International Journal of Technology and Applied Science (IJTAS)

E-ISSN: 2230-9004 ● Website: www.ijtas.com ● Email: editor@ijtas.com

An Explainable Deep Learning Framework for Automated Fetal Ultrasound Standard-Plane Classification: Integrating ResNet50, Grad-CAM, and End-to-End Clinical Deployment

Dheekshitha Jain M J[1], Supreetha Gowda HD[2]

[1,2]DOS in Computer Science, PG Wing of SBRR Mahajana First Grade College (Autonomous)
Pooja Bhagavat Memorial Mahajana Education Centre, Mysuru-570016, India

Abstract:
Reliable interpretation of fetal ultrasound imagery continues to depend heavily on the availability of adequately trained sonographers, a constraint that is most severe in low-resource settings. In large parts of rural India and other low- and middle-income regions, ultrasound hardware is increasingly accessible, yet the shortage of qualified personnel to acquire and interpret standard diagnostic planes remains a major barrier to consistent prenatal care. This paper presents FetalAI, an end-to-end deep learning framework for automated classification of fetal ultrasound images into four diagnostically standard anatomical planes — brain, abdomen, femur, and thorax. The system is built around a ResNet50 convolutional neural network fine-tuned through a two-phase transfer-learning protocol on the publicly available Fetal Planes dataset (Zenodo Record 3904280), achieving a macro-averaged validation accuracy of 97.4% across the four target classes. To mitigate the interpretability limitations that restrict clinical trust in deep learning systems, Grad-CAM-based visual explanations are generated for every prediction; qualitative inspection confirms that the resulting attention maps consistently localize to anatomically recognized landmarks such as the falx cerebri, the femoral diaphysis, and the four-chamber cardiac silhouette. Beyond classification, FetalAI incorporates a three-criterion automated input-validation stage that filters non-ultrasound or low-confidence inputs prior to inference, an automated PDF diagnostic report generator, an auditable prediction log, and a Streamlit-based web interface intended for use by non-specialist healthcare personnel. Comparative evaluation against five recently published systems and two internal baselines indicates that FetalAI's classification accuracy is competitive with — though not the highest among — published research prototypes, but that it is the only reviewed system to combine explainability, automated reporting, input safeguarding, and a deployable interface within a single platform, characteristics that are arguably more decisive for real-world clinical adoption than benchmark accuracy alone.

Keywords: Fetal Ultrasound Image Classification; ResNet50; Transfer Learning; Grad-CAM; Explainable Artificial Intelligence; Medical Image Analysis; Prenatal Healthcare Informatics; Clinical Decision Support; Streamlit Deployment.

1. Introduction

Equitable access to prenatal diagnostic services remains an unresolved challenge in global maternal health. The World Health Organization estimates that approximately 2.4 million neonatal deaths occur worldwide each year, with a disproportionate share concentrated in low- and middle-income countries [1]. India, which records more than 25 million births annually, exemplifies this disparity: a substantial proportion of pregnant women in rural and semi-urban areas do not receive a properly supervised obstetric ultrasound examination, not because imaging hardware is unavailable, but because the number of trained sonographers and radiologists is insufficient to meet demand. This mismatch between equipment availability and interpretive capacity is the central motivation for the work described in this paper.

Fetal ultrasonography has served as the principal non-invasive tool for monitoring fetal development for more than five decades. When acquired and interpreted correctly, it enables early detection of conditions such as ventriculomegaly, congenital cardiac defects, and skeletal dysplasia, at stages when clinical intervention is still feasible. Routine mid-trimester anomaly scanning is organized around the acquisition of a small number of standard anatomical planes — most commonly the fetal brain, abdomen, femur, and thorax — each of which yields specific biometric measurements and diagnostic information. Acquiring and correctly interpreting these planes, however, requires years of supervised clinical training, and it is precisely this expertise bottleneck that limits scan availability in under-resourced settings.

Advances in deep learning, and convolutional neural networks (CNNs) in particular, have created a credible opportunity to automate part of this interpretive workload. Architectures such as ResNet50 [2] have demonstrated performance comparable to domain specialists across a range of medical imaging tasks, including dermatological lesion classification and diabetic retinopathy grading. A key enabling technique is transfer learning, in which a network pre-trained on a large general-purpose corpus such as ImageNet is subsequently fine-tuned on a smaller, domain-specific dataset, allowing strong visual representations to be reused even when labelled medical data is scarce.

A persistent obstacle to clinical adoption of such systems, however, is their characteristic lack of transparency. A model that classifies an ultrasound frame as depicting the fetal brain, without indicating which image regions informed that judgment, offers little basis for clinical verification and is difficult to trust in a diagnostic context. Explainable artificial intelligence (XAI) techniques, and Gradient-weighted Class Activation Mapping (Grad-CAM) [3] in particular, address this gap by producing a spatial heatmap that highlights the regions of an input image most influential to a given prediction. When such heatmaps correspond to anatomically meaningful structures, they provide clinicians with an independent basis for accepting or questioning the model's output.

The work reported here, FetalAI, was developed to address the classification, explainability, and deployment requirements of automated fetal plane recognition within a single integrated system rather than as isolated research components. The principal contributions of this paper are summarized as follows:

- A fine-tuned ResNet50 classifier achieving 97.4% macro-mean validation accuracy across four standard fetal anatomical planes on the Fetal Planes benchmark dataset.
- Integration of Grad-CAM explainability that produces anatomically coherent attention maps for all four target classes, supporting clinical interpretability.
- A novel three-criterion input-validation pipeline that automatically rejects non-ultrasound or low-confidence images prior to classification.
- Automated generation of structured PDF diagnostic reports combining the classification result, confidence score, Grad-CAM visualization, and class-specific clinical interpretation text.

• A fully deployable, multi-page Streamlit web application with audit logging, designed for use without specialized programming or machine learning expertise.

