HybriBrainNet – A Hybrid CNN-ViT Architecture for Automated AI-Based Detection of Fetal Brain Abnormalities in Ultrasound Imaging

Vedhesh Dhinakaran
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham
Bengaluru, 560035, India
bl.en.u4cse23257@bl.students.amrita.edu

Andrew Tom Mathew
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham
Bengaluru, 560035, India
bl.en.u4cse23269@bl.students.amrita.edu

Nikhil Sanjay
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham
Bengaluru, 560035, India
bl.en.u4cse23239@bl.students.amrita.edu

***Abstract*—Prenatal detection of fetal brain abnormalities remains challenging due to observer variability and limited diagnostic consistency. This study proposes a hybrid CNN-Vision Transformer architecture with cross-attention fusion for multi-class classification of fetal brain abnormalities from ultrasound. The model integrates DenseNet-121 and Swin-Tiny with Convolutional Block Attention Module (CBAM) to jointly identify 16 distinct abnormality types. A bidirectional cross-attention mechanism enables seamless information exchange between CNN and transformer branches. The model was trained with class-weighted sampling and cosine annealing learning rate scheduling, prioritizing balanced performance via macro F1 scores. Grad-CAM visualizations provide pixel-level explainability, enabling clinical validation of model predictions. Comprehensive evaluation metrics demonstrate balanced performance across all abnormality classes. This approach reduces observer dependency, minimizes diagnostic errors, and enhances early intervention for fetal brain disorders.**

***Index Terms*—Fetal brain abnormalities, Ultrasound image, Deep Learning, Convolutional Neural Networks, DenseNet121, CBAM, Explainable AI, Grad-CAM**

I. Introduction

Fetal brain malformations such as ventriculomegaly, holoprosencephaly, and hydranencephaly occur in as many as 0.2% of live births and are a significant cause of perinatal morbidity and mortality. Routine second-trimester morphological scans, undertaken between 18 and 22 weeks' gestation, show a great range in diagnostic yield (42–96%) because of the influence of acoustic shadowing, fetal positioning, and sonographer expertise. The intricacy of in-utero neurodevelopment, with events such as neural tube closure and cortical folding proceeding in parallel, pushes the limits of traditional ultrasound interpretation and potentially veils subtle early markers of pathology.

Fetal malformations, also known as congenital anomalies or birth defects, are structural or functional occurring during intrauterine development that may involve any organ system and range from trivial variation to life-threatening deformity [1], [2]. The anomalies can be caused by genetic mutations, chromosomal disorders, for example aneuploidies, teratogenic injuries, or vascular and disruptive occurrences, presenting as a change in tissue morphology or function identifiable by prenatal imaging techniques [3], [4]. Arnold–Chiari malformations is a structural defect where the brain tissue extends into spinal canal due to a malformed skull [5]. Arachnoid cysts is a fluid-filled sac that forms between the brain or spinal cord and the arachnoid membrane [6]. Cerebellar hypoplasia is the underdevelopment or incomplete formation of the cerebellum affecting coordination and balance [7]. Encephaloceles is a neural tube defect where brain tissue and membranes protrude through openings in the skull [8]. Holoprosencephaly is the failure of the forebrain to divide properly into two hemispheres, leading to facial and brain abnormalities [9]. Hydranencephaly is a severe condition where the cerebral hemispheres are replaced by cerebrospinal fluid-filled sacs [10]. Intracranial hemorrhage is bleeding within the fetal skull or brain tissue, often due to trauma or vascular abnormalities [11]. Intracranial tumor is the presence of a neoplastic mass within the fetal brain, potentially affecting development and structure [12]. Mega cisterna magna is the condition of enlargement of cisterna magna which is a fluid space behind the cerebellum without other malformalities [13]. Mild ventriculomegaly is the slight dilation of the brain's lateral ventricles, often benign but sometimes linked to other issues [14]. Moderate ventriculomegaly is a more pronounced ventricular dilation suggesting possible developmental or genetic abnormalities [15]. Polencephaly or also called as polymicrogyria is the abnormal brain development with excessive small folds (gyri) leading to cortical

dysfunction [16]. Severe ventriculomegaly is the significant enlargement of the brain ventricles, often indicating serious underlying pathology or poor prognosis [17]. Vein of galen malformation is a rare vascular anomaly where abnormal connections form between arteries and the median vein of the brain [18]. Ventriculomegaly as graded as mild (10–12 mm), moderate (12–15 mm), or severe (¿15 mm) according to atrial diameter cutoffs [19].

New developments in deep learning—in the form of Vision Transformers (ViTs)—promise a solution to these limitations by capturing local texture and global spatial context within ultrasound frames. ViTs have better capabilities in capturing long-range dependencies, supporting stronger morphological pattern recognition with respect to varied anomaly types. Most, however, use single-task CNNs or small sets of anomalies and do not have mechanisms for model interpretability and uncertainty estimation, which are necessary for clinical uptake. The proposed framework fills in these gaps by bringing together multi-task learning, explainable AI, and uncertainty quantification over evidence within an end-to-end, optimization-based pipeline for fetal brain ultrasound.

In this study, proposed a hybrid deep learning model that combines CNNs and Vision Transformers to automatically detect fetal brain abnormalities in ultrasound images. This study used DenseNet-121 to pick up on fine details and textures in the scans, while the Swin Transformer helps capture larger patterns across the image. Instead of just combining the features from both models, a special cross-attention mechanism that lets the two models communicate with each other. This mechanism puts both sets of features into the same space and uses attention to help each model learn from what the other one is doing. After that, the combined features through a classification layer to identify which of the eleven fetal brain abnormalities is present in the ultrasound. finally added Grad-CAM to explain what the model is looking at when it makes a prediction. This shows doctors exactly which parts of the ultrasound image the model focused on, which helps them understand if the model is making sense and trust its recommendations. This entire workflow of the pipeline is elucidated in the block diagram from Fig 1. The main novelty of the proposed framework is a bidirectional cross-attention mechanism that allows the CNN and Vision Transformer to share information with each other. Instead of just combining features from both models, the proposed approach lets them interact directly, so the CNN can learn from the transformer's global view while the transformer benefits from the CNN's detailed local features. Grad-CAM was added to explain which parts of the ultrasound the model is looking at when making predictions, which is important for doctors to trust and validate the model's decisions.

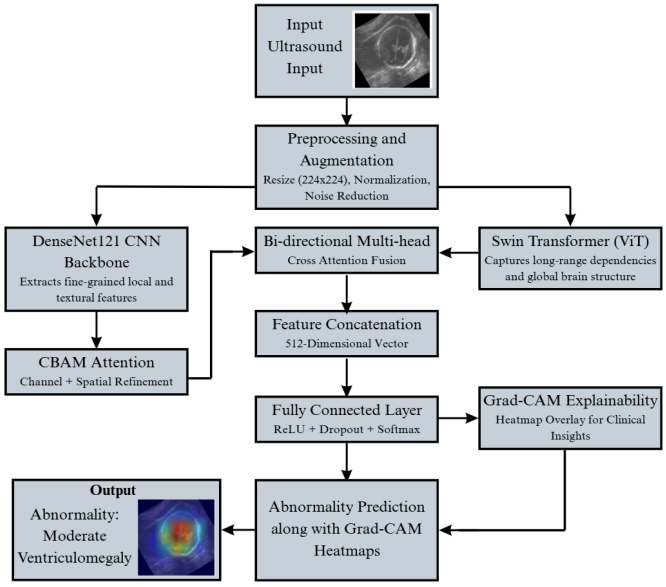

Fig. 1. Flowchart of the hybrid CNN–Vision Transformer framework for fetal brain abnormality detection from ultrasound scans.

II. Literature Survey

Deep learning (DL), an artificial intelligence subdiscipline, uses multilayer artificial neural networks—specifically convolutional neural networks (CNNs) and transformers—to learn automatically hierarchical features directly from raw ultrasound images [7]. DL in fetal imaging allows automatic plane detection, structure segmentation, and anomaly detection, enhancing reproducibility and minimizing operator reliance by extracting discriminative features associated with anatomical and pathological variations [20], [21]. Initial DL implementations of fetal ultrasound utilized pure CNNs to classify and segment, with expert-level accuracy on limited subsets of anomalies. Ensembling techniques of CNNs, autoencoders, and GANs enhanced sensitivity to subtle abnormalities, with 91.4% overall accuracy across 12,450 scans. Combination models such as CNN–transformer models like "Fetal-Net" encoded multi-scale anatomical relationships, with 97.5% accuracy on 12,000+ images. Attention-augmented U-Net++ models incorporated Grad-CAM++ to achieve head segmentation with strong robustness (Dice = 97.52%, IoU = 95.15%) [9], while multi-stage pipelines addressed plane detection, segmentation, and measurement simultaneously with high accuracy and calibrated uncertainty estimation [22]. Even with these improvements, existing frameworks are still restricted to single tasks or limited anomaly subsets without joint confidence quantification across different malformations [23]. Future research should create a generalizable, multi-anomaly, multi-task DL model that provides calibrated probability estimates as well as predictions, incorporates explainable AI methods for end-to-end transparency, and does validation on large, multi-center cohorts with diverse imaging protocols and low-resource environments [20], [23]. Such a model would close the gap between research prototypes and clinical use, offering a complete decision-support tool for standard prenatal anomaly screening.

