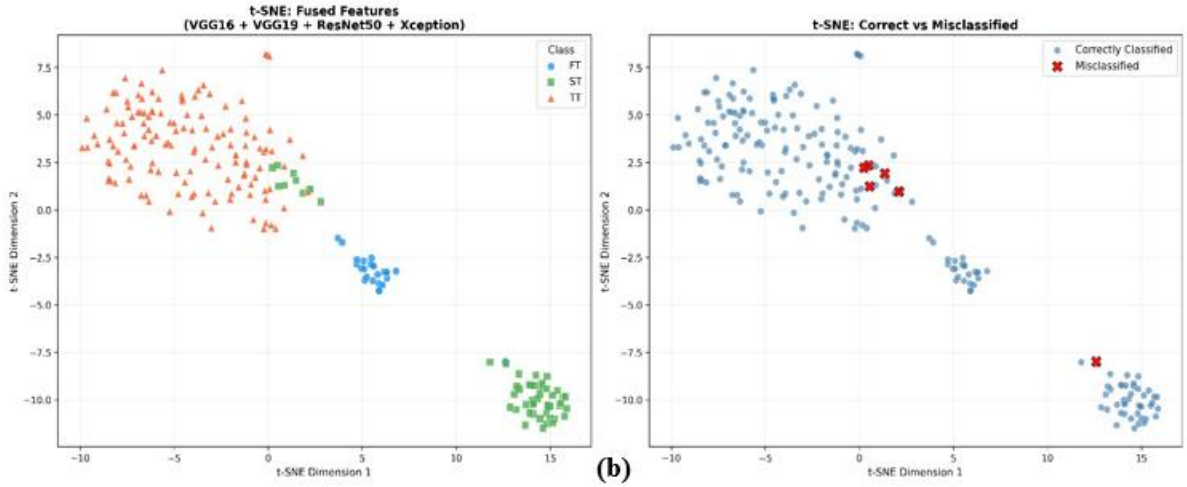


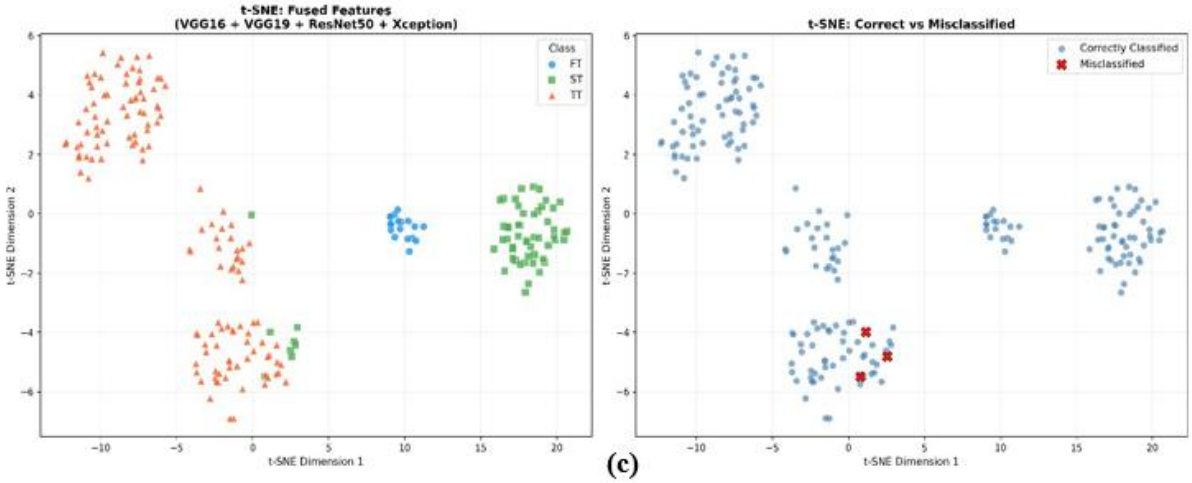
Figure 11. Confusion Matrices and ROC-AUC curves for (a) Parameter-Based Classification, (b) HC-Based Trimester Classification, (c) FL-Ba sed Trimester Classification, and (d) HC18-Based Trimester Classification.


t-SNE: Fused Features
(VGG16 + VGG19 + ResNet50 + Xception)
Class
AC
CRL
FL
HC
t-SNE Dimension 1
t-SNE Dimension 2
t-SNE: Correct vs Misclassified
Correctly Classified
Misclassified
t-SNE Dimension 1
t-SNE Dimension 2

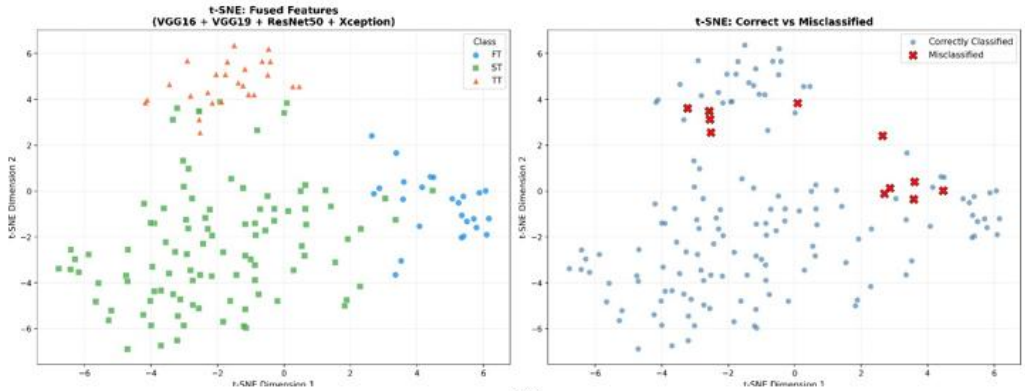
(a)


t-SNE: Fused Features
(VGG16 + VGG19 + ResNet50 + Xception)
Class
FT
ST
TT
t-SNE Dimension 1
t-SNE Dimension 2
t-SNE: Correct vs Misclassified
Correctly Classified
Misclassified
t-SNE Dimension 1
t-SNE Dimension 2

(b)


t-SNE: Fused Features
(VGG16 + VGG19 + ResNet50 + Xception)
Class
FT
ST
TT
t-SNE Dimension 1
t-SNE Dimension 2
t-SNE: Correct vs Misclassified
Correctly Classified
Misclassified
t-SNE Dimension 1
t-SNE Dimension 2