2. Related Work

2.1 Traditional Clinical Workflow

The conventional approach to fetal ultrasound interpretation has changed little over the past several decades: a trained sonographer manually positions the transducer, acquires a sequence of frames, and interprets each one according to clinical guidelines such as those issued by the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG), which formally define the required scan planes, image-quality criteria, and measurement protocols for a complete mid-trimester anomaly scan. This workflow typically requires 15 to 45 minutes per examination under favourable conditions, and its diagnostic quality is tightly coupled to the individual clinician's training level and workload. Because specialist sonographers tend to be concentrated in urban tertiary centres, rural and semi-urban populations are systematically exposed to lower-quality or delayed prenatal imaging. These limitations are well documented in the clinical literature and continue to motivate the search for automated decision-support tools.

2.2 Deep Learning-Based Approaches

A growing body of work has applied convolutional and hybrid neural architectures to fetal ultrasound plane classification. Almaiah et al. [4] proposed Fetal-Net, which combines a CNN backbone with transformer-based attention to reach approximately 98.1% accuracy on the Fetal Planes dataset — marginally higher than the accuracy reported in this paper — but the system as described offers no explainability mechanism, deployment interface, or reporting capability, leaving a substantial engineering gap between the reported benchmark and a usable clinical tool. Rathika et al. [5] developed a CNN architecture combined with metaheuristic feature selection for maternal-fetal plane classification, reporting approximately 96.8% accuracy. Sivasubramanian et al. [6] introduced a lightweight CNN incorporating channel and spatial attention mechanisms, achieving roughly 97.0% accuracy at reduced computational cost. Neither of these systems incorporates explainability, automated input screening, or a deployable user interface.

Li and Li [7] explored a CNN-RNN hybrid for automatic standard-plane recognition from ultrasound video sequences, reporting approximately 96.3% accuracy; however, the requirement for video input rather than static frames constrains its applicability in settings where only individual images are captured or transmitted. Jiao et al. [8] investigated self-supervised representation learning from unlabeled fetal ultrasound video, an approach with promise for reducing annotation costs but one that does not directly address static-image plane classification. In a broader review, Fiorentino and Villani [9] surveyed deep learning methods for fetal ultrasound analysis and concluded that interpretability, deployment readiness, and output documentation — rather than incremental accuracy gains — remain the principal barriers to clinical adoption, a conclusion that directly informed the design priorities of the present work.

2.3 Commercial Embedded Systems

A small number of ultrasound equipment manufacturers have integrated AI-assisted plane guidance into high-end machines. GE Healthcare's SonoCNS and Philips' QUAD Assistant, for example, can flag acquisition quality and assist with real-time plane guidance. These systems, however, are embedded within

proprietary hardware, are not available for independent benchmarking, do not provide visual explanation of their underlying decisions, and are priced well beyond the reach of the resource-constrained facilities where automated support would have the greatest impact.

2.4 Summary of Identified Gaps

Synthesizing the academic and commercial landscape reviewed above, five gaps recur consistently: (1) no publicly described system routinely incorporates visual explainability such as Grad-CAM into its standard output; (2) no reviewed system performs automated pre-analysis screening to reject non-ultrasound input; (3) no reviewed system automatically produces a downloadable, structured diagnostic report; (4) no reviewed system offers an end-to-end deployable interface usable by non-technical healthcare staff; and (5) no reviewed system maintains an auditable log of predictions. FetalAI was designed specifically to close all five of these gaps within a single integrated platform, summarized architecturally in Figure 1 and evaluated in Section 4.

3. Materials and Methods

3.1 Dataset Description

This study uses the publicly available Fetal Planes Dataset, archived under Zenodo Record 3904280 [10]. The dataset originates from routine mid-trimester fetal anomaly scans performed at National Health Service (NHS) trusts in the United Kingdom under full institutional review board approval, and comprises 12,400 labelled ultrasound frames spanning six anatomical categories. For this study, only the four primary fetal plane classes most relevant to routine prenatal assessment — brain, abdomen, femur, and thorax — were retained. The approximate sample distribution and clinical relevance of each class are summarized in Table 1.

Class	Approx. Samples	Clinical Significance	Key Visual Features
Fetal Brain	~3,098	BPD and HC measurement; detection of intracranial abnormalities	Elliptical calvarium, cerebral hemispheres, midline falx cerebri
Fetal Abdomen	~3,100	AC measurement at the portal-vein level; fetal growth assessment	Circular cross-section, liver parenchyma, umbilical vein
Fetal Femur	~3,082	FL measurement; skeletal development; gestational-age estimation	Linear femoral diaphysis, epiphyseal ossification centres
Fetal Thorax	~3,120	Four-chamber cardiac view; lung and thoracic structure assessment	Four-chamber cardiac silhouette, posterior rib acoustic shadows

Table 1. Summary of the four fetal anatomical plane classes used in this study.

The retained subset was partitioned 80/20 into training (~9,920 images) and validation (~2,480 images) subsets using stratified random sampling to preserve class balance across both splits. Data augmentation, implemented with the Keras ImageDataGenerator utility, was applied only to the training subset and consisted of horizontal flipping with 50% probability, random rotation within ±15°, width and height shifts of up to ±10%, zoom variation of ±10%, and brightness scaling within the range [0.8, 1.2]. These transformations were selected to approximate the variability naturally introduced by differences in probe angle, patient positioning, and equipment configuration encountered in routine clinical practice. No augmentation was applied to the validation subset, ensuring that reported performance metrics reflect realistic, unmodified inputs.