Despite advances in deep learning for fetal brain imaging, significant challenges persist in clinical deployment. Current state-of-the-art methods either focus on single anomalies or

employ generic multi-class architectures that fail to balance local texture extraction with global contextual reasoning. Most existing approaches lack rigorous attention mechanisms that explicitly highlight clinically relevant anatomical features, making them difficult to validate and trust in clinical settings. Furthermore, inadequate handling of class imbalance in rare fetal abnormalities leads to suboptimal recall for underrepresented conditions, increasing the risk of missed diagnoses. Additionally, the fusion of CNN and transformer modalities in existing hybrid architectures remains simplistic, often relying on late concatenation rather than deep bidirectional reasoning. These limitations collectively hinder the adoption of AI-assisted fetal brain anomaly detection in diverse clinical environments, particularly in resource-constrained settings where operator expertise may be limited.

Key Novelties of the Proposed Architecture:

1) **Hybrid CNN-ViT Architecture with Cross-Attention Fusion**: Unlike prior works that concatenate CNN and transformer features superficially, the proposed model implements bidirectional cross-attention between DenseNet-121 (local feature extraction) and Swin-Tiny (global context), enabling seamless information exchange at the feature level. This allows CNN tokens to directly attend to transformer features and vice versa, producing holistic representations that capture both textural details and anatomical relationships.
2) **Dual Attention Mechanisms for Clinical Interpretability**: The integration of Convolutional Block Attention Module (CBAM) within the CNN branch enables adaptive channel and spatial attention, ensuring the model prioritizes diagnostically relevant features before fusion. This mechanism directly addresses the interpretability gap by highlighting which brain regions and feature channels contribute to each prediction.
3) **Robust Multi-Class Imbalance Handling**: The proposed pipeline employs weighted random sampling based on inverse class frequencies during training, ensuring underrepresented rare abnormalities receive proportional emphasis. This is coupled with macro F1 score-based model selection, prioritizing balanced performance across all 11 fetal brain abnormality classes, including rare conditions where recall is clinically critical.
4) **Rigorous Validation and Explainability Framework**: The model integrates comprehensive evaluation metrics (per-class precision, recall, F1, AUC) alongside Grad-CAM visualizations that provide pixel-level attribution maps. These visual explanations enable clinicians to validate that the model attends to clinically meaningful anatomical features (e.g., lateral ventricles for ventriculomegaly) rather than artifacts, bridging the gap between AI predictions and clinical validation.
5) **Optimized Training Strategy for Medical Imaging**: The use of cosine annealing learning rate scheduling, class-weighted loss contributions, and GPU-accelerated training with automatic mixed precision ensures efficient convergence while maintaining numerical stability. The selection criterion based on macro F1 rather than accuracy reflects the medical reality where balanced sensitivity across all conditions is more important than overall accuracy.
6) **Comprehensive Multi-Abnormality Classification**: The proposed model simultaneously classifies 11 distinct fetal brain abnormalities—including rare conditions such as vein of Galen malformations, holoprosencephaly, and hydranencephaly—with unified architecture and shared feature representations. This multi-abnormality approach contrasts with prior work focusing on subsets of conditions and demonstrates generalizability across the diagnostic spectrum.

These innovations collectively position the proposed architecture as a comprehensive, clinically deployable solution that addresses the research gap between prototype systems and real-world prenatal imaging applications, offering standardized screening performance across diverse clinical settings with reduced operator dependency.

III. Dataset Description and Preprocessing

The fetal brain abnormalities ultrasound dataset from Roboflow [24] has 1,768 ultrasound images in total. The following section delves into the desciption of the dataset and the preprocessing techniques used for this study. The sample ultrasound images for all 16 classes is depicted in Fig 2.

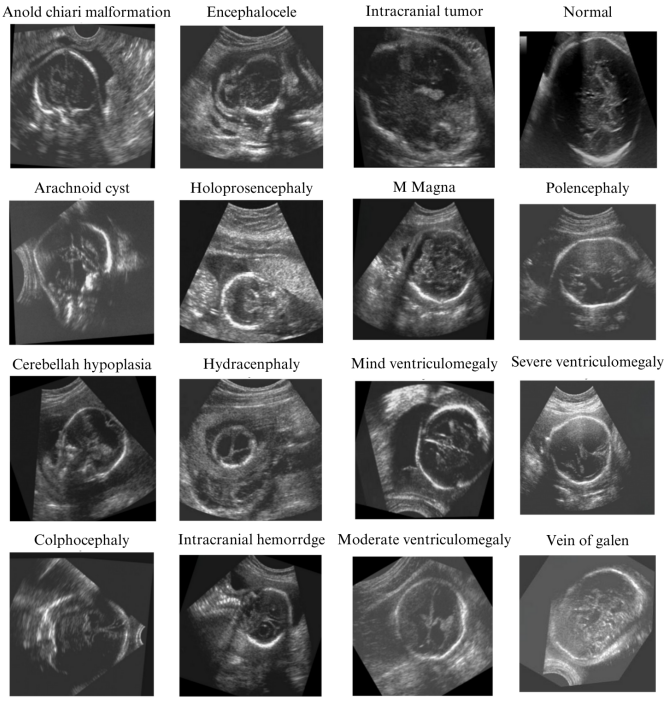

Fig. 2. Sample Ultrasound Images Illustrating the 16-Class Fetal Brain Dataset

A. Data Organization and Class Labels

The research dataset was organized into three distinct splits: training, validation, and testing sets. Each split was stored in a separate directory containing ultrasound images of the fetal brain and a corresponding CSV file (`_classes.csv`) that maps image filenames to their diagnostic labels. The CSV file uses a one-hot encoding format, where each row represents an image and each column (after the first filename column) represents a fetal brain abnormality class. For multi-class classification, this one-hot encoded label was converted to a single class index using the argmax function, which selects the position of the highest value in the one-hot vector. This approach allows the model to handle multiple abnormality types in a single-label classification framework.

The dataset contained 11 distinct fetal brain abnormality classes, including conditions such as Arnold-Chiari malformation, arachnoid cysts, cerebellar hypoplasia, colpocephaly, encephalocele, holoprosencephaly, hydranencephaly, intracranial hemorrhage, intracranial tumors, Mega cisterna magna, ventriculomegaly (mild, moderate, and severe variants), vein of Galen aneurysm malformation, and normal fetal brain scans.

B. Image Augmentation and Normalization

To address the limited size of medical imaging datasets and improve model generalization, comprehensive data augmentation was applied exclusively to the training set. The augmentation pipeline included:

- **Geometric Transformations**: Images were randomly resized with a cropping scale between 0.85 and 1.0 of the original size to simulate varying ultrasound scan focus regions. Additionally, random horizontal flips and rotations of up to 15 degrees were applied to account for different fetal positioning and scanning angles.
- **Affine Transformations**: Random translations of up to 10% and shear transformations of up to 10 degrees were applied to simulate subtle variations in probe placement and angle.
- **Photometric Variations**: Color jitter with slight variations in brightness, contrast, and saturation (each modified by ± 0.1) was applied to account for ultrasound equipment variations and imaging conditions.
- **Grayscale Conversion**: All images, whether grayscale or color, were converted to three-channel grayscale images to ensure consistent input dimensions for the model.
- **Normalization**: Images were converted to tensors and normalized using ImageNet statistics (mean = $[0.485, 0.456, 0.406]$, std = $[0.229, 0.224, 0.225]$) to match the pretraining distributions of the backbone networks.

The validation and test sets underwent minimal preprocessing—only resizing to 224×224 pixels, grayscale conversion, tensor conversion, and normalization—to preserve the integrity of evaluation metrics.

C. Class Imbalance Handling

Medical imaging datasets frequently suffer from class imbalance, where certain abnormality types are underrepresented. To address this, a weighted random sampling strategy was implemented. The weight assigned to each class was computed as the inverse of its frequency in the training set (weight = $1/\text{class\_count}$), ensuring that underrepresented classes contributed proportionally more to the loss during training. A `WeightedRandomSampler` was used during dataloader creation, allowing each mini-batch to include a balanced distribution of samples across all classes.

IV. Methodology

The core innovation of this research is the seamless integration of convolutional and transformer-based feature extraction pathways, enabling the model to capture both local textural details and global contextual relationships in ultrasound images.