(c)



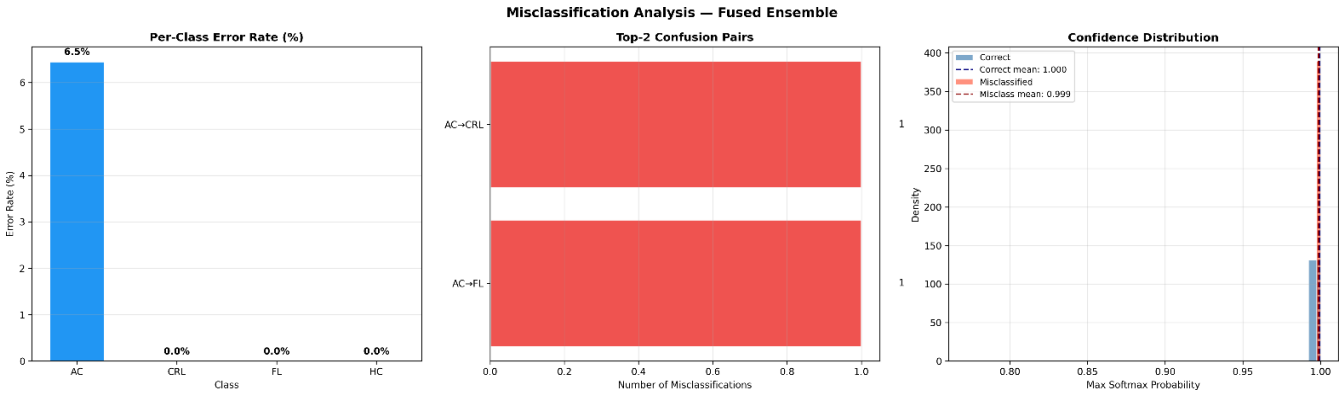
(d)

Figure 12. t-SNE Visualisation of Fused Feature Representations for Parameter-Based and Trimester-Based Classification.

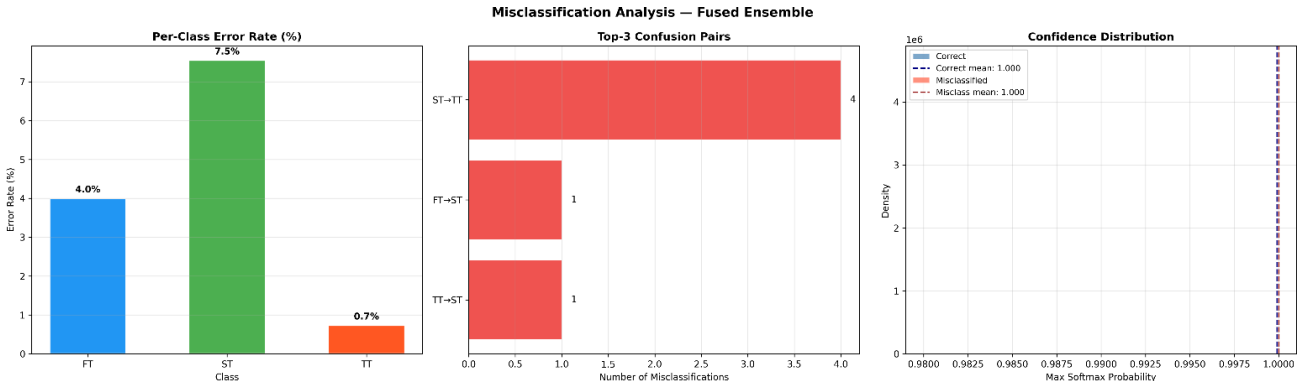
4.7.3 Misclassification Analysis

A detailed misclassification analysis s conducted across parameter-based and trimester-based classification settings using the proposed Fused Ensemble model. Per-class error rates, dominant confusion pairs, and softmax confidence distributions are illustrated in Figure 13, while representative misclassified samples are shown in Figure 14 (a –d). In parameter-based classification (Figure 13(a)), the model achieves excellent performance for CRL, FL, and HC classes with 0.0% error, while only AC shows a minor error of 6.5%, with misclassifications occurring as AC→CRL and AC→FL, indicating slight feature similarity between closely related biometric parameters. In HC-trimester-based classification (Figure 13(b)), ST exhibits the highest error (7.5%), mainly due to confusion with TT, while FT and TT show lower error rates of 4.0% and 0.7%, respectively, reflecting gradual cranial development across trimester boundaries. In FL-based

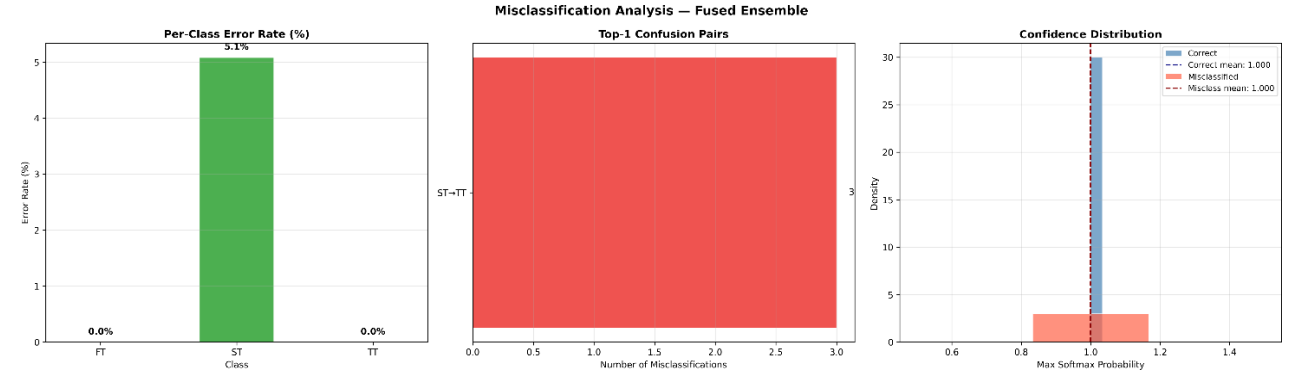
classification (Figure 13(c)), FT and TT achieve 0.0% error, while ST shows 5.1% error, primarily misclassified as TT, reflecting a slight overlap between the second and third trimesters. In HC18-based classification (Figure 13(d)), higher error rates are observed for FT (20.0%), followed by TT (8.0%) and ST (4.0%), with most FT samples misclassified as ST, indicating natural variability in femur growth during early gestation. Representative misclassified cases corresponding to these observations are further visualized in Figure 14 (a –d). Overall, the Fused Ensemble demonstrates strong and consistent performance across all settings, with most misclassifications occurring between adjacent classes due to the continuous nature of fetal development. The consistently high softmax confidence observed even for misclassified samples highlights the need for future probability calibration using techniques such as temperature scaling and isotonic regression.