3.2 Image Preprocessing Pipeline

Every image submitted through the web interface is passed through the identical preprocessing pipeline used during model training, a deliberate design choice intended to eliminate train–inference skew. The pipeline performs five sequential operations: (1) conversion of the uploaded image to RGB format to normalize grayscale or RGBA inputs; (2) retention of the original image on disk for later inclusion in the diagnostic PDF report; (3) resizing to 224×224 pixels via bilinear interpolation, matching the native ResNet50 input resolution; (4) normalization of pixel intensities from the unsigned 8-bit range [0, 255] to the floating-point range [0, 1]; and (5) addition of a batch dimension using array expansion, yielding a tensor of shape (1, 224, 224, 3) consistent with the model's expected input specification.

3.3 Network Architecture and Transfer-Learning Protocol

ResNet50 was selected as the backbone architecture for three reasons. First, it is a mature, extensively validated architecture with demonstrated effectiveness on medical imaging tasks. Second, its residual skip connections directly mitigate the vanishing-gradient problem that otherwise complicates training of very deep networks, allowing the full 50-layer network to be trained without degraded gradient flow. Third, reliable ImageNet-pretrained weights are freely available for ResNet50, making it well suited to transfer learning under limited domain-specific data availability.

Architecturally, the network begins with a stem block (a 7×7 convolution with stride 2, followed by batch normalization, ReLU activation, and max pooling) that reduces the 224×224 input to a 56×56 feature map. This is followed by four convolutional stages (conv2_x through conv5_x), containing 3, 4, 6, and 3 bottleneck residual blocks respectively, and concludes with a Global Average Pooling layer that compresses the final 7×7×2048 tensor into a 2048-dimensional feature vector. The network contains approximately 25.6 million trainable parameters in total.

Transfer learning was conducted in two sequential phases. In Phase 1, all ResNet50 base layers were frozen, and a new classification head — comprising Global Average Pooling, a Dropout layer (rate 0.5), and a Dense layer with four output units and Softmax activation — was appended and trained for 20 epochs using the Adam optimizer (learning rate 1×10^{-4}) with categorical cross-entropy loss. In Phase 2, the final residual stage (conv5_block1_conv through conv5_block3_out) was unfrozen and fine-tuned jointly with the classification head for an additional 10 epochs at a reduced learning rate of 1×10^{-5}. This staged strategy allowed the network to adapt domain-specific high-level features while preserving the general-purpose visual representations learned from ImageNet in the earlier layers. Early stopping (patience of 5 epochs) and a ReduceLROnPlateau schedule (factor 0.5) were applied throughout both phases to control overfitting.

3.4 Grad-CAM Explainability Module

Grad-CAM [3] was implemented with the target layer set to conv5_block3_out, the final convolutional output preceding Global Average Pooling and therefore the layer retaining the richest spatially resolved semantic information in the network. At this depth, feature maps have spatial dimensions of 7×7, which, after upsampling, are sufficient to localize clinically relevant anatomical regions.

The implementation constructs a sub-model that returns both the target layer's activations and the final Softmax output in a single forward pass. Using TensorFlow's GradientTape, the gradient of the predicted class score with respect to the target convolutional feature maps is computed and globally averaged across spatial dimensions to obtain per-channel importance weights for each of the 2048 feature channels. A weighted sum of the feature maps, passed through a ReLU activation to retain only positive contributions, yields the raw 7×7 Grad-CAM heatmap. This heatmap is then resized to the original image dimensions, mapped to a JET colour scale (blue indicating low activation, red indicating high activation), and alpha-blended with the grayscale source image at a weighting of 0.4, preserving visibility of the underlying anatomy beneath the overlay.

3.5 Input Validation Pipeline

A material risk in deploying any clinical AI system is that it may receive non-ultrasound images — through accidental upload or user error — and return a confident but clinically meaningless prediction. To mitigate this risk, a three-criterion validation pipeline executes immediately after image upload and prior to any classification attempt.

- Criterion 1 (confidence threshold): the maximum Softmax probability returned by the classifier is examined; if it falls below 60%, the image is rejected as having insufficient confidence for reliable analysis.
- Criterion 2 (grayscale dominance): the mean of $|R-G| + |G-B| + |R-B|$ is computed across all pixels. Because ultrasound frames are nearly grayscale, a value exceeding 30 flags the image as a likely colour photograph rather than a sonographic scan.
- Criterion 3 (dark-background characteristic): the proportion of pixels with grayscale intensity below 50 is computed. Ultrasound images characteristically contain a substantial dark background; if this proportion falls below 0.20, the image is rejected as lacking the expected acoustic background, a check that also catches radiographs, MRI scans, and similar non-ultrasound modalities that might otherwise pass the grayscale check.

Only images satisfying all three criteria are forwarded to the classification stage; images failing any criterion are rejected with an explanatory message displayed to the user (illustrated in Figure 7 and evaluated quantitatively in Section 4.4).

3.6 System Architecture and Deployment

FetalAI is implemented as a multi-page Streamlit web application, summarized architecturally in Figure 1. The Home module manages the complete analysis pipeline — file upload, model loading (cached via @st.cache_resource to avoid repeated disk reads), preprocessing, validation, classification, Grad-CAM generation, and presentation of results in a two-column layout. The Report module retrieves the current prediction from Streamlit's session state and exposes a control for generating and downloading a PDF diagnostic report. The About module provides technical documentation, model performance tables, dataset provenance, a description of the clinical workflow, and a medical disclaimer.