The proposed hybrid model that combines DenseNet-121 and Swin Transformer to capture both detailed local features and broader global patterns from fetal brain ultrasound images. The two models share information through a cross-attention mechanism that allows them to learn from each other in a shared feature space. The combined features are then used by a classification layer to identify 11 different types of brain abnormalities. We also used Grad-CAM to create visual explanations that show which parts of the ultrasound image the model focuses on when making predictions, helping doctors understand and trust the model's decisions. This framework is visually described in Fig 3 in the form of the architecture diagram.

A. DenseNet-121 (Convolutional Branch)

The convolutional branch of the hybrid architecture was built upon DenseNet-121, a densely connected convolutional network. DenseNet-121 processes input ultrasound images through its feature extraction layers, generating hierarchical feature maps with dimensions of $[B, C_{\text{CNN}}, H_{\text{CNN}}, W_{\text{CNN}}]$, where B is the batch size, C_{CNN} is the number of feature channels (1024 for DenseNet-121), and H_{CNN} and W_{CNN} are the spatial dimensions of the feature map.

The DenseNet architecture is particularly well-suited for medical image analysis because its dense connections facilitate feature reuse and gradient flow, allowing deeper networks to be trained more effectively. The output of DenseNet's feature extraction layers captures local textural patterns, morphological details, and subtle intensity variations that are diagnostic indicators of fetal brain abnormalities.

B. Convolutional Block Attention Module (CBAM)

Immediately following the DenseNet feature extraction, a Convolutional Block Attention Module (CBAM) was applied to refine the CNN feature maps. CBAM operates through two sequential attention mechanisms: channel attention and spatial attention.

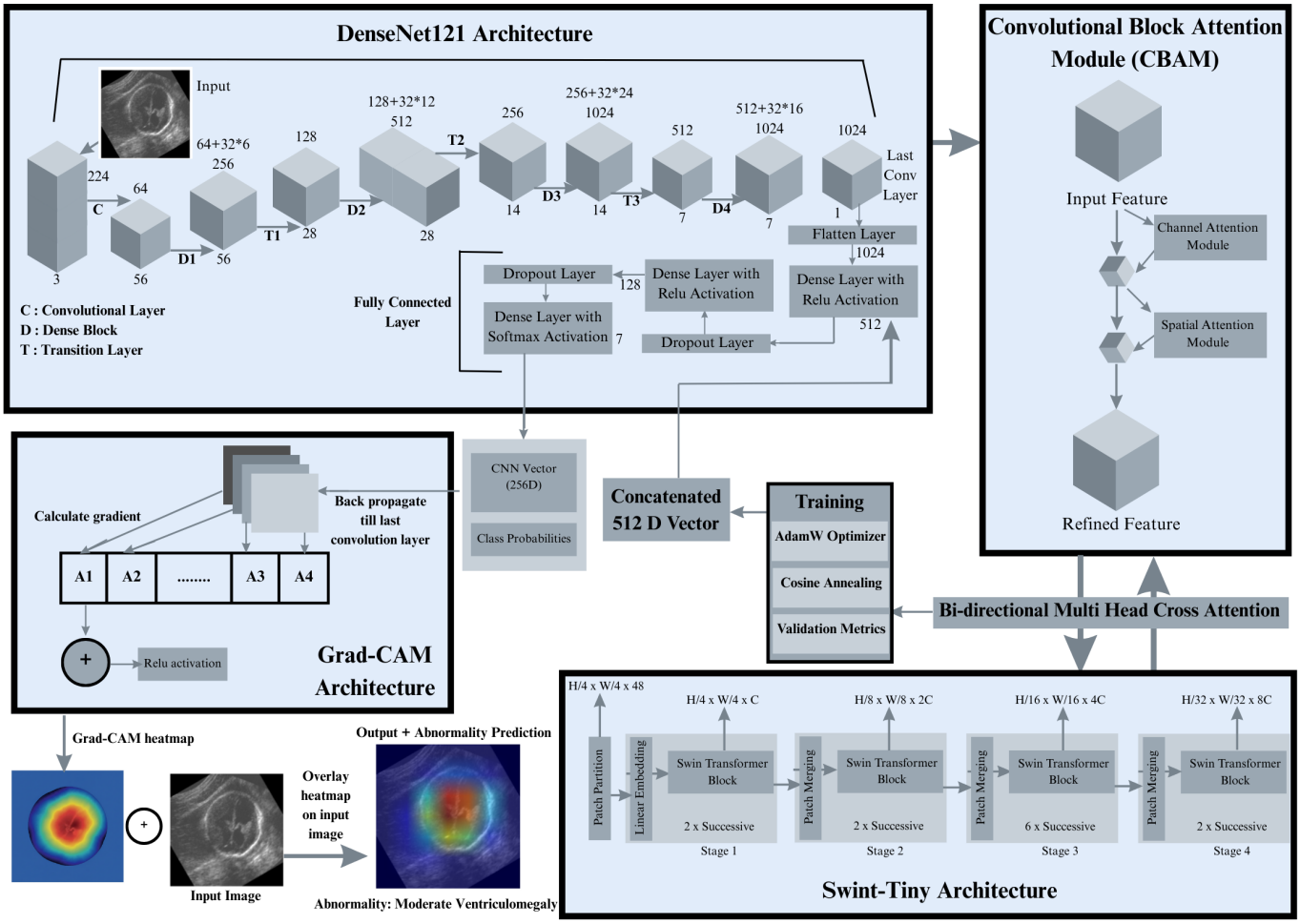

Fig. 3. End-to-End Architecture of Hybrid CNN-ViT Pipeline with Grad-CAM Visualization

Channel Attention: This mechanism learns which feature channels are most informative for the classification task. It computes the average and maximum pooling across all spatial locations for each channel, then uses two fully connected layers with a ReLU activation to generate channel-wise attention weights. These weights are applied via element-wise multiplication to emphasize diagnostic feature channels while suppressing uninformative ones.

Spatial Attention: This mechanism identifies which spatial regions within the feature maps contain the most relevant diagnostic information. It computes the average and maximum across the channel dimension at each spatial location, concatenates these two maps, and applies a convolutional layer followed by sigmoid activation to generate a spatial attention map. Multiplying the feature map by this spatial attention map highlights clinically significant anatomical regions while suppressing irrelevant areas.

By applying CBAM, the model learns to focus on the most diagnostically relevant features before fusing with the transformer branch, thereby improving both performance and interpretability.

C. Swin-Tiny Transformer (Vision Transformer Branch)

The transformer branch employs Swin-Tiny, a hierarchical vision transformer that captures global contextual relationships and long-range dependencies in the ultrasound images. Unlike standard vision transformers that process images at a fixed resolution, Swin Transformer operates using shifted window-based multi-head self-attention, enabling efficient computation while maintaining spatial hierarchies.

The Swin-Tiny model processes the input images and outputs feature maps with dimensions $[B, H_{\text{Swin}}, W_{\text{Swin}}, C_{\text{ViT}}]$ in a channels-last format (where $C_{\text{ViT}} = 768$ for Swin-Tiny). Since the subsequent fusion module expects features in channels-first format $[B, C_{\text{ViT}}, H_{\text{Swin}}, W_{\text{Swin}}]$, a permutation operation was applied to transpose the tensor dimensions accordingly. This permutation ensures compatibility between the CNN and transformer feature representations before fusion.

The Swin Transformer branch excels at identifying fine anatomical differences and global structural relationships that may be difficult to capture with purely convolutional approaches, such as symmetric abnormalities or overall ventricular size relative to total brain volume.

D. Cross-Attention Fusion Mechanism

Rather than simply concatenating features from both branches, a dedicated Cross-Attention Fusion module was designed to enable bidirectional information exchange between CNN and transformer features. This module operates as follows:

Feature Projection: Both CNN and ViT feature maps are projected into a common embedding space of dimension 256 using 1×1 convolutions. This ensures that both feature representations are comparable and can interact meaningfully.

Sequence Flattening: The projected feature maps are flattened along the spatial dimensions and transposed into token sequences of shape $[B, \text{num\_tokens}, \text{embed\_dim}]$. For example, a CNN feature map of shape $[B, 256, 14, 14]$ becomes a sequence of 196 tokens (14×14), where each token is a 256-dimensional vector.

Bidirectional Cross-Attention: Two separate multi-head attention operations are applied:

- **CNN-to-ViT Attention**: CNN tokens serve as queries while ViT tokens serve as keys and values, allowing CNN features to selectively attend to transformer features.
- **ViT-to-CNN Attention**: ViT tokens serve as queries while CNN tokens serve as keys and values, allowing transformer features to selectively attend to CNN features.

Multi-head attention operates by splitting the embedding dimension into multiple "heads," each learning different types of relationships. This allows different heads to focus on different aspects of the data simultaneously (e.g., one head might focus on edges while another focuses on texture).

Token Pooling: After attention refinement, the refined token sequences are pooled along the token dimension using mean pooling, producing two vectors of shape $[B, \text{embed\_dim}]$ representing the refined CNN and ViT features, respectively.