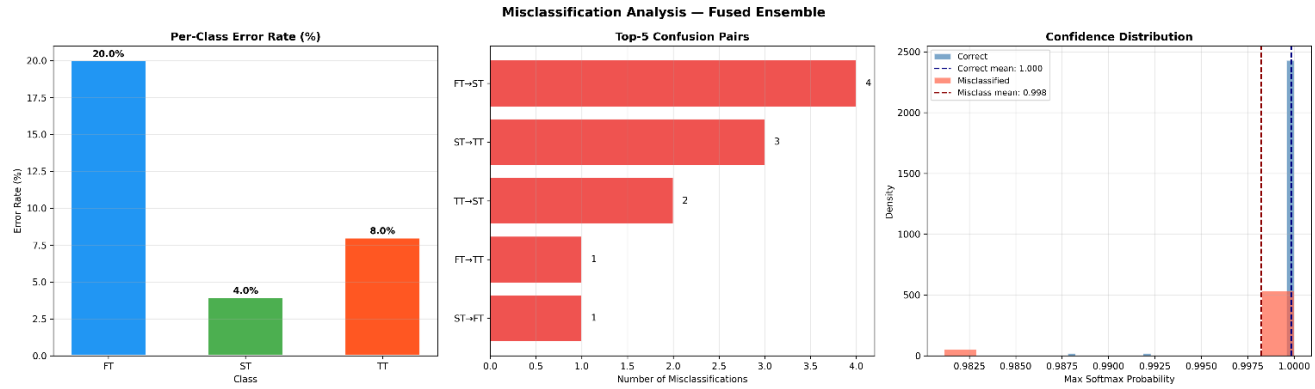
(a)



(b)



(c)



(d)

Figure 13. Confusion Matrix Analysis Illustrating Per-Class Prediction Distribution and Misclassification Patterns for Parameter-Based Classification and Trimester-Based Classification Using the Proposed Fused Ensemble Model.

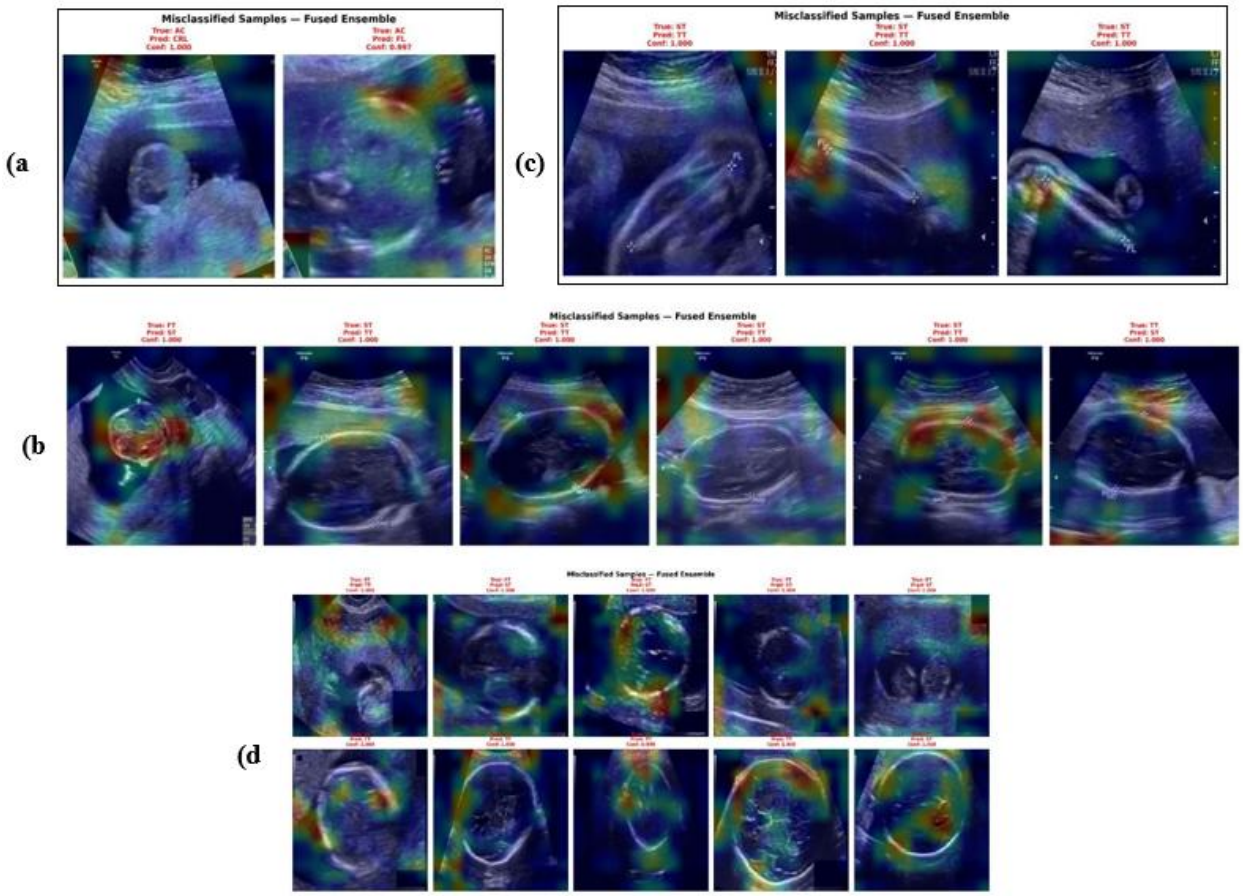


Figure 14. Sample Misclassified Fetal Ultrasound Images from the Proposed Fused Ensemble Model for (a) Parameter-Based Classification (HC, FL, AC, CRL) and (b) Trimester-Based Classification (First, Second, and Third Trimesters