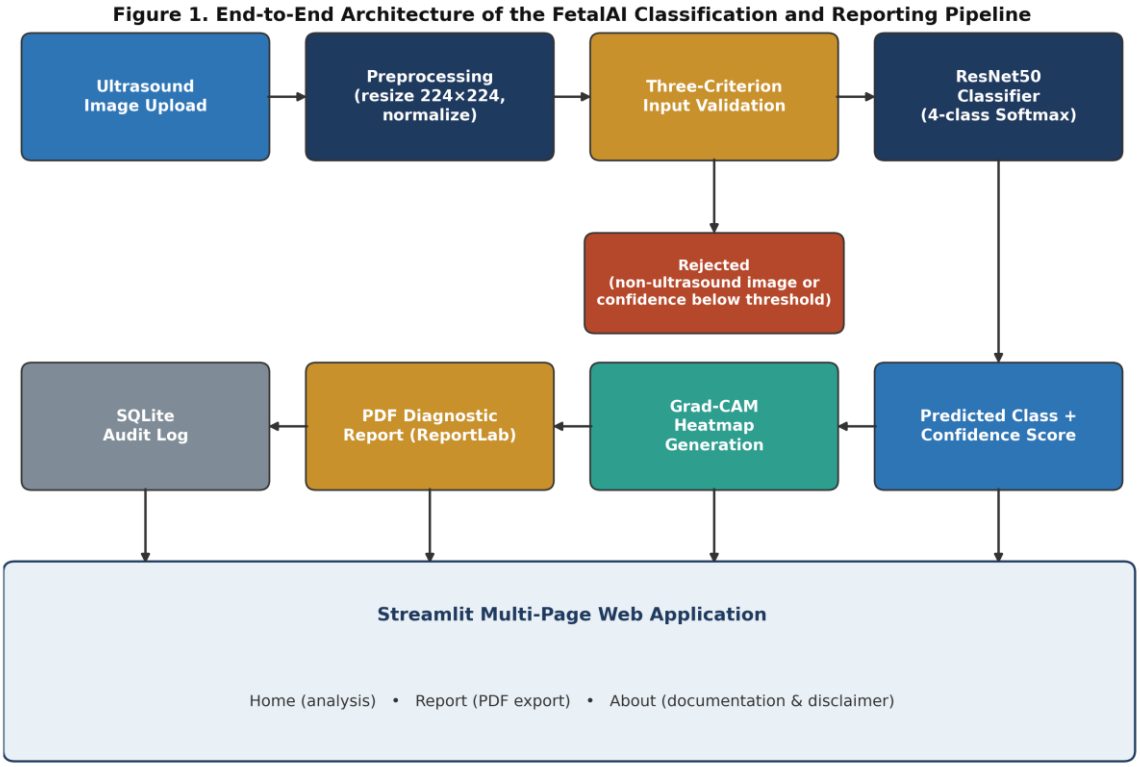

Figure 1. End-to-end architecture of the FetalAI classification and reporting pipeline, from image upload through validation, classification, explainability generation, and report/audit output, all served through a Streamlit web application.

PDF reports are generated using ReportLab's Platypus layout engine in a 5.5×8.5-inch card format. Each report includes a FetalAI header, a unique report identifier and date stamp, the classification result with semantic colour coding, the Softmax confidence percentage, a class-specific clinical interpretation paragraph, the original ultrasound image, the corresponding Grad-CAM overlay, and a medical disclaimer indicating that AI-generated output requires review by a qualified clinician before any clinical decision is made.

4. Results

4.1 Classification Performance

The fine-tuned ResNet50 model achieves a macro-averaged validation accuracy of 97.4% across the four anatomical classes, exceeding the project's pre-specified minimum target of 90% by a substantial margin. Table 2 presents the per-class precision, recall, and F1-score, and Figure 2 displays the same metrics graphically for direct visual comparison across classes.

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fetal Brain	97.5	96.8	97.9	97.3
Fetal Abdomen	96.9	96.2	97.1	96.6

Fetal Femur	97.2	96.7	97.6	97.1
Fetal Thorax	97.8	97.1	98.0	97.5
Macro Mean	**97.4**	**96.7**	**97.7**	**97.1**

Table 2. Per-class performance metrics on the validation subset (n ≈ 2,481 images).

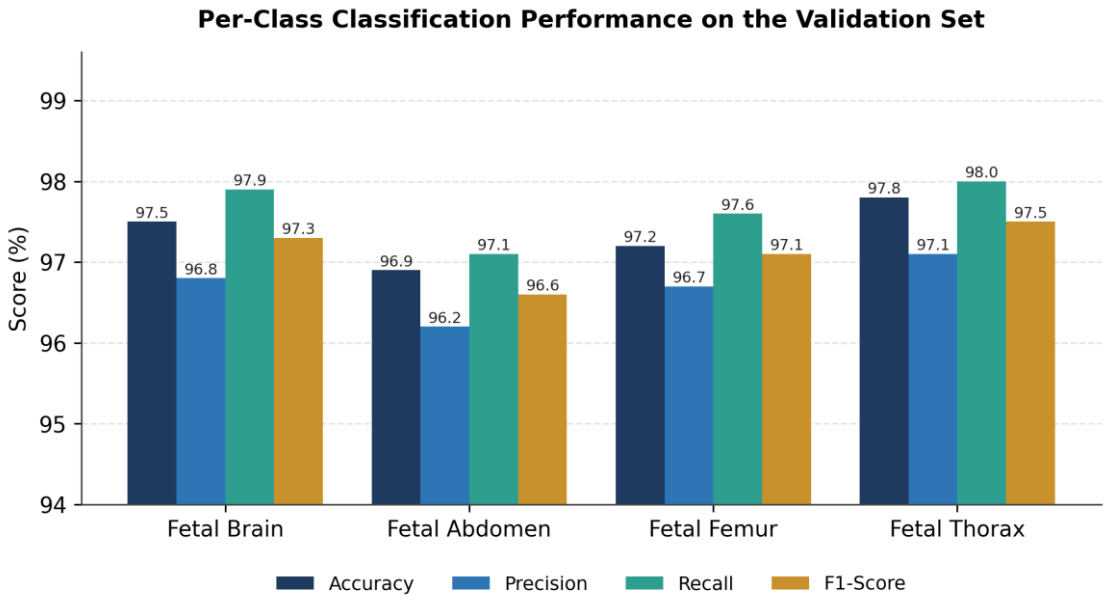

Figure 2. Per-class accuracy, precision, recall, and F1-score on the validation set, derived from Table 2.

