Automated Maternal Fetal Ultrasound Image Identification Using a Hybrid Vision Transformer Model

Thunakala Bala Krishna[1], Ajay Kumar Reddy Poreddy[1], Kolla Gnapika Sindhu[2], and Priyanka Kokil[1](✉)

[1] Advanced Signal and Image Processing (ASIP) Lab, Department of Electronics and Communication Engineering, Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, Chennai 600127, India
{ec21d0004,edm20d012,priyanka}@iiitdm.ac.in

[2] Amrita School of Computing, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Chennai, Tamil Nadu, India

Abstract. Ultrasound (US) technology has revolutionized prenatal care by offering noninvasive, real-time visualization of maternal-fetal anatomy. The accurate classification of maternal-fetal US planes is a critical segment of effective prenatal diagnosis. However, the inherent inter-class variance among different fetal US images presents a significant hurdle, making fetal anatomy detection a laborious and time-consuming task, even for experienced sonographers. This paper proposes a novel approach using a Hybrid Vision Transformer (H-ViT) for automated fetal anatomical plane classification to address these challenges. The proposed method utilizes hierarchical features extracted from DenseNet-121, which are then inputted into the vision transformer to analyze complex spatial relationships and patterns within fetal US images. By incorporating both global and local features, the proposed method enhances feature discriminability, thus alleviating low inter-class variance. The effectiveness of the H-ViT is evaluated using the largest publicly available maternal-fetal US image dataset. The experimental results rigorously demonstrate the superiority of our approach, achieving an accuracy of 96.60% compared to other state-of-the-art techniques.

Keywords: Fetal ultrasound classification · Vision transformer · Convolutional neural network · Deep learning · Maternal–fetal planes

1 Introduction

Ultrasound (US) imaging has revolutionized the field of prenatal care, providing healthcare professionals with a non-invasive and real-time visualization tool to assess the well-being of both the mother and the developing fetus [1]. It allows for examining critical anatomical structures, tracking fetal growth, and identifying potential abnormalities at various stages of pregnancy [2]. Traditionally, the

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 253–269, 2025.
https://doi.org/10.1007/978-3-031-78195-7_17

identification of maternal-fetal US planes has relied on the expertise of highly skilled sonographers, who manually analyze US images to locate and interpret specific anatomical landmarks [3]. However, this process is time-consuming, subject to inter-observer variability, and heavily reliant on the operator's experience and training. As a result, there is a growing need for automated and standardized approaches to enhance the efficiency and accuracy of maternal-fetal US plane detection. From a medical practitioner's point of view, the evaluation of fetal US images can be quite demanding because they often contain distortions like acoustic shadows, speckle noise, motion blur, and unclear boundaries. These distortions arise due to the intricate interplay between US waves and the tissues of both the mother and the fetus [4]. The comprehensive study and accurate classification of US fetal planes, encompassing crucial aspects such as the fetal abdomen, fetal brain, fetal femur, fetal thorax, maternal cervix, and other pertinent planes, hold immense potential to transform healthcare and positively impact society. They constitute a proactive approach towards enhancing prenatal care, estimating fetal weight [5], promoting maternal and fetal health [6], and fostering a higher quality of life for expectant mothers and their offspring [7]. Figure 1 presents the most frequently employed fetal anatomical structures in prenatal screening to estimate fetal well-being. Classifying US fetal planes aids in gestational age prediction by providing critical anatomical markers and measurements, enabling a more accurate estimation of the fetus's developmental stage and age [9].