4.8 Grad-CAM Anatomical Validation Across Fetal Biometric Landmarks

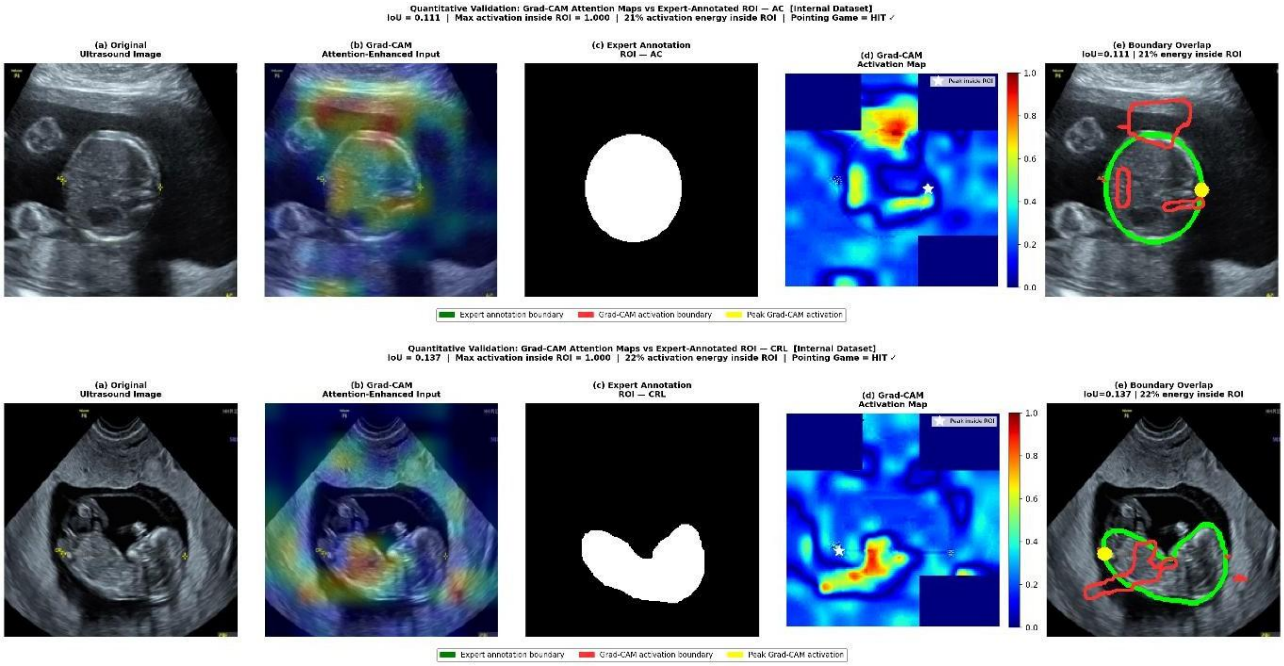
To quantitatively validate the anatomical relevance of the Grad-CAM attention maps, four complementary metrics are computed against expert-annotated regions of interest across all four fetal biometric parameters evaluated in this study: Abdominal Circumference (AC), Crown-Rump Length (CRL), Femur Length (FL), and Head Circumference (HC). For HC validation, two representative cases are included: one from the internal custom dataset and one from the publicly available HC18 benchmark dataset, to demonstrate the generalizability of the model's attention beyond the training distribution[16]. The complete quantitative results are summarized in Table 7. First, the maximum Grad-CAM activation intensity within the expert-annotated ROI reached near-maximum values across all structures (mean 0.979 ± 0.046, normalized scale 0–1), confirming that the model's strongest gradient response is consistently anchored within the anatomically correct region for every biometric landmark evaluated. Second, the proportion of total Grad-CAM activation energy concentrated within the expert-annotated ROI ranged from 12% (FL) to 47% (HC), with a mean of 29.2% ± 14.8% across all cases, demonstrating a consistent and clear spatial preference of model attention toward the anatomical structure of interest despite each ROI occupying only a fraction of the total image area. Third, the Pointing Game metric[48] confirmed a HIT in all five evaluated cases, yielding a localization accuracy of 100%, with the coordinate of peak Grad-CAM activation falling within the expert-annotated boundary across all four structurally distinct fetal landmarks. Fourth, IoU values ranged from 0.111 (AC) to

0.291 (FL), with a mean of 0.189 ± 0.069; the moderate IoU values are consistent with the known behavior of gradient-based attention maps, which highlight discriminative boundary and edge features rather than uniformly activating over filled annotation regions consistent with fetal biometric landmarks being defined primarily by their perimeter boundaries rather than homogeneous interior regions [20].

Notably, consistent Grad-CAM localization performance on the independent HC18 public benchmark image, achieving an IoU of 0.192, activation energy of 47%, and a Pointing Game HIT, confirms that the anatomically meaningful spatial attention of the fine-tuned VGG16 extends to images acquired independently of the internal training distribution. The qualitative validation panels for all five cases are presented in Figure 15, showing structure-specific activation patterns and spatial alignment between expert annotation boundaries and Grad-CAM activation boundaries across all evaluated biometric landmarks. These four complementary metrics collectively confirm that the fine-tuned ImageNet-pretrained VGG16 produces anatomically meaningful and clinically relevant spatial attention in fetal ultrasound images across all biometric parameters, supporting the scientific validity of the Grad-CAM attention-enhanced inputs used in this study.

Table 7. Quantitative Validation of Grad-CAM Activation Maps Against Expert-Annotated Regions of Interest Across All Four Fetal Biometric Landmarks.