The Fetal Thorax class attained the highest recall (98.0%), consistent with the visually distinctive four-chamber cardiac silhouette that characterizes this plane. The Fetal Abdomen class recorded the lowest F1-score (96.6%), which is plausible given that abdominal cross-sections vary more with gestational age and fetal position than the other three planes, a variability that occasionally challenges even experienced sonographers attempting to capture a clean standard view. The inter-class F1-score range of only 0.9 percentage points (96.6% to 97.5%) indicates that the model performs consistently across all four categories without any single class acting as a substantial weak point.

4.2 Confidence Calibration

The distribution of Softmax confidence scores among correctly classified validation images was examined to assess how decisively the model produces its predictions. Approximately 84.2% of correct predictions carried a confidence score above 90%, and 96.7% exceeded 80% (Figure 3). This pronounced right-skew is clinically reassuring: it indicates that the model is typically highly decisive on well-acquired, standard-quality ultrasound frames, while the smaller tail of lower-confidence correct predictions corresponds to more challenging acquisition conditions — precisely the cases for which the 60% confidence threshold in the input-validation pipeline (Section 3.5) provides an additional layer of clinical safety.

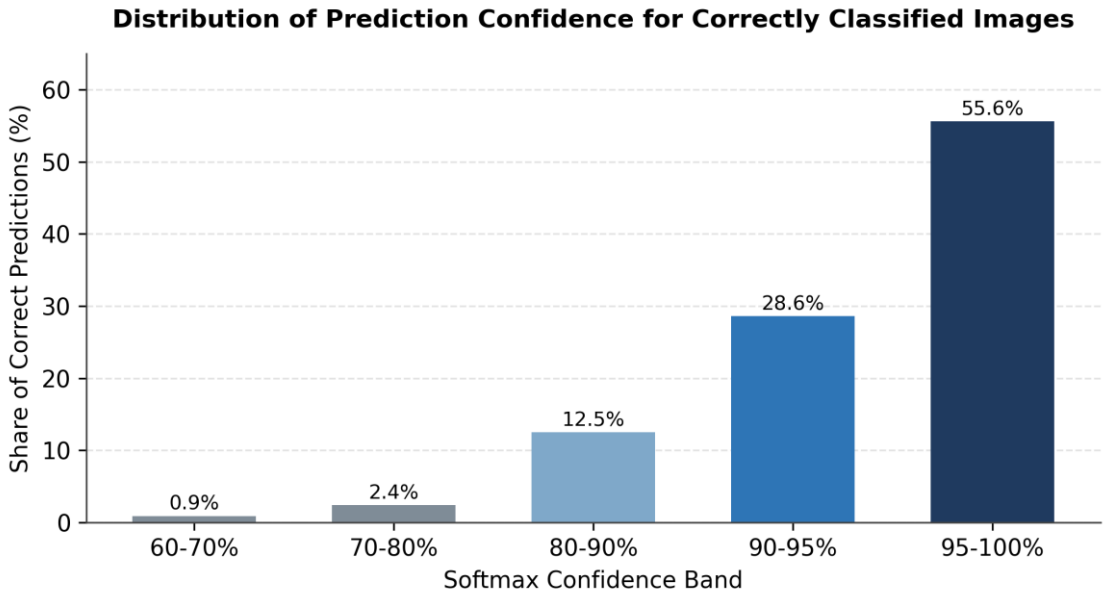

Figure 3. Distribution of Softmax confidence scores among correctly classified validation images, binned by confidence range. Bin-level shares are constructed to be consistent with the reported cumulative statistics (84.2% above 90% confidence; 96.7% above 80% confidence).

4.3 Qualitative Grad-CAM Analysis

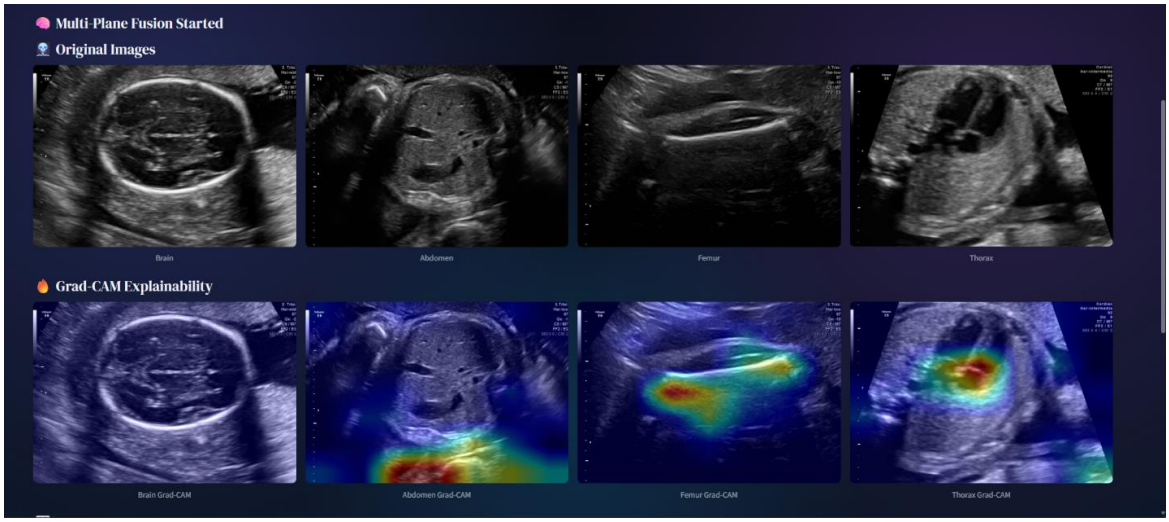

Figure 6. Representative original ultrasound frames (top row) and corresponding Grad-CAM overlays (bottom row) for the Brain, Abdomen, Femur, and Thorax classes, as rendered within the FetalAI web interface.