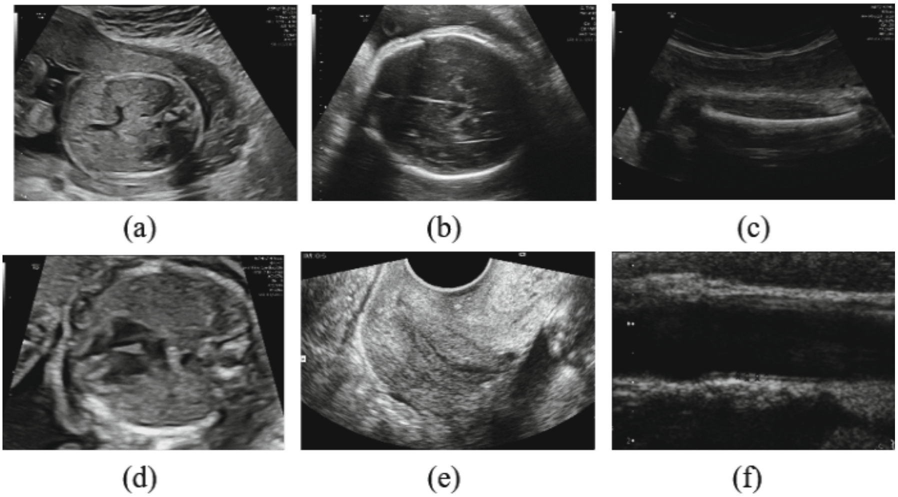

Fig. 1. Illustration of the commonly used fetal US images: (a) fetal abdomen (b) fetal brain (c) fetal femur (d) fetal thorax (e) fetal cervix (f) other [8].

Deep learning techniques, especially convolutional neural networks (CNNs), have demonstrated remarkable success in medical imaging tasks [10]. CNNs excel at learning complex hierarchical representations from large datasets, enabling them to extract meaningful features and classify images with high accuracy [11]. Their application in US imaging has shown promising results in various areas, such as fetal biometry, organ segmentation, and anomaly detection. Xavier *et al.* [12] evaluated various deep CNNs on fetal US images and concluded that DenseNet-169 has a close correlation with human technicians. Sridar *et al.* [13] computed the local features of US images using pre-trained CNNs and the global attributes using the discriminant regions of the US planes. Further, the final decision is made by fusing the decisions computed from the individual attributes using the support vector machine (SVM) classifier. Rasheed *et al.* [14] computed the frame level predictions of fetal US videos using AlexNet and segmented the fetal head frames using UNET. Further, the segmented frames are utilized to calculate the biparietal diameter (BPD) and head circumference (HC) via segmented contours.

Thunakala *et al.* [15] developed a feature-fused model for fetal US planes using ResNet-50 and the AlexNet models. Further, the fused features are fed to SVM for final prediction. Yu *et al.* [16] developed a CNN model for fetal facial detection using augmentation and fine-tuning techniques. In [17], the authors utilized SonoNet architecture to automatically detect 13 fetal standard views and a bounding box localization mechanism using weak supervision from image-level labels. Yang *et al.* [18] developed a radial component mechanism (RCM) that visualizes the key geometric characteristics of fetal abdominal planes. Further, the critical attributes identified from RCM are given to a random forest (RF) classifier to identify fetal abdominal and non-abdominal classes. In [19], the authors extracted dense and region of interest attributes of fetal US images using Fisher vector, transform descriptor, and the Gaussian mixture model. Further, the computed attributes are given to SVM to identify sagittal, axial, and coronal US planes. In [20], the authors computed the spatial features of fetal US videos using MobileNet architecture and the temporal attributes using recurrent neural networks. Further, these cues are fused using a two-stage mechanism to identify the four classes of fetal US videos. In [21], the authors fused the deep features computed from AlexNet and VGG-19 model and fed to multilayer perceptron to classify the six planes of fetal US images. In [9,22], the authors developed an ensemble network by fine-tuning the top-performing CNNs, and the predictions from stacked CNNs are given to the majority voting classifier for final prediction. In [23], the authors reduced the interclass variance among the fetal US planes using VGG-16 and adaptive weighting mechanism, and then the features were fed to the softmax classifier to identify six classes of US planes. Although existing fetal US image classification methods [9,12–16,18–22] have produced satisfactory results, several limitations remain. The following highlights the limitations of the existing methods:

1. Only certain fetal anatomical structures, such as the fetal brain and maternal cervix, are accurately interpreted. However, other prominent planes like the fetal femur, abdomen, and thorax are often misclassified as non-standard planes.

2. The inherent low interclass variance of fetal US images needs to be addressed more effectively.
3. There is a necessity to develop DL models that focus primarily on significant regions of the fetal anatomy while minimizing attention to artifacts such as background information.