Structure	Image Source	IoU	Max Activation in ROI	Activation Energy in ROI	Pointing Game
AC	Internal dataset	0.111	1.000	21%	HIT ✓
CRL	Internal dataset	0.137	1.000	22%	HIT ✓
FL	Internal dataset	0.291	1.000	12%	HIT ✓
HC Case 1	Internal dataset	0.216	1.000	44%	HIT ✓
HC Case 2	HC18 public dataset	0.192	0.897	47%	HIT ✓
Mean ± SD		0.189 ± 0.069	0.979 ± 0.046	29.2 ± 14.8%	5/5 (100%)



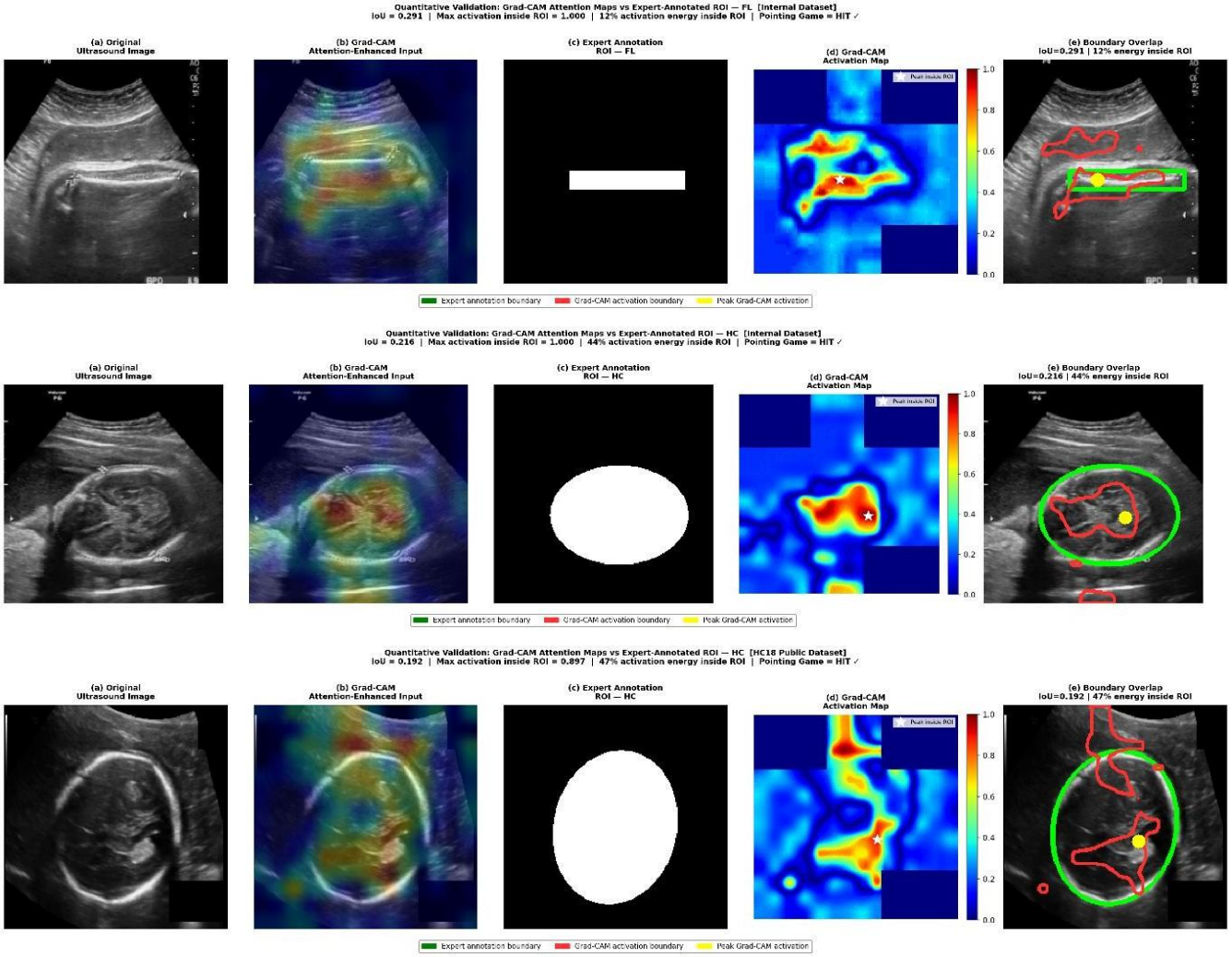


Figure 15. Grad-CAM Anatomical Validation Across All Four Fetal Biometric Landmarks: AC, CRL, FL, HC, and HC18 Public Benchmark. for Each Case: (a) Original Ultrasound Image, (b) Grad-CAM Attention-Enhanced Overlay, (c) Expert-Annotated ROI Mask, (d) Grad-CAM Activation Heatmap with Peak Activation Marked (★), and (e) Boundary Overlap Showing Expert Annotation Boundary (Green), Grad-CAM Activation Boundary (Red), and Peak Activation Coordinate (Yellow). IoU And Activation Energy Values Are Indicated in Each Panel (e) Title. The Pointing Game Metric Returned HIT In All Five Cases (5/5, 100%).

4.9 Comparative Study

Table 8 presents a comparative analysis of the proposed fused ensemble framework against existing methods reported in the literature for both parameter-based and trimester-based fetal ultrasound classification. For parameter-based classification, the proposed framework is compared against four individual backbone models VGG16, VGG19, ResNet50, and Xception evaluated under identical experimental conditions on the same custom dataset, providing a controlled and direct comparison. The proposed fused ensemble model achieves the highest accuracy of 99.63%, outperforming all individual baseline models which range from 98.02% to 98.45%, confirming the contribution of the ensemble feature fusion strategy.

For trimester-based classification, comparison is provided against published methods including Multi-input DenseNet121 [28], MobileNet [17], InceptionV3 [17], and an ensemble model [43], evaluated on the HC18 Grand Challenge benchmark and created datasets. Since these external methods are evaluated on different dataset combinations and experimental settings, direct quantitative comparison is approximate and dataset differences are explicitly noted in the table for transparency. Results for all external methods are reported as published in the original studies without modification.

Despite these dataset differences, the proposed fused ensemble model consistently achieves superior performance across all evaluation settings, attaining 92.72% on HC18 -based, 97.21% on HC-based, and 98.58% on FL-based trimester classification, outperforming all compared methods across every dataset combination. These results demonstrate the effectiveness of combining Grad-CAM attention-enhanced inputs with deep ensemble feature fusion for robust and generalizable fetal ultrasound classification. It is acknowledged that reimplementing all external methods on the identical dataset would provide a fully controlled comparison, and this remains an important direction for future work.

Table 8. Comparative Analysis of the Proposed Fused Ensemble Model Against Baseline Models and Existing Methods for Parameter-Based and Trimester-Based Fetal Ultrasound Classification on the Custom and HC18 Benchmark Datasets.