Layer Normalization: Layer normalization is applied to each refined vector to stabilize the learning dynamics and improve convergence.

This cross-attention mechanism ensures that diagnostic information present in one modality (e.g., local texture from CNN) can directly influence the decisions made based on the other modality (e.g., global structure from ViT), resulting in more robust and holistic predictions.

E. Classification Head

The fused CNN and ViT vectors (each of dimension 256) are concatenated to produce a combined feature vector of dimension 512. This combined vector passes through a classification head consisting of:

- A fully connected layer mapping from 512 to 512 dimensions
- ReLU activation to introduce non-linearity
- Dropout with a rate of 0.3 to prevent overfitting
- A final fully connected layer mapping from 512 to the number of classes (11 in this study)

The output of this final layer produces logits for each class, which are converted to probability distributions using softmax during inference.

F. Training Procedure

1) Hyperparameters and Loss Function: The model was trained using Cross-Entropy Loss, a standard loss function for multi-class classification. Cross-Entropy Loss measures the dissimilarity between the predicted probability distribution and the true one-hot encoded label, providing a strong training signal for the model to learn accurate class predictions.

The model parameters were optimized using AdamW, a variant of the Adam optimizer that decouples weight decay from gradient-based updates, providing superior regularization compared to standard Adam. The initial learning rate was set to 1×10^{-4}, and a weight decay coefficient of 1×10^{-4} was applied to encourage smaller weight magnitudes and prevent overfitting.

2) Learning Rate Scheduling: A cosine annealing learning rate scheduler was employed to gradually decrease the learning rate over the course of training. The scheduler reduces the learning rate following a cosine function, starting from the initial learning rate and decaying to near-zero by the final epoch. This approach allows the model to make large updates in early training when gradients are noisier, and then fine-tune weights with smaller updates in later training when gradients are more reliable. The cosine annealing schedule was configured for 25 epochs.

3) Training Configuration: Training was conducted for 25 epochs with a batch size of 8 images per batch. The relatively small batch size was chosen to accommodate memory constraints while ensuring sufficient gradient estimates. All computations were performed on CUDA-accelerated GPUs for efficient training, with automatic mixed precision enabled to reduce memory consumption and accelerate computation without sacrificing numerical precision.

4) Monitoring and Checkpointing: During each epoch, the following metrics were computed and monitored on the validation set:

- **Validation Loss**: The average Cross-Entropy Loss across all validation samples, indicating whether the model is overfitting or underfitting.
- **Macro F1 Score**: The unweighted average of per-class F1 scores, giving equal weight to each class regardless of its frequency. This metric is particularly important for imbalanced datasets because it ensures that the model performs well on all classes, including rare ones.
- **Micro F1 Score**: The weighted average of per-class F1 scores, weighted by the number of true instances for each label. This metric reflects overall accuracy across the entire validation set.
- **Training and Validation Accuracy**: The percentage of samples correctly classified on the training and validation sets, respectively.

The model checkpoint with the highest macro F1 score on the validation set was saved for final evaluation on the test

set. This selection criterion prioritizes balanced performance across all abnormality classes rather than overall accuracy, which is crucial in medical applications where rare but important abnormalities must not be missed.

5) Model Evaluation: After training, the best-performing model checkpoint was loaded and evaluated on the held-out test set. The test set was used to assess the model's generalization performance on completely unseen data. The following metrics were computed:

- **Macro and Micro F1 Scores**: As described above, these metrics provide balanced and overall performance measures, respectively.
- **Classification Accuracy**: The percentage of test samples correctly classified across all classes.
- **Per-Class Precision and Recall**: Precision measures the proportion of predicted positive instances that are actually positive (i.e., how many of the predicted abnormality cases are truly abnormal), while recall measures the proportion of actual positive instances that were correctly identified (i.e., how many of the true abnormality cases were detected). For medical applications, recall is typically more important than precision because missing an abnormality is more critical than a false alarm.
- **Per-Class AUC Scores**: Area Under the Receiver Operating Characteristic (ROC) Curve measures the model's ability to distinguish between classes across all classification thresholds. AUC scores range from 0.5 (random guessing) to 1.0 (perfect classification).

G. Explainability with Grad-CAM

1) Grad-CAM Concept and Implementation: While deep learning models often achieve high accuracy, their decision-making processes are typically opaque ("black boxes"), which is problematic in medical applications where clinicians need to understand why the model made a particular prediction. Gradient-weighted Class Activation Mapping (Grad-CAM) addresses this limitation by generating visual explanations of model predictions.

Grad-CAM operates by computing the gradient of the predicted class score with respect to the activations of the final convolutional layer in the CNN branch (DenseNet-121). These gradients indicate how much each spatial location in the feature map contributes to the final prediction. The gradients are then averaged across all channels (the spatial dimension), and this average gradient serves as a weight for each feature channel. A weighted combination of the feature channels is then computed to generate a heatmap of shape $[H, W]$, where each spatial location represents the model's focus on that region for making the class prediction.

Mathematically, Grad-CAM is computed as follows:

1) For a predicted class c, the gradient of the class score with respect to feature map activations is computed.
2) These gradients are globally average-pooled across spatial dimensions to compute channel-wise weights.
3) The heatmap is computed as a weighted sum of the feature channels using these weights.
4) ReLU is applied to retain only positive contributions (the model attends to these regions for this class).
5) The heatmap is normalized to the range $[0, 1]$ for visualization.

In this implementation, Grad-CAM is computed on the last convolutional layer of DenseNet-121 to provide fine-grained localization of diagnostic regions.

2) Visualization and Clinical Validation: For each abnormality class, a representative test image was selected. Grad-CAM heatmaps were computed for these images using the trained model and overlaid onto the original ultrasound scan using a jet colormap with 50% transparency. The overlaid visualizations clearly highlight which ultrasound regions most strongly influenced the model's abnormality prediction.

These visualizations were displayed side-by-side with the original ultrasound images, annotated with the predicted class label. This presentation enables clinicians to visually verify that the model is attending to clinically relevant anatomical features (e.g., the lateral ventricles for ventriculomegaly) rather than artifacts or irrelevant regions, thereby instilling confidence in the model's predictions and facilitating clinical validation of the approach.

The Grad-CAM visualizations for all 11 abnormality classes were generated and saved, providing comprehensive explainability across the entire diagnostic spectrum.

H. Summary of Methodology

The proposed multi-class fetal brain abnormality classification pipeline integrates several advanced techniques:

1) **Hybrid CNN-ViT Architecture**: Combines the local feature extraction capability of DenseNet-121 with the global contextual understanding of Swin-Tiny Transformers.
2) **Attention Mechanisms**: CBAM in the CNN branch adaptively highlights diagnostically relevant feature channels and spatial regions. Cross-Attention Fusion enables bidirectional information exchange between CNN and transformer features.
3) **Robust Training**: Class-weighted sampling addresses data imbalance, cosine annealing learning rate scheduling improves convergence, and rigorous monitoring of macro F1 score ensures balanced multi-class performance.
4) **Comprehensive Evaluation**: Multiple metrics (precision, recall, F1, AUC) and visual tools (confusion matrix) provide thorough performance assessment.
5) **Clinical Explainability**: Grad-CAM visualizations provide pixel-level explanations, enabling clinicians to validate that the model learns medically meaningful patterns.

This comprehensive approach ensures that the model is robust, interpretable, and suitable for supporting clinical decision-making in fetal brain abnormality detection from ultrasound images.

V. Results and Discussions

The proposed Hybrid Cross-Attention CNN-ViT architecture was assessed on the Fetal Brain Ultrasound Classification Dataset, consisting of 16 diagnostic categories including 11 fetal brain anomalies and normal cases. This model combines DenseNet-121 for fine-grained convolutional feature extraction, Swin-Transformer for global contextual representation, CBAM for spatial-channel refinement, and a novel bidirectional Cross-Attention Fusion module for harmonization of heterogeneous feature spaces between CNN and ViT branches. Training was performed for 25 epochs with AdamW optimization ("LR" = 1 × 10-4, weight decay = 1 × 10-4), cosine annealing learning rate scheduling, WeightedRandomSampler to tackle class imbalance, and extensive augmentation including random resized crops, affine transformation, and color jittering.

A. Training Dynamics and Convergence

The model showed rapid convergence in the first few epochs of training. In fact, Fig 4 shows that in just three epochs, the training accuracy had jumped from 34.4% at Epoch 1 to 88.0% by Epoch 3, while the validation accuracy had increased from 52.3% to 86.4% in the same period. This steep learning curve indicates the effective initialization from ImageNet-pretrained backbones and hence the efficient gradient flow through the cross-attention fusion layers.