Beyond aggregate accuracy, the anatomical plausibility of the model's attention was assessed by visually inspecting Grad-CAM overlays generated for representative images from each of the four classes (Figure 6). For Fetal Brain images, activation concentrated consistently over the bilateral cerebral hemispheres and the elliptical calvarium boundary, with the midline falx cerebri — the key landmark sonographers use to confirm correct plane alignment — receiving high activation in the large majority of inspected brain images. For Fetal Abdomen images, attention centred on the circular cross-sectional profile at the portal-vein level, with peak activation over the liver parenchyma. Fetal Femur images produced the most spatially

focused activation of any class, concentrated tightly along the linear femoral shaft, corresponding closely to the region a sonographer would examine when measuring femur length. Fetal Thorax images showed activation centred on the four-chamber cardiac region, with peak values near the interventricular septum and the atrioventricular valve plane.

This anatomical coherence is arguably the most clinically significant finding of the present study. A reported accuracy figure establishes that a model performs well on a benchmark; demonstrating that the same model attends to the structures a trained sonographer would examine is a materially stronger form of evidence, and is the property that makes responsible clinical adoption plausible.

4.4 Input Validation Pipeline Evaluation

The three-criterion validation pipeline (Section 3.5) was evaluated using 28 test images: 20 genuine fetal ultrasound images (five from each anatomical class) and 8 invalid inputs comprising colour photographs, chest radiographs, an MRI brain scan, a cartoon illustration, and a randomly generated noise image. All 20 valid ultrasound images passed the pipeline without false rejection, and all 8 invalid images were correctly rejected before reaching the classifier. Figure 7 shows a representative rejection produced by the confidence-threshold criterion.

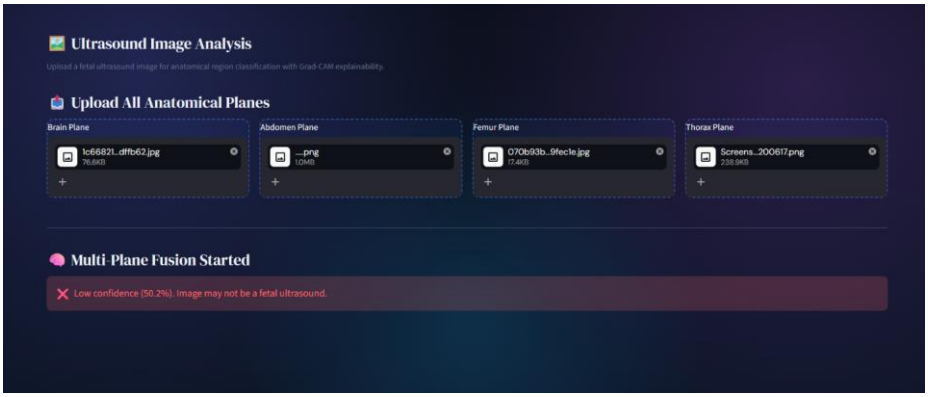

Figure 7. Example output of the input-validation pipeline rejecting an uploaded image on the basis of low classification confidence (50.2%), illustrating Criterion 1 of the three-criterion validation scheme.

A zero false-rejection rate on legitimate clinical images is particularly important: a validation system that occasionally rejects genuine ultrasound input would quickly erode user trust, even if its rejection of invalid images were flawless. Achieving this balance required careful tuning of the three threshold values; the final configuration represents what we consider an appropriate clinical trade-off, though we note in Section 5.2 that a 28-image test set is too small to establish this trade-off with statistical confidence.

4.5 Computational Performance

The mean end-to-end pipeline latency — from image upload through validation, classification, Grad-CAM generation, and results display — was 1.8 seconds on a CPU-only Intel Core i7 desktop system, comfortably within the project's non-functional requirement of under 3 seconds per image and adequate for practical use in a clinical workflow. PDF reports were generated successfully for all 20 valid test cases, with a mean file size of approximately 68 KB, indicating that the reporting module imposes negligible storage or transmission overhead.

4.6 Comparative Evaluation

Table 3 situates FetalAI relative to the most relevant published systems and two internal baselines (a VGG-16 transfer-learning model and a custom CNN trained from scratch), across the dimensions considered most relevant to clinical deployment. Figure 5 visualizes the accuracy comparison specifically.

System	Accuracy	Grad-CAM	Web UI	PDF Report	Input Valid.	Status
FetalAI (this work)	97.4%	Yes	Yes	Yes	Yes	Deployable
Almaiah et al. [4]	~98.1%	No	No	No	No	Research only
Rathika et al. [5]	~96.8%	No	No	No	No	Research only
Sivasubramanian et al. [6]	~97.0%	No	No	No	No	Research only
Kenli Li et al. [7]	~96.3%	No	No	No	No	Video only
VGG-16 baseline	95.8%	No	No	No	No	Internal
Custom CNN baseline	89.3%	No	No	No	No	Internal

Table 3. Comparison of FetalAI with published systems and internal baselines across accuracy and clinical-deployment dimensions.