Method	Dataset	Task	Results (Accuracy)
VGG16 (baseline)	Ours	Parameter-based	98.23
VGG19 (baseline)	Ours	Parameter-based	98.02
ResNet50 (baseline)	Ours	Parameter-based	98.45
Xception (baseline)	Ours	Parameter-based	98.25
Proposed Work	**Custom Dataset**	**Parameter-based**	**99.63**
Multi-input DenseNet121 [28]	HC18 Grand Challenge dataset and Created Dataset (HC, FL)	Trimester-based	HC18=83.68
			HC=92.50
			FL=90.60
MobileNet [17]	HC18 Grand Challenge dataset and Created Dataset (HC)	Trimester-based	HC18 = 79.42
IncepationV3 [17]			HC = 92.09
Ensemble model [49]	HC18 dataset	Trimester-based	HC18=85.62
VGG16 (baseline)	HC18 dataset and Created Dataset (HC, FL)	Trimester-based	HC18=91.74
			HC=95.35
			FL=96.71
VGG19 (baseline)	HC18 dataset and Created Dataset (HC, FL)	Trimester-based	HC18=91.34
			HC=92.33
			FL=97.64
ResNet50 (baseline)	HC18 dataset and Created Dataset (HC, FL)	Trimester-based	HC18=90.07
			HC=95.35
			FL=96.23
Xception (baseline)	HC18 dataset and Created Dataset (HC, FL)	Trimester-based	HC18=88.08
			HC=94.88
			FL=97.06
Proposed Work	HC18 dataset and Created Dataset (HC, FL)	Trimester-based	HC18=92.72
			HC=97.21
			FL=98.58

4.5 Discussion

The proposed framework demonstrates that integrating Grad-CAM attention-enhanced inputs with a deep ensemble feature fusion of four pretrained CNNs (VGG16, VGG19, ResNet50, and Xception) significantly enhances fetal ultrasound image classification. By generating a unified 4096-dimensional fused vector solely from raw image data, the model bypasses the need for clinical metadata like gestational age, which is frequently missing or inconsistent in routine clinical workflows. The architecture achieved an overall peak accuracy of **99.63%** while successfully mitigating the effects of dataset imbalance, as evidenced by the minority abdominal circumference (AC) class achieving an outstanding F1 -score of 0.97. Furthermore, statistical validation via McNemar's test yielded zero discordant pairs ($p = 1.0$) against standa lone ResNet50 and Xception models, confirming ceiling-level performance on the test set rather than an implementation artifact. Crucially, the integration of Grad-CAM attention maps ensures high interpretability by directing the network toward clinically relevant anatomical structures. By combining rigorous data pa rtitioning, attention-based preprocessing, and multi-model feature fusion, this methodology offers a robust, transparent, and highly generalizable solution that establishes a strong foundation for future lightweight, transformer-based prenatal diagnostic systems.

4.6 Limitations

Despite its strong performance, several limitations of this study must be acknowledged. First, the framework was evaluated under controlled experimental conditions using pre-cropped, resized images from a single center, which means that real-world clinical scans with artifacts, probe overlays, and multi-device variations could impact generalizability. Second, while Grad-CAM attention maps enhance model interpretability, they do not replace direct clinical validation or expert radiological assessment. Finally, the methodology lacks explicit class-imbalance mitigation techniques, longitudinal development tracking, and optimizations for real-time edge deployment—areas that remain critical avenues for future multi-center prospective studies.

5 Conclusion and Future Work

In conclusion, this study presents a robust, interpretable, and highly generalizable deep learning framework that significantly advances automated fetal ultrasound image classification. By fusing the complementary hierarchical feature representations of four pretrained CNNs (VGG16, VGG19, ResNet50, and Xception) with Grad-CAM attention-enhanced inputs, the model achieves a peak classification accuracy of **99.63%** and demonstrates high resilience against dataset imbalance, as evidenced by an F1-score of 0.97 for the minority abdominal circumference (AC) class. Operating entirely on raw image data without relying on external clinical metadata, the framework is uniquely suited for seamless integration into real-world clinical workflows. Furthermore, the incorporation of Grad-CAM visualization provides crucial clinical interpretability, transforming the ensemble model into a transparent tool for clinical decision support. Future efforts will prioritize large-scale multi-center validation, lightweight architectural optimization for real-time edge deployment, and the exploration of transformer-based vision architectures to further scale the framework's clinical utility and diagnostic robustness.

Declarations

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflicts of interest.

Availability of Data and Materials: The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Acknowledgment

The authors express their gratitude to KSTEPS, DST, Government of Karnataka, for their financial support and for granting a Ph.D. fellowship, which made this research possible. We also extend our sincere thanks to Dr. Vanita B. Metgud and Dr. Satwik B. Metgud from Metgud Hospital - Advanced Laparoscopy Centre and IVF, Belagavi, Karnataka, India, for providing fetal ultrasound images and sharing valuable insights into fetal analysis and abnormalities, which greatly contributed to this study.

References:

1. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. IEEE Trans. Med. Imaging. 36, 2204–2215 (2017). https://doi.org/10.1109/TMI.2017.2712367 .