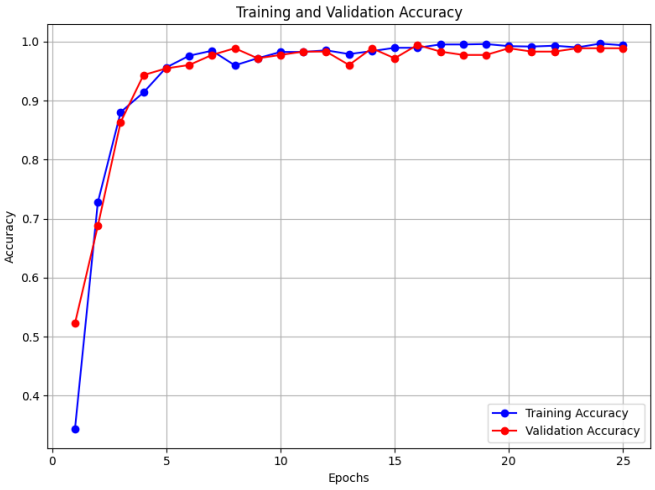

Fig. 4. Training and validation accuracy curves over 25 epochs.

By Epoch 5, the model achieved 95.6% training accuracy and 95.4% validation accuracy, with a macro F1-score of 0.9722-interpreted as a strong multi-class discrimination capability. The training and validation curves remained tightly coupled throughout all epochs, with minimum divergence observed even at convergence: Epoch 25, train accuracy = 99.4%, validation accuracy = 98.9%. This behavior suggests that the combination of aggressive data augmentation, dropout regularization (p = 0.3), and the attention mechanisms effectively avoided overfitting despite the high model capacity.

B. Loss Convergence

Fig 5 presents the loss trajectories of training and validation. The cross-entropy loss dropped abruptly from 2.18 for the training set and 1.76 for the validation set to 0.29 and 0.22, at Epoch 4. After Epoch 10 the convergence stabilized, with final losses at 0.015 for the training set and 0.020 for the validation set at the final epoch.

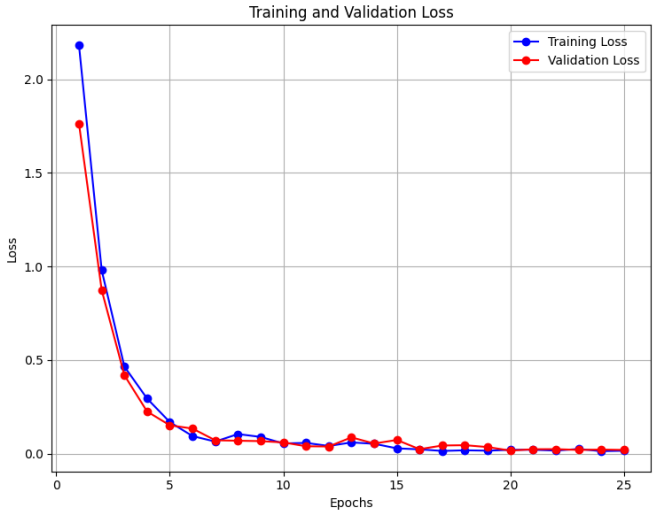

Fig. 5. Training and validation loss curves.

C. Test Set Performance

The best checkpoint, which was on epoch 16, was tested on the held-out test set. The results obtained are elucidated in Table I.

TABLE I
Overall Model Performance Metrics

Metric	Value
Test Accuracy	96%
Macro F1-Score	0.9953
Micro F1-Score	0.9943
Overall Accuracy (avg. over classes)	0.995
Overall Recall (avg. across classes)	0.994

These metrics hence confirm near-perfect alignment between validation and test performances, therefore it can be said the model generalized well to unseen ultrasound images without overfitting to the specific dataset.

D. Confusion Matrix Analysis

The confusion matrix gives a fine view into per-class classification performance, which is shown in Fig 6. Strong diagonal dominance is visible across all 16 categories with only minimal off-diagonal errors.

Notable observations include:

1) **Perfect Classification:** The classes, normal (24/24), mild ventriculomegaly (24/24), moderate ventriculomegaly (22/22), severe ventriculomegaly (19/19), cerebellah hypoplasia (16/16), encephalocele (14/14), and

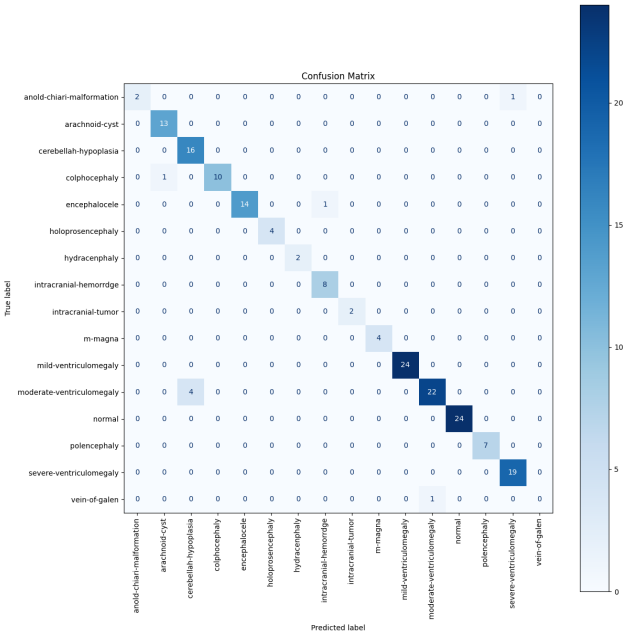

Fig. 6. Confusion matrix for the test set of all 16 classes of the fetal brains. The diagonal element stands for correct classification, while entries which are off diagonal are misclassifications.

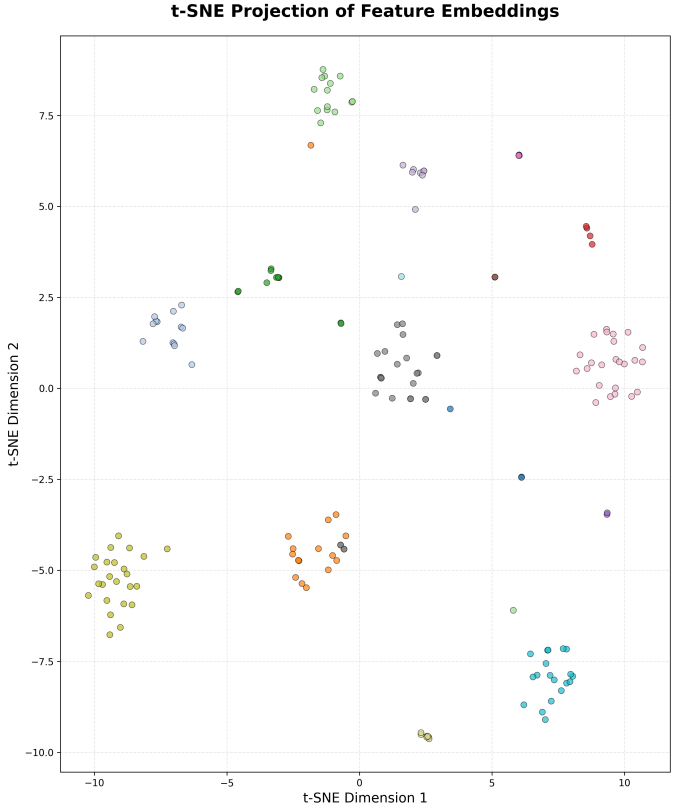

Fig. 7. t-SNE 2D visualization of learned feature embeddings for test data, demonstrating class-discriminative clustering across 16 fetal brain conditions.

arachnoid cyst (13/13) were all classified with 100% accuracy.

2) **Challenging Cases:** There is one misclassification for Colpocephaly, which is predicted as arachnoid cyst, probably due to the overlap of ventricular morphology in both the cases. The moderate-ventriculomegaly had 4 cases misclassified as cerebellar-hypoplasia, which can be reflected by the subtle anatomical similarities in posterior fossa imaging. Arnold Chiari malformation had 1 misclassification into severe-ventriculomegaly, again in keeping with a shared posterior fossa pathology.
3) **Rare Classes:** Even with a small number of training samples, the following rare conditions were recognized correctly: holoprosencephaly (4/4); m-magna (4/4); hydranencephaly(2/2); intracranial-tumor (2/2). WeightedRandomSampler is effective in mitigating class imbalance.
4) **Ventriculomegaly Spectrum:** The model was able to effectively differentiated between mild, moderate, and severe grades/cases of ventriculomegaly with very little confusion, a clinically critical capability for prenatal counseling and treatment planning.