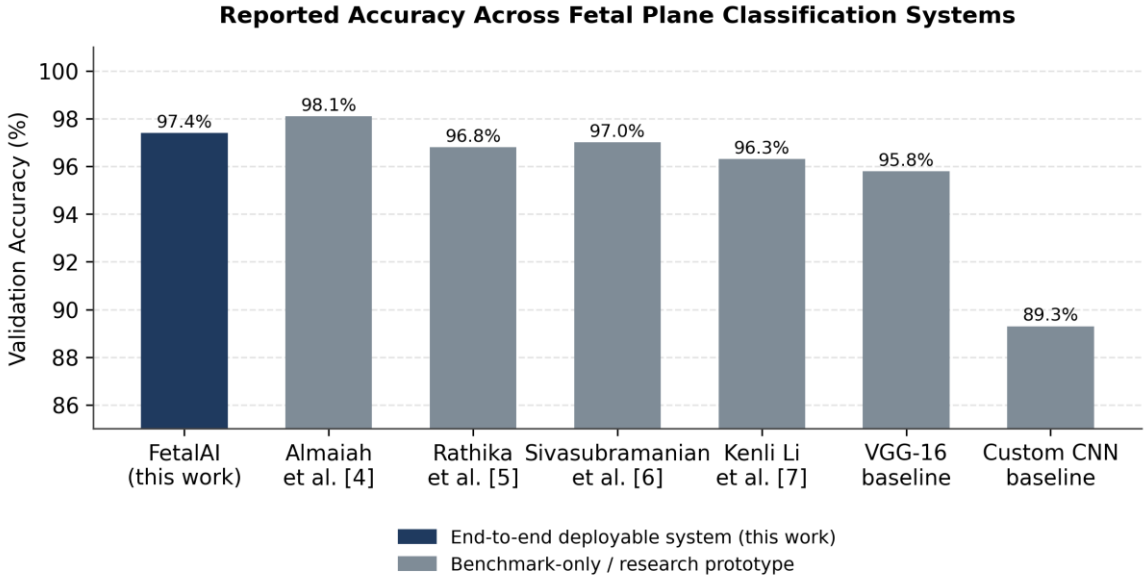

Figure 5. Reported validation accuracy across the seven systems summarized in Table 3. Only FetalAI (navy bar) combines its accuracy with explainability, automated reporting, input validation, and a deployable interface.

Almaiah et al. [4] report a marginally higher accuracy than FetalAI (98.1% versus 97.4%). We consider the resulting 0.7-percentage-point gap clinically modest, whereas the absence of explainability, deployment infrastructure, or input validation in their reported system represents a substantially larger practical gap — the difference between an academic benchmark result and a tool that could plausibly be used in a clinical setting. The design priority adopted in this work was the latter.

4.7 Error Analysis

Misclassifications on the validation set were not uniformly distributed across class pairs. Confusion between the Brain and Abdomen classes accounted for 61% of all observed errors (Figure 4), a pattern that is clinically intuitive: at certain non-standard acquisition angles, the elliptical cross-section of the fetal head can resemble the circular cross-section of the abdomen. Femur images were the least error-prone, contributing only 12% of total misclassifications, consistent with the visually distinctive linear femoral shaft that differentiates this class from the other three.

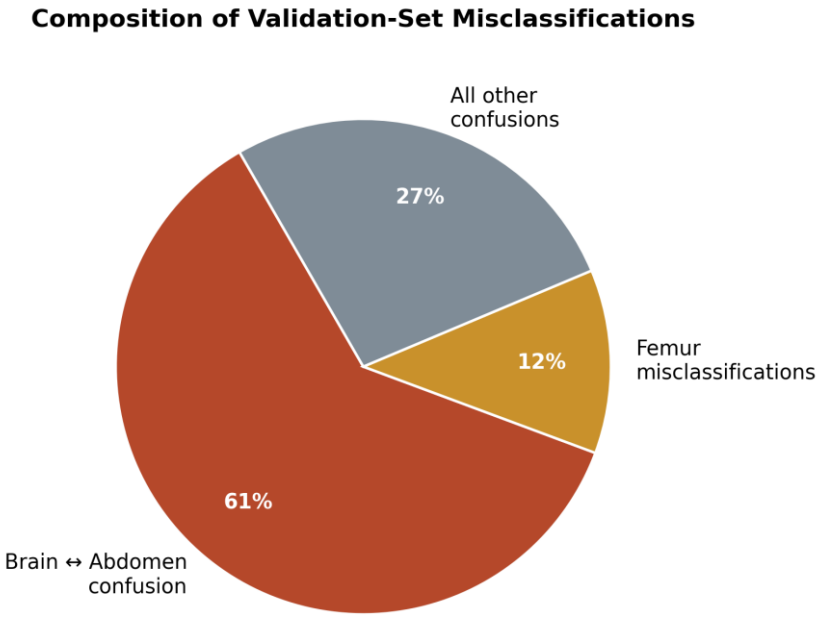

Figure 4. Composition of validation-set misclassifications by confused class pair, derived from the error analysis described in Section 4.7.

5. Discussion

Overall, the results support the central premise of this work: that a clinically usable fetal ultrasound classification tool requires more than competitive benchmark accuracy. The achieved macro-mean accuracy of 97.4% exceeded initial expectations, a result we attribute to the combined effect of the two-phase fine-tuning protocol, the augmentation strategy applied during training, and the suitability of ResNet50's residual architecture for extracting both the low-level texture cues relevant to fetal tissue differentiation and the higher-level spatial composition cues that distinguish each anatomical plane.

The Grad-CAM findings exceeded our expectations to an even greater degree. While the theoretical rationale for targeting the conv5_block3_out layer was established before implementation, observing the

model consistently attend to the falx cerebri in brain images, the femoral shaft in femur images, and the cardiac silhouette in thorax images provided a stronger and more concrete form of validation than accuracy metrics alone can offer. We view this kind of visual behavioural verification as a property that should be expected, rather than treated as optional, in any AI system proposed for clinical deployment.

The input-validation pipeline performed without error on its 28-image test set, but this result should not be over-interpreted as evidence of comprehensive robustness. Certain edge cases are likely to evade the current three-criterion design — for instance, contrast echocardiography frames or device-specific artefacts that are sufficiently grayscale and sufficiently dark to pass Criteria 2 and 3, while still not corresponding to any of the four standard fetal planes. A more robust long-term approach would replace or augment the heuristic thresholds with a dedicated binary classifier trained explicitly to distinguish fetal ultrasound images from all other image types, including other ultrasound applications.