Recently, the introduction of hybrid vision transformer models has further advanced the field of image classification. Vision transformers (ViTs), originally proposed for natural image analysis, leverage the transformer architecture's ability to capture long-range dependencies and self-attention mechanisms. By incorporating both convolutional and transformer-based layers, hybrid Vision transformers (H-ViTs) combine the strengths of both approaches, potentially improving the performance of image classification tasks. In this paper, we have proposed to combine the effectiveness of CNN architectures and ViTs to automate the detection of common maternal-fetal US planes. We hypothesize that combining CNNs and ViTs will leverage the benefits of spatial feature extraction from CNNs and the attention-based capabilities of transformers, leading to enhanced accuracy and addressing the inherent low interclass variance of fetal US images. To the best of our knowledge, none of the works in the existing literature [9,12–24] proposed a hybrid model based on deep CNN and transformer to classify fetal US planes. This has motivated us to develop an H-ViT model to reduce inter-subject variability among the critical anatomical structures of fetal US images. Our study comprehensively evaluates these models on a diverse dataset, considering various factors such as computational efficiency, interpretability, and generalization performance. By automating the detection of common maternal fetal ultrasound planes, we envision a future where healthcare professionals can benefit from standardized and efficient analysis, enabling more accurate diagnoses, improved patient care, and ultimately, better outcomes for both the mother and the fetus. The following are the contributions of the proposed method:

1. A deep learning model called H-ViT, which improves the classification of commonly used fetal anatomical structures by enhancing the interclass variance among fetal planes is proposed.
2. The H-ViT model combines DenseNet-121 and Vision Transformer (ViT). Spatial feature maps extracted from the DenseNet-121 backbone are fed into the ViT, which further refines the fetal US attributes using the attention-based capabilities of transformers.
3. To ensure the robustness and reliability of the proposed model, we further assessed its performance using a speckle-introduced fetal US image dataset, notably under noisy conditions.

The remainder of the paper is structured as follows: Section 2 delve into the dataset details and introduce the H-ViT architecture proposed for classifying fetal US images. Section 3 reports the experimental results and validation studies, and Section 4 concludes the paper by summarizing the findings and discussing potential future research directions.

2 Methodology

This section describes the fetal US dataset, the preprocessing steps applied to the fetal US images, the feature extraction process, and an elucidation of the proposed H-ViT for fetal US image classification.

Table 1. Distribution of dataset over different classes

Fetal Plane	No of Samples
Fetal Abdomen	711
Fetal Femur	1040
Fetal Brain	3092
Fetal Thorax	1718
Maternal Cervix	1626
Other	4216
Total	12400

2.1 Dataset Description

A large and diverse dataset comprising routinely acquired maternal-fetal screening US images is utilized to evaluate the effectiveness of the proposed H-ViT [8]. The dataset was collected from two hospitals involving multiple operators and US machines. An expert maternal-fetal clinician has meticulously labeled all the images in the dataset to ensure accuracy. The dataset consists of 12,400 images, categorized into six distinct classes as mentioned in Table 1. Four classes represent the most commonly used fetal anatomical planes: abdomen, brain, femur, and thorax. Additionally, there is a class named the maternal cervix, often utilized for prematurity screening. Finally, a general category encompasses any other less frequently encountered image plane.

2.2 Data augmentation

In order to enhance the robustness and generalization of the proposed classification method, a systematic data augmentation strategy is employed on the maternal-fetal screening US dataset. We have increased the number of samples in each class of the fetal US dataset using data augmentation techniques such as cropping, rotating, translating, and flipping images [25]. Data augmentation enhances dataset diversity and reduces data imbalance, enabling assessing the robustness and generalization of the classification model [26]. This augmentation approach aims to diversify the dataset by generating variations of the original images while preserving their anatomical and clinical characteristics [25]. The original fetal US dataset exhibits variations in the number of samples across the distinct fetal plane categories, with some classes having considerably fewer samples compared to others, as shown in Table 1. After augmenting the fetal US image dataset, the number of images corresponding to each class has increased to 2000. Additionally, the augmented fetal US dataset is divided into 70% for training, 15% for validation, and 15% for testing the CNN and ViT models.