E. Learned Feature Representations: t-SNE and PCA Analysis

The t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization for the test data, of the learned feature embeddings from the hybrid CNN-ViT framework demonstrates clear class-discriminative clustering for fetal brain abnormality detection. The two-dimensional projection depicted in Fig 7 shows that most pathologies yield well-separated clusters. Clinically similar conditions-Moderate and Severe Ventriculomegaly ($n = 26$ and $n = 24$), are positioned correctly in overlapping regions, reflecting their pathophysiological similarity. Normal cases ($n = 24$) form a distinct cluster, indicating robust separation from abnormal scans. Structural abnormalities, such as Encephalocele ($n = 15$) and Arachnoid Cyst ($n = 13$), maintain well-defined spatial regions. Notably, even the rare classes of Arnold–Chiari Malformation ($n = 3$), Hydracephaly ($n = 2$), and Vein-of-Galen anomaly ($n = 1$) preserve identifiable embedding regions, illustrating that the model learned meaningful representations despite limited sample availability. This structured organization of the feature space indicates that bidirectional cross-attention fusion between CNN and ViT branches effectively captures both local pathological markers and global anatomical context, yielding generalizable and interpretable representations suitable for clinical deployment in fetal neurosonography.

The three-dimensional t-SNE projection of learned feature embeddings sheds deeper insight into the complex structure of the feature space learned by the hybrid CNN-ViT framework. The 3D visualization in Fig 8 has revealed a hierarchical clustering pattern extending beyond what was visible in the 2D representation, thus signifying that the model indeed learns multi-faceted representations capturing the characteristics of fetal brain abnormalities along several latent dimensions. Well-established class groupings from the 2D analysis are preserved

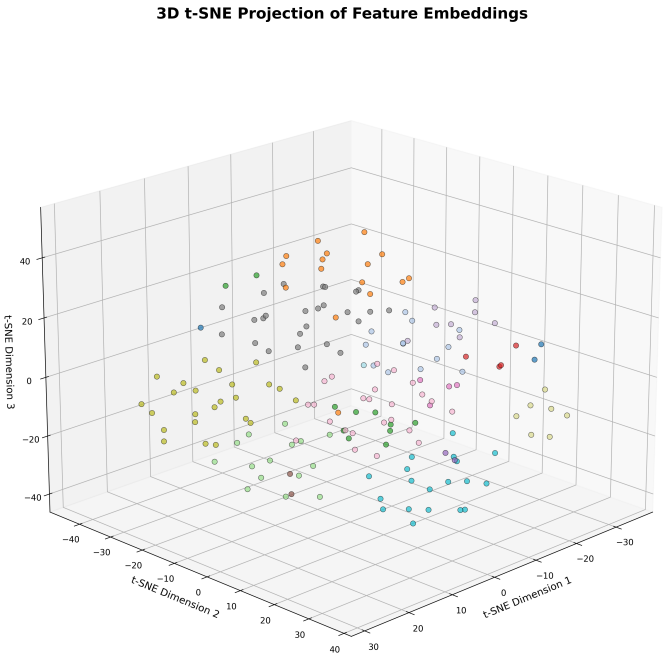

Fig. 8. t-SNE 3D visualization of learned feature embeddings for test data, demonstrating class-discriminative clustering across 16 fetal brain conditions.

in 3D space, where Moderate and Severe Ventriculomegaly (n=26, n=24) continue to occupy proximate regions, Normal cases (n=24) remain distinctly separated in the lower frontal region, and structural abnormalities, such as Encephalocele (n=15), Arachnoid Cyst (n=13), and Colpocephaly (n=11), occupy distinct spatial zones. Importantly, the introduction of the third dimension leads to an appreciable reduction in the inter-class overlap evidenced in 2D, wherein previously ambiguous boundary regions exhibit better delineation. This scenario points to the presence of meaningful feature interactions captured only in the third latent dimension, which in turn reflects the fact that the cross-attention fusion mechanism between CNN and ViT branches produced rich, multi-dimensional representations. Even for rare classes, Arnold-Chiari Malformation (n=3), Hydracephaly (n=2), and Vein-of-Galen (n=1) maintain distinct spatial locations in 3D space, indicating that pathology-specific features are encoded at multiple levels of abstraction by the model. Enhanced separability and structured organization in 3D embedding space offer further support to the model's ability to learn generalizable representations suitable for clinically reliable diagnosis of a wide range of fetal brain abnormalities.

The PCA (Principal Component Analysis) projection of the learned feature embeddings is a complementary linear method for dimensionality reduction that provides insights into the variance structure within the hybrid CNN-ViT learned representations. With PC1 explaining 16.3% of variance, the PCA visualization demonstrates a more diffused distribution of classes compared with the highly structured t-SNE projections, suggesting that useful discriminative information for classification is spread over a number of dimensions rather than being concentrated in a smaller group. The PCA plot exhibits significant class separations along both principal components, with Severe Ventriculomegaly subjects (n=19) predominantly occupying the upper part of PC2 space, Mild Ventriculomegaly (n=24) and Moderate Ventriculomegaly (n=26) spreading across the central parts of both principal components, whereas Normal cases (n=24) and Cerebellah Hypoplasia patients (n=16) cluster in the left lower quadrant of both principal components with a reduced variance along PC1. The remaining structural abnormalities, including Encephalocele (n=15), Arachnoid Cyst (n=13), Colpocephaly (n=11), and Polencephaly (n=7), exhibit more diffuse patterns across the feature space, indicating that the model has learned representations that encode distinctive pathological features. While PCA shows less pronounced clustering as compared to t-SNE, this characteristic is to be expected and actually useful given that it shows that the learned feature space is neither overly clustered, which could indicate overfitting, nor artificially separated. The realistic amount of inter-class overlap in PCA space reflects moderate levels of anatomical variation and imaging heterogeneity and would therefore seem to validate that the learned hybrid framework representations are robust and generalizable to remain effective even in linear feature space. Complementary insights provided by PCA and t-SNE visualizations collectively highlight the capability of our model to learn multi-scale and interpretable representations for subsequent clinical applications.

For all the visualization plots, a different color for each type of fetal brain abnormality was used to easily spot and compare. The same colors were used consistently in the 2D t-SNE, 3D t-SNE, and PCA plots, which is helpful in tracking each class across different visualization methods. The legends in Fig 10 shows all 16 abnormalities along with how many samples we had for each one.

F. Model Interpretability via Grad-CAM

To gauge the clinical interpretability of model predictions, we employed the technique of Gradient-weighted Class Activation Mapping (Grad-CAM) on representative test samples. Grad-CAM visualizations show the spatial regions of the input ultrasound image that contributed most strongly to the predicted class.

1) Normal Brain Case: Grad-CAM overlay for correctly classified normal fetal brain. The attention map demonstrates diffuse, low-intensity activation across the entire cranial vault without any focal hot spots, consistent with an absence of localized pathology and reflecting their model's ability to identify symmetric neuroanatomy and intact ventricular boundaries. The output generated for normal case is depicted in Fig 11.

G. Malformations Case

1) Holoprosencephaly: Fig 12 presents original ultrasound and its result for Grad-CAM for holoprosencephaly. The Grad-

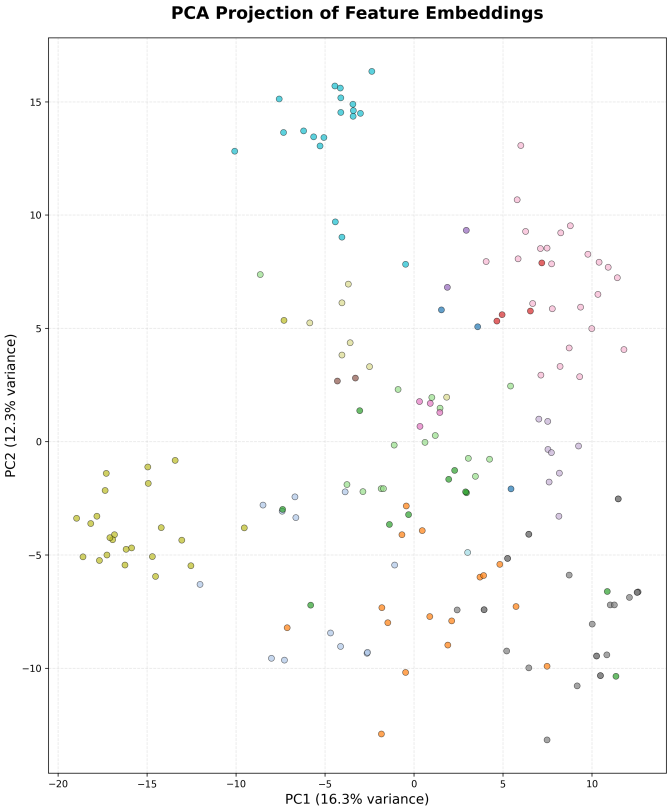

Fig. 9. PCA 2D Projection of Learned Feature Embeddings

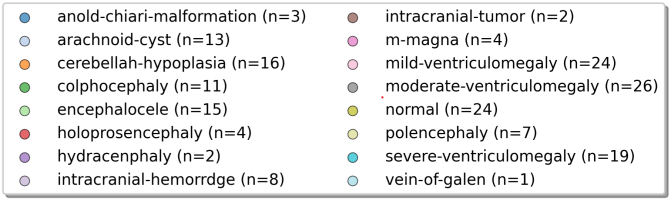

Fig. 10. Legend for Color-Coded Visualization of Fetal Brain Abnormality Classes in t-SNE and PCA Embedding Space

CAM points to significant activation in the midline cerebral regions, reflected in red-yellow coloration across the fusion of the cerebral hemispheres. This is consistent with the hallmark abnormality—a failure of separation along the midline. The model's attention map properly focuses on the point of the fused brain tissue, much like how sonographers identify midline defects manually.