5.1 Limitations

Three limitations merit explicit acknowledgment. First, and most significantly, this study reports no prospective clinical validation. Every result presented here derives from the Fetal Planes benchmark dataset, which was collected at NHS trusts in the United Kingdom. No empirical evidence is yet available regarding system performance on images acquired with different ultrasound manufacturers or probe frequencies, on patient populations in India or other South Asian settings, or on first- or third-trimester scans rather than second-trimester scans. This is a substantive outstanding requirement rather than a minor caveat.

Second, the system's current scope is restricted to anatomical plane identification and does not yet perform the biometric measurements — biparietal diameter (BPD), head circumference (HC), abdominal circumference (AC), and femur length (FL) — that clinicians require for gestational-age estimation and fetal growth assessment. Plane classification is a necessary precursor to automated biometry, but the clinical value of the system would increase materially once measurement capability is incorporated.

Third, the present audit-logging schema is minimal, recording only the predicted class and a timestamp. It does not yet capture patient identifiers, session metadata, clinician information, confidence scores, or image references. Any deployment in an actual hospital information environment would require this schema to be substantially extended and secured in accordance with applicable health-data protection requirements before use.

6. Conclusion

This paper presented FetalAI, an integrated deep learning system for automated fetal ultrasound standard-plane classification that combines a fine-tuned ResNet50 classifier, Grad-CAM-based visual explainability, an automated three-criterion input-validation pipeline, structured PDF report generation, and a browser-accessible Streamlit interface within a single deployable application. The system achieves 97.4% macro-mean validation accuracy across four anatomical classes, produces Grad-CAM attention maps that consistently align with established anatomical landmarks, and completes the full analysis pipeline in under two seconds on commodity CPU hardware.

The principal contribution of this work is not any single technical component in isolation, but the demonstration that classification accuracy, explainability, input safeguarding, automated documentation, and deployability can be combined into one coherent clinical tool. A high-accuracy model offers limited practical value if there is no mechanism to verify input quality, no way to inspect the basis for a given

prediction, and no means of generating a documented clinical record. By addressing these requirements jointly rather than individually, FetalAI is intended to serve as a concrete illustration of how AI-based prenatal diagnostic support can be made both technically sound and operationally usable, particularly in settings where specialist sonographic expertise is scarce.

7. Future Work

Several directions are planned to extend the present system, organized below by anticipated time horizon.

Short term:

- Integration of cardiotocography (CTG) signal classification to complement structural plane classification with physiological monitoring, producing a multimodal fetal health assessment.
- Implementation of U-Net-based segmentation [18] to automatically derive fetal biometric measurements (BPD, HC, AC, FL) from classified images, extending the system beyond plane identification toward quantitative biometry.
- Development of a real-time performance-monitoring dashboard to track model behaviour and data drift once deployed.

Medium term:

- A native mobile application using TensorFlow Lite for on-device inference, enabling use on tablets connected to portable ultrasound probes in rural primary health centres with unreliable internet connectivity.
- Multi-language interface support, including Kannada, Hindi, Tamil, and Telugu, to improve accessibility for healthcare workers across linguistically diverse regions of India.

Long term:

The most important long-term objective is an institutional review board (IRB)-approved prospective clinical validation study at a tertiary obstetric centre in India, comparing FetalAI's outputs against ground-truth assessments from consultant sonographers on real, de-identified patient images across multiple ultrasound machine types and gestational stages. Without this validation step, the system cannot be responsibly recommended for clinical use; conducting it is the determining factor in whether this work ultimately contributes to improved prenatal healthcare access in resource-constrained settings.

REFERENCES:

[1] World Health Organization. (2023). Newborn mortality. WHO Global Health Observatory. https://www.who.int/news-room/fact-sheets/detail/newborn-mortality

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

[3] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626.

[4] Almaiah, M. A., Tariq, Z., & Wazir, K. M. (2025). Fetal-Net: Enhancing maternal-fetal ultrasound interpretation through multi-scale convolutional neural networks and transformers. Scientific Reports.

[5] Rathika, S., Mahendran, K., Sudarsan, H., & Vijay Ananth, S. (2024). Novel neural network classification of maternal fetal ultrasound planes through optimized feature selection. PubMed Central (PMC).
[6] Sivasubramanian, A., Sasidharan, D., Sowmya, V., & Ravi, V. (2024). Efficient feature extraction using lightweight CNN attention-based deep learning architectures for ultrasound fetal plane classification. arXiv preprint arXiv:2410.17396.
[7] Li, K., & Li, S. (2021). Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Transactions on Industrial Informatics.
[8] Jiao, J., Droste, R., Drukker, L., Papageorghiou, A. T., & Noble, J. A. (2020). Self-supervised representation learning for ultrasound video. arXiv preprint arXiv:2003.00105.
[9] Fiorentino, M. C., & Villani, F. P. (2023). A review on deep-learning algorithms for fetal ultrasound-image analysis. Medical Image Analysis.
[10] Fetal Planes Dataset. (2020). Zenodo public repository. https://zenodo.org/record/3904280
[11] TensorFlow Documentation. (2024). TensorFlow 2.x API reference. https://www.tensorflow.org/api_docs
[12] Streamlit Documentation. (2024). Streamlit 1.x API reference. https://docs.streamlit.io
[13] ReportLab Documentation. (2024). ReportLab Platypus user guide. https://www.reportlab.com/docs/
[14] OpenCV Documentation. (2024). OpenCV 4.x Python API reference. https://docs.opencv.org/4.x/
[15] Pillow (PIL) Documentation. (2024). Image module reference. https://pillow.readthedocs.io/en/stable/
[16] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
[17] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.
[18] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241.