2.3 Preprocessing

The proposed approach employs appropriate preprocessing techniques on the fetal US image dataset to facilitate training and evaluation. Preprocessing steps include rescaling the pixel values of the US images to a range of 0 to 1, shearing the images to introduce geometric variations, applying zooming transformations, and horizontal flipping. These steps enhance the ability of the classification model to generalize and enhance its performance. Furthermore, the images are resized to a standardized resolution of 224×224×3 pixels, ensuring consistency across the dataset. Additionally, as part of the preprocessing pipeline, speckle noise is introduced to the US images. Speckle noise, a multiplicative noise commonly encountered in US imaging due to interference patterns, is simulated to evaluate the proposed model strength to this artifact and assess its robustness in real-world scenarios.

2.4 Feature Extraction

The feature extraction process is crucial for capturing informative representations from input images in the proposed methodology. To achieve this, the pre-trained DenseNet-121 [27] CNN is utilized as our backbone feature extraction network. DenseNet-121 has demonstrated its effectiveness in various computer vision tasks due to its dense connectivity and feature reuse characteristics. The DenseNet-121 is pre-trained on a large ImageNet dataset to acquire a generalized understanding of low- to mid-level features. The pre-trained backbone CNN is fine-tuned on the target fetal US image dataset, enabling the extraction of hierarchical features with varying levels of abstraction [28]. During feature extraction, fetal US images are passed through the DenseNet-121 backbone, undergoing a series of convolutional and pooling layers. The dense blocks within the architecture generate feature maps that contain increasingly abstract representations as we move deeper into the network, as shown in Figure 2. These feature maps capture both low-level details and high-level semantic information, making them suitable for a wide range of downstream tasks.

2.5 Proposed H-ViT model

In this section, we introduce a novel and synergistic methodology for precise image classification, leveraging the remarkable capabilities of the ViT architecture. The ViT represents a groundbreaking approach in computer vision and image analysis [29]. Unlike traditional CNNs, ViT relies on self-attention mechanism, allowing it to effectively capture long-range dependencies in images, as shown in Fig 2. This innovative architecture has shown remarkable performance in various image-related tasks, demonstrating its potential to revolutionize the field of image classification.

Building upon the strengths of ViT, our approach combines it with the computational prowess of the DenseNet-121 backbone [30]. The harmonious fusion of cutting-edge techniques results in an efficacious and proficient model explicitly designed for achieving precise and accurate image classification.

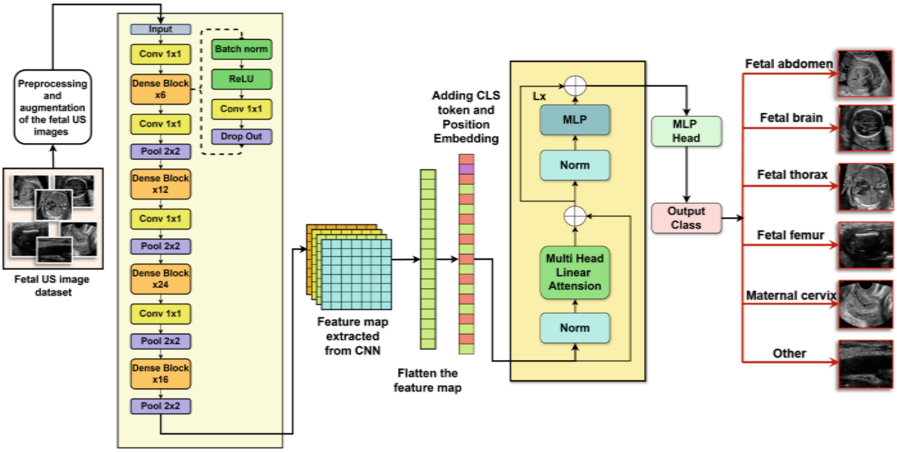

Fig. 2. Proposed model for Fetal Plane classification

Patch-based representation: Upon computing feature extraction by the backbone DenseNet-121 model, we strategically partition the resultant feature maps into discrete patches. The feature map partitioning is the foundation of our patch-based representation strategy. It facilitates a synergistic amalgamation of localized context and holistic perceptual insights derived from fetal US images. The partitioning strategy aims to capture fine-grained local intricacies and global contextual cues, fostering a comprehensive and nuanced understanding of fetal US image information.

Position embedding and class embedding: Position embedding is a strategy utilized in transformer models to encode the position of each image patch. CNNs, through their architecture, intrinsically preserve the