2. Al-Ra zgan, M., Ali, Y.A., Awwad, E.M.: Enhancing Fetal Medical Image Analysis through Attention-guided Convolution: A Comparative Study with Established Models. Journal of Disability Research. 3, (2024). https://doi.org/10.57197/jdr-2024-0005.
3. Kiserud, T., Piaggio, G., Carroli, G., Widmer, M., Carvalho, J., Neerup Jensen, L., Giordano, D., Cecatti, J.G., Abdel Aleem, H., Talegawkar, S.A., Benachi, A., Diemert, A., Tshefu Kitoto, A., Thinkhamrop, J., Lumbiganon, P., Tabor, A., Kriplani, A., Gonzalez Perez, R., Hecher, K., Hanson, M.A., Gülmezoglu, A.M., Platt, L.D.: The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight. PLoS Med. 14, (2017). https://doi.org/10.1371/journal.pmed.1002220 .
4. Papageorghiou, A.T., Ohuma , E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., Noble, J.A., Pang, R., Victora, C.G., Barros, F.C., Carvalho, M., Salomon, L.J., Bhutta, Z.A., Kennedy, S.H., Villar, J.: International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. The Lancet. 384, 869–879 (2014). https://doi.org/10.1016/S0140-6736(14)61490-2.
5. Unterscheider, J., Daly, S., Geary, M.P., Kennelly, M.M., McAuliffe, F.M., O'Donoghue, K., Hunter, A., Morrison, J.J., Burke, G., Dicker, P., Tully, E.C., Malone, F.D.: Optimizing the definition of intrauterine growth restriction: the multicenter prospective PORTO Study. Am. J. Obstet. Gynecol. 208, 290.e1- 290.e6(2013). https://doi.org/10.1016/j.ajog.2013.02.007 .
6. Sarris, I., Ioannou, C., Ohuma , E.O., Altman, D.G., Hoch, L., Cosgrove, C., Fathima, S., Salomon, L.J., Papageorghiou, A.T.: Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. BJOG. 120, 33–37 (2013). https://doi.org/10.1111/1471-0528.12315.
7. Salomon, L.J., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S.L., Kalache, K., Leung, K.Y., Malinger, G., Munoz, H., Prefumo, F., Toi, A., Lee, W.: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics and Gynecology. 37, 116–126 (2011). https://doi.org/10.1002/uog.8831.
8. 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. 12, (2022). https://doi.org/10.3390/diagnostics12092229 .
9. 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. 12, 18–45 (2024). https://doi.org/10.14738/tecs.123.16985 .
10. Muehlematter, U.J., Daniore, P., Vokinger, K.N.: Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit. Health. 3, e195–e203 (2021). https://doi.org/10.1016/S2589-7500(20)30292-2.
11. Xie, J., Hu, J., Li, A., Liu, X.: DAFS-Net: Lightweight medical image segmentation via fusing spatial and frequency domains. Signal Image Video Process. 20, (2026). https://doi.org/10.1007/s11760-026-05346-x.
12. Kanna, S.K.R., Shajin, F.H., Rajesh, P., Mannepalli, K.: A multi-branch multi-scale convolutional neural network using automatic detection of fetal arrhythmia. Signal Image Video Process. 18, 87–96 (2024). https://doi.org/10.1007/s11760-024-03133-0.
13. Shen, D., Wu, G., Suk, H. Il: Deep Learning in Medical Image Analysis. Annu. Rev. Biomed. Eng. 19, 221 – 248 (2017). https://doi.org/10.1146/annurev-bioeng-071516-044442.
14. Zhang, Y., Zhu, J., Long, S., Bai, J., Lu, Y., Chen, G.: An automatic measurement of cervix dilation in intrapartum ultrasound image. Signal Image Video Process. 19, (2025). https://doi.org/10.1007/s11760-024-03759-0.
15. Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S.: A Review on Deep-Learning Algorithm s for Fetal Ultrasound-Image Analysis. (2022). https://doi.org/10.1016/j.media.2022.102629 .
16. van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., van Ginneken, B.: Automated measurement of fetal head circumference using 2D ultrasound images. PLoS One. 13, (2018). https://doi.org/10.1371/journal.pone.0200412 .
17. Gornale, S.S., Kamat, P., Hiremath, P.S., Kumar, S., Goh, K.W.: Automated Segmentation and Trimester-Based Classification of Fetal Head Circumference in Ultrasound Images Using Deep Learning Techniques. Intern. J. Pattern Recognit. Artif. Intell. (2026). https://doi.org/10.1142/s021800142552038x.
18. . Faghihi, A., Fathollahi, M. & Rajabi, R. Diagnosis of skin cancer using VGG16 and VGG19 based transfer learning models. Multimed Tools Appl 83, 57495–57510 (2024). https://doi.org/10.1007/s11042-023-17735-2
19. Kamal, K., EZ-ZAHRAOUY, H.: A comparison between the VGG16, VGG19 and ResNet50 architecture frameworks for classification of normal and CLAHE processed medical images, https://www.researchsquare.com/article/rs-2863523/v1, (2023). https://doi.org/10.21203/rs.3.rs-2863523/v1.
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. (2016). https://doi.org/10.1007/s11263-019-01228-7.
21. Ishikawa, G., Xu, R., Ohya , J., Iwata, H.: Detecting a Fetus in Ultrasound Images using Grad CAM and Locating the Fetus in the Uterus. In: International Conference on Pattern Recognition Applications and Methods. pp. 181–189. Science and Technology Publications, Lda (2019). https://doi.org/10.5220/0007385001810189 .
22. Rahaman, M.M., Li, C., Yao, Y., Kulwa, F., Wu, X., Li, X., Wang, Q.: DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. Comput. Biol. Med. 136, (2021). https://doi.org/10.1016/j.compbiomed.2021.104649 .