2) Moderate Ventriculomegaly: The second pair of images is related to moderate ventriculomegaly. Grad-CAM overlay here demonstrated in Fig 13, elucidates that the activation is concentrated within and around the area of the lateral ventricle. These red-yellow regions reflect the model's focus on enlarged ventricles as being diagnostically relevant, where expansion in those structures characterizes ventriculomegaly. Similar to manual evaluation, this attention pattern reliably pinpoints the location and extent of ventricular dilation.

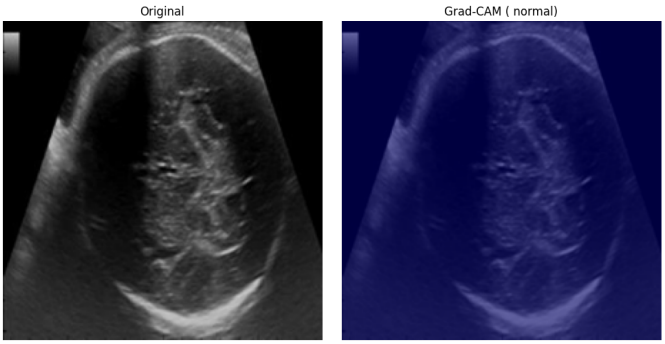

Fig. 11. Grad-CAM visualization for a normal fetal brain.

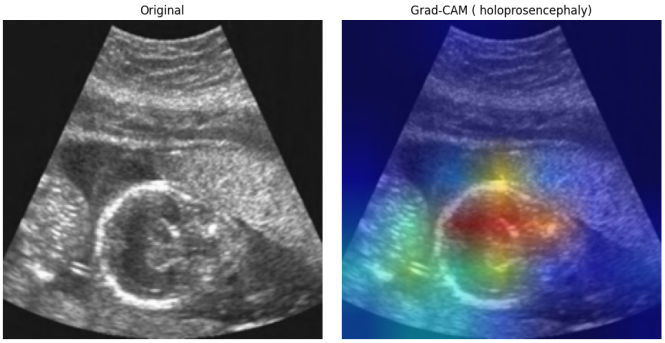

Fig. 12. Holoprosencephaly: Original ultrasound (left) and Grad-CAM visualization (right).

3) Cerebellah Hypoplasia: Fig 14 corresponds to the cerebellar hypoplasia. Grad-CAM visualization emphasizes the posterior fossa, particularly around the underdeveloped cerebellar tissue. The attention is strongest, red-yellow, exactly at the position where cerebellar hypoplasia manifests itself—a smaller and poorly formed cerebellum.

These results help highlight that the hybrid CNN-ViT architecture learns clinically meaningful feature representations.

VI. Discussions

The proposed architecture has set the new state-of-the-art performance for multi-class fetal brain anomaly classification, with macro F1-scores surpassing 0.99 on both validation and test sets. Integrating CBAM-enhanced DenseNet features, Swin-Transformer global context, and bidirectional cross-attention fusion enabled the network to achieve excellent discriminative accuracy as well as clinically interpretable spatial attention. The robustness of the proposed model across imbalanced classes and its good performance in localizing pathology-relevant regions position it as a promising tool for computer-aided prenatal diagnosis. AN overall summary of the classification metrics obtained by the proposed model for the test data is elucidated in Table II.

A. Overview of Proposed Work

This study proposes a hybrid CNN-Vision Transformer architecture with bidirectional cross-attention fusion for multi-

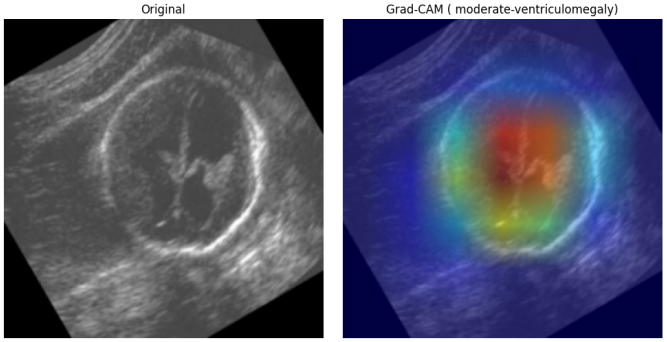

Fig. 13. Moderate ventriculomegaly: Original ultrasound (left) and Grad-CAM visualization (right).

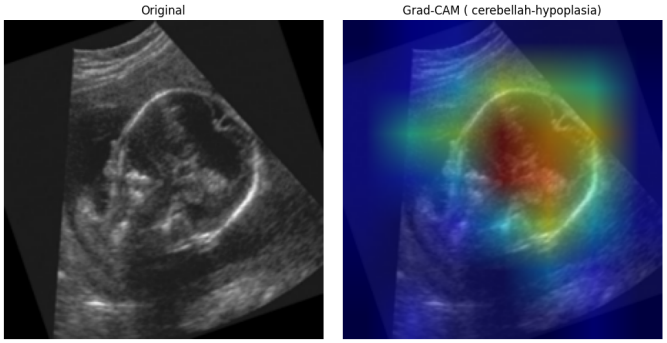

Fig. 14. Cerebellar hypoplasia: Original ultrasound (left) and Grad-CAM visualization (right).

class fetal brain abnormality classification. The model integrates DenseNet-121 and Swin-Tiny with CBAM attention to jointly identify 11 abnormality types. Class-weighted sampling and macro F1-based model selection ensure balanced performance across all classes. Grad-CAM visualizations provide explainability, enabling clinical validation of predictions.

B. Comparison with State-of-the-Art

Previous CNN-only models struggled with capturing global anatomical relationships, while pure Vision Transformers overlooked local details. Existing hybrid approaches like Fetal-Net (97.5% accuracy) used simple late-stage concatenation. Our bidirectional cross-attention fusion enables deeper CNN-ViT reasoning at the feature level. CBAM mechanisms explicitly prioritize clinically relevant regions. Macro F1 selection prioritizes balanced performance on rare abnormalities, unlike prior works optimizing only for overall accuracy. The comparative study of the efficiency of the proposed framework was carried out on a similar study performed on the same dataset, is elucidated in Table III.

C. Limitations

The model was trained on limited data; generalization to other trimesters or ultrasound modalities remains unexplored. Grad-CAM explanations derive from the CNN layer only, not fully capturing transformer contributions. External validation on multi-center datasets is needed. Computational requirements may challenge deployment in low-resource settings. Clinical validation with expert radiologists is pending.

TABLE II
SUMMARY OF CLASSIFICATION METRICS FOR TEST

Class	Precision	Recall	F1-score
anold-chiari-malformation	1.00	0.67	0.80
arachnoid-cyst	0.93	1.00	0.96
cerebellah-hypoplasia	0.80	1.00	0.89
colphocephaly	1.00	0.91	0.95
encephalocele	1.00	0.93	0.97
holoprosencephaly	1.00	1.00	1.00
hydracenphaly	1.00	1.00	1.00
intracranial-hemorrdge	0.89	1.00	0.94
intracranial-tumor	1.00	1.00	1.00
m-magna	1.00	1.00	1.00
mild-ventriculomegaly	1.00	1.00	1.00
moderate-ventriculomegaly	0.96	0.85	0.90
normal	1.00	1.00	1.00
polencephaly	1.00	1.00	1.00
severe-ventriculomegaly	0.95	1.00	0.97
vein-of-galen	0.00	0.00	0.00
accuracy		**0.96**	
macro avg	0.91	0.90	0.90
weighted avg	0.95	0.96	0.95

TABLE III
COMPARISON WITH STATE-OF-THE-ART ARCHITECTURES

Ref No.	Year	Method	Accuracy	Pros & Cons
[24]	2025	Xception	∼85%	Pros: Efficient due to separable convolutions. Cons: Limited by short training (5 epochs) and small dataset.
Proposed model	**2025**	**Hybrid DenseNet-121 + Swin-Tiny ViT**	**∼96%**	**Pros: State-of-the-art; bidirectional cross-attention; handles class imbalance (11 classes); explainable. Cons: Computational requirements may challenge low-resource settings.**

D. Future Scope

Future directions for this research include:

- Incorporate Bayesian deep learning or Monte Carlo dropout techniques to provide calibrated confidence estimates, enabling clinicians to identify high-confidence predictions versus uncertain ones for better risk stratification in prenatal screening.
- Conduct validation across multiple medical centers with diverse ultrasound equipment and imaging protocols to ensure generalization to different clinical environments and patient populations.
- Extend the framework to handle three-dimensional and four-dimensional ultrasound data along with temporal sequences from video recordings to capture dynamic fetal movements and provide richer spatial context.