23. Rauf, F., Attique Khan, M., Albarakati, H.M., Jabeen, K., Alsenan, S., Hamza, A., Teng, S., Nam, Y.: Artificia l intelligence assisted common maternal fetal planes prediction from ultrasound images based on information fusion of customized convolutional neural networks. Front. Med. (Lausanne). 11, (2024). https://doi.org/10.3389/fmed.2024.1486995 .
24. Li, M., Jiang, Y., Zhang, Y., Zhu, H.: Medical image analysis using deep learning algorithms. Front. Public Health. 11, (2023). https://doi.org/10.3389/fpubh.2023.1273253 .
25. Xiao, S., Zhang, J., Zhu, Y., Zhang, Z., Cao, H., Xie, M., Zhang, L.: Application and Progress of Artificia l Intelligence in Fetal Ultrasound, (2023). https://doi.org/10.3390/jcm12093298 .
26. Sivasubramanian, A., Sasidharan, D., Sowmya, V. et al. Efficient feature extraction using light-weight CNN attention-based deep learning architectures for ultrasound fetal plane classification. Phys Eng Sci Med 48,1079–1093(2025). https://doi.org/10.1007/s13246-025-01566-6
27. Ghabri, H., Alqahtani, M.S., Ben Othman, S., Al-Rasheed, A., Abbas, M., Almubarak, H.A., Sakli, H., Abdelkarim, M.N.: Transfer learning for accurate fetal organ classification from ultrasound images: a potential tool for maternal healthcare providers. Sci. Rep. 13, (2023). https://doi.org/10.1038/s41598-023-44689-0.
28. Gornale, S.S., Kamat, P.C., Hiremath, P.S., Siddalingappa, R.: A Hybrid Ensemble of Denoising Autoencoders and Deep Learning Models for Fetal Image Analysis. Cureus Journal of Computer Science. (2025). https://doi.org/10.7759/s44389-025-09506-x.
29. Rathika, S., Mahendran, K., Sudarsan, H., Ananth, S.V.: Novel neural network classification of maternal fetal ultrasound planes through optimized feature selection. BMC Med. Imaging. 24, (2024). https://doi.org/10.1186/s12880-024-01453-8.
30. Harikumar, A., Surendran, S., Gargi, S.: Explainable AI in Deep Learning Based Classification of Fetal Ultrasound Image Planes. In: Procedia Computer Science.pp. 1023–1033. Elsevier B.V. (2024). https://doi.org/10.1016/j.procs.2024.03.291 .
31. Hasan, M.N., Aowlad Hossain, A.B.M.: Fetal Brain Planes Classification Using Deep Ensemble Transfer Learning from U-Net Segmented Fetal Neurosonography Images. International Journal of Image, Graphics and Signal Processing. 16, 74–86 (2024). https://doi.org/10.5815/ijigsp.2024.04.06 .
32. Liang, J., Huang, R., Kong, P., Li, S., Wang, T., Lei, B.: SPRNet: Automatic fetal standard plane recognition network for ultrasound images. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 38–46. Springer (2019). https://doi.org/10.1007/978-3-030-32875-7_5.
33. Regmi, B., Shah, C.: Classification Methods Based on Machine Learning for the Analysis of Fetal Hea lth Data(2023). https://doi.org/10.48550/arXiv.2311.10962
34. 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.31, 2053–2066 (2022). https://doi.org/10.1109/TIP.2022.3150294 .
35. Fiorentino, M.C., Migliorelli, G., Villani, F.P., Frontoni, E., Moccia, S.: Contrastive prototype federated learning against noisy labels in fetal standard plane detection. International Journal of Computer Assisted Radiology and Surgery . 20, 1431–1439 (2025). https://doi.org/10.1007/s11548-025-03400-6.
36. Ghelich Oghli, M., Shabanzadeh, A., Moradi, S., Sirjani, N., Gerami, R., Ghaderi, P., Sanei Taheri, M., Shiri, I., Arabi, H., Zaidi, H.: Automatic fetal biometry prediction using a novel deep convolutional network architecture. Physica Medica. 88, 127–137 (2021). https://doi.org/10.1016/j.ejmp.2021.06.020 .
37. Ambsdorf, J., Munk, A., Llambias, S., Christensen, A.N., Mikolaj, K., Balestriero, R., Tolsgaard, M.G., Feragen, A., Nielsen, M.: General Methods Make Great Domain-Specific Foundation Models: A Case-Study on Fetal Ultrasound. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., and Park, J. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 271–281. Springer Nature Switzerland, Cham (2026).https://doi.org/10.1007/978-3-032-04981-0_26
38. Salomon, L.J., Alfirevic, Z., Da Silva Costa, F., Deter, R.L., Figueras, F., Ghi, T., Glanc, P., Khalil, A., Lee, W., Napolitano, R., Papageorghiou, A., Sotiradis, A., Stirnemann, J., Toi, A., Yeo, G.: ISUOG Practice Guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound in Obstetrics and Gynecology. 53, 715–723 (2019). https://doi.org/10.1002/uog.20272.
39. Gülmez, B.: A novel deep neural network model based Xception and genetic algorithm for detection of COVID-19 from X-ray images. Ann. Oper. Res. 328, 617–641 (2023). https://doi.org/10.1007/s10479-022-05151-y.
40. Gornale, S.S., Patravali, P.U., Hiremath, P.S.: Early Detection of Osteoarthritis based on Cartilage Thickness in Knee X-ray Images. International Journal of Image, Graphics and Signal Processing. 11, 56–63 (2019). https://doi.org/10.5815/ijigsp.2019.09.06.
41. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? ArXiv. abs/1411.1792, (2014). https://doi.org/10.48550/arXiv.1411.1792 end automation using Deep Learning. Nat. Commun. 14, (2023). https://doi.org/10.1038/s41467-023-42438-5.
42. Lo, J., Lim, A., Wagner, M.W., Ertl-Wagner, B., Sussman, D.: Fetal Organ Anomaly Classification Network for Identifying Organ Anomalies in Fetal MRI. Front. Artif. Intell. 5, (2022). https://doi.org/10.3389/frai.2022.832485 .

43. Patil, G., Palaiahnakote, S., Gornale, S.S., Lopresti, D.P.: Altered Handwritten Text Detection in Document Ima ges Using Deep Learning. Intern. J. Pattern Recognit. Artif. Intell. 38, (2024). https://doi.org/10.1142/S0218001424520062 .
44. Kuzu, A., Santur, Y.: Early Diagnosis and Classification of Fetal Health Status from a Fetal Cardiotocography Dataset Using Ensemble Learning. Diagnostics. 13, (2023). https://doi.org/10.3390/diagnostics13152471 .
45. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-Down Neural Attention by Excitation Backprop. Int. J. Comput. Vis. 126, 1084 –1102 (2018). https://doi.org/10.1007/s11263-017-1059-x.
46. Gornale, S.S., Kamat, P.C., Siddalingappa, R., Goh, K.W., Li kefang: View of Two-Stage Machine Learning Pipeline for Fetal Head Analysis. International Journal on Advanced Electrical and Computer Engineering. (2026). https://doi.org/10.65521/ijaece.v15i1S.1368
47. Raghu, M., Zhang, C., Kleinberg, J.M., Bengio, S.: Transfusion: Understanding Transfer Learning for Medical Imaging. In: Neural Information Processing Systems(2019). https://doi.org/10.48550/arXiv.1902.07208
48. Ennab, M., Mcheick, H.: Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models. Mach. Learn. Knowl. Extr. 7, (2025). https://doi.org/10.3390/make7010012.
49. Slimani, S., Hounka, S., Mahmoudi, A., Rehah, T., Laoudiyi, D., Saadi, H., Bouziyane, A., Lamrissi, A., Jalal, M., Bouhya, S., Akiki, M., Bouyakhf, Y., Badaoui, B., Radgui, A., Mhlanga, M., Bouyakhf, E.H.: Fetal biometry and amniotic fluid volume assessment end-to-