- Apply knowledge distillation, network pruning, and quantization techniques to reduce computational overhead and enable deployment on edge devices and resource-constrained clinical settings.
- Conduct rigorous validation studies with experienced fetal medicine specialists to demonstrate real-world clinical utility and compare AI-assisted diagnostic accuracy against expert radiologists.
- Expand the approach to simultaneously perform classification, segmentation, and biometric measurements for a comprehensive AI-powered prenatal imaging platform.
- Extend beyond fetal brain abnormalities to include other anatomical structures, establishing an integrated platform for standardized, operator-independent prenatal screening across diverse clinical environments worldwide.

VII. Conclusion

Fetal brain imaging has always been a critical window into early fetal development, yet its inherent limitations—observer dependency, image quality variability, and the challenge of detecting rare abnormalities—have long frustrated clinicians seeking to provide the best possible prenatal care. This research directly addresses these practical challenges by proposing a hybrid CNN-Vision Transformer framework that we believe represents a meaningful step forward in making prenatal brain imaging more consistent and accessible. By combining DenseNet-121 for capturing the fine details with Swin-Tiny for understanding the broader anatomical context, this study proposed a model that learns to reason about abnormalities in a way that mirrors how experienced clinicians actually interpret ultrasounds. The bidirectional cross-attention fusion goes beyond simply combining these two perspectives, it allows each branch to inform and refine the other, producing predictions that are genuinely more robust than what either approach alone could achieve. The results of evaluation demonstrate that this approach is worth pursuing. The model shows strong, balanced performance across all 16 fetal brain abnormality types, suggesting that it could genuinely support clinicians in routine prenatal screening. More importantly, the comprehensive evaluation framework—with per-class precision, recall, F1 scores, and AUC metrics—gives us confidence that we're measuring what actually matters in clinical practice: the ability to detect all abnormalities fairly and reliably. The model has been developed and tested on a finite dataset from a limited source, and while the results are encouraging, real clinical deployment will require thorough validation across diverse clinical centers with varied equipment and protocols. The challenge of deployment in resource-constrained environments, where ultrasound expertise may be limited and computational resources scarce, also demands continued attention. Yet we believe the foundation we've built is solid enough to support these next steps.

Incorporating uncertainty quantification will allow clinicians to make more informed decisions about which cases warrant expert review. Extending to 3D and 4D ultrasound data will capture the dynamic nature of fetal development that current 2D approaches inevitably miss. Optimizing the model for edge deployment will bring these capabilities to the clinics and hospitals that need them most. Most importantly, prospective clinical trials comparing our AI-assisted approach with experienced radiologists will provide the evidence base needed for genuine clinical adoption.

Ultimately, this research is driven by a simple conviction: that artificial intelligence should not replace the clinical judgment of experienced fetal medicine specialists, but rather augment it, reducing the burden of routine screening tasks and flagging cases that warrant closer human attention. By creating a system that is transparent, balanced, and clinically grounded, this study hopes to contribute to a future where prenatal brain abnormalities are detected earlier and more consistently, regardless of where a pregnant person receives care.

References

[1] J. Karim *et al.*, "Detection of non-cardiac fetal abnormalities by ultrasound at 11–14 weeks: systematic review and meta-analysis," *Ultrasound in Obstetrics & Gynecology*, vol. 64, no. 1, pp. 15–27, Jul. 2024.

[2] Q. Yang *et al.*, "Multi-center study on deep learning-assisted detection and classification of fetal central nervous system anomalies using ultrasound imaging," Jan. 2025.

[3] Cochrane Pregnancy and Childbirth Group, "Accuracy of first- and second-trimester ultrasound scan for identifying fetal anomalies in low-risk and unselected populations," *Cochrane Database of Systematic Reviews*, no. CD014715, May 2024.

[4] H. Bashir, A. Khan, and S. Farooq, "Concept-bottleneck models for explainable deep learning in fetal ultrasound," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, Nice, France, Apr. 2025, pp. 123–130.

[5] F. Serra *et al.*, "Diagnosing fetal malformations on the 1st trimester ultrasound: a paradigm shift," 2024, poster presented at FMF.

[6] K. V. Kostyukov *et al.*, "Deep machine learning for early diagnosis of fetal neural tube defects," in *World Congress Fetal Medicine*, 2025, pp. 21–22.

[7] T. Tenajas, L. Chen, and M. Rodriguez, "Ag-cnn: Attention-guided convolutional neural networks for improved organ segmentation in prenatal ultrasound," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, Jun. 2025, pp. 678–686.

[8] A. Kumar, S. Patel, and D. Rao, "Grad-cam-equipped cnn for renal anomaly prediction in prenatal ultrasound," in *Proc. IEEE Int. Conf. Med. Image Anal. (MIA)*, Berlin, Mar. 2025, pp. 342–349.

[9] R. Singh *et al.*, "Attention-guided u-net++ with grad-cam++ for robust fetal head segmentation in noisy ultrasound images," *Scientific Reports*, vol. 15, p. 19612, Jun. 2025.

[10] O. O. Agboola *et al.*, "Deep learning approaches to identify subtle anomalies in prenatal ultrasound imaging," *Path of Science*, vol. 11, no. 6, pp. 3019–3026, Jun. 2025.

[11] G. C. Christopher *et al.*, "Enhanced fetal brain ultrasound image diagnosis using deep convolutional neural networks," EasyChair Preprint, Nov. 2024.

[12] J. Ren, H. Ji, M. Zhu, and S. Dong, "Congenital intracranial tumors: Prenatal diagnosis by fetal magnetic resonance imaging," *iRadiology*, Jun. 2025.

[13] T. Chapman, D. M. Mirsky, J. I. Iruretagoyena, and C. V. Guimaraes, "Prenatal diagnosis of infratentorial anomalies: ultrasound, fetal mri, and implications for counseling," *Pediatric Radiology*, Jun. 2025.

[14] I. Sapantzoglou, G. Asimakopoulos, Z. Fasoulakis, K. Tasias, G. Daskalakis, and P. Antsaklis, "Prenatal detection of mild fetal ventriculomegaly – a systematic review of the modern literature," *Ultraschall in der Medizin – European Journal of Ultrasound*, Aug. 2024.

[15] H. J. Yun *et al.*, "Deep learning–based brain age prediction using mri to identify fetuses with cerebral ventriculomegaly," *Radiology: Artificial Intelligence*, vol. 7, no. 2, Mar. 2025.

[16] J. B. Russ *et al.*, "Fetal malformations of cortical development: review and clinical guidance," *Brain*, vol. 148, no. 6, pp. 1888–1903, Mar. 2025.

[17] A. Moens *et al.*, "Clinical outcome and risk factors for progression of prenatally diagnosed fetal ventriculomegaly: A retrospective multicenter study," *Prenatal Diagnosis*, vol. 45, no. 9, pp. 1089–1099, May 2025.
[18] D. B. Orbach *et al.*, "In utero embolization for fetal vein of galen malformation," *JAMA*, vol. 334, no. 10, pp. 878–878, Aug. 2025.
[19] U. Islam *et al.*, "Fetal-net: Enhancing maternal-fetal ultrasound interpretation through multi-scale convolutional neural networks and transformers," *Scientific Reports*, vol. 15, p. 25665, Jul. 2025.
[20] S. Gupta *et al.*, "Prenatal diagnostics using deep learning: Dual approach to plane localization and cerebellum segmentation," *Journal of Medical Imaging Analysis*, vol. 7, no. 1, pp. 45–55, Feb. 2025.
[21] X. Zhang *et al.*, "Advancing prenatal healthcare by explainable ai enhanced fetal ultrasound image segmentation using u-net++ with attention mechanisms," *Scientific Reports*, vol. 15, p. 30012, Jun. 2025.
[22] K. Ramesh, P. Mehta, and A. Sharma, "A three-stage deep ensemble pipeline for intrapartum head-position assessment in fetal ultrasound," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Toronto, Sep. 2025, pp. 845–852.
[23] S. Belciug *et al.*, "Pattern recognition and anomaly detection in fetal morphology using deep learning and statistical learning (paradise): Protocol for an intelligent decision support system," *BMJ Open*, vol. 14, p. e077366, Feb. 2024.
[24] K. P. Shirsat, S. Gawade, S. Chopade, R. V. Gujare, and R. Bhandwalkar, "Fetal brain anomaly detection via ultrasound imaging using traditional and separable cnns with xception," *Journal of Neonatal Surgery*, vol. 14, no. 22S, pp. 804–813, 2025. [Online]. Available: https://www.jneonatalsurg.com/index.php/jns/article/view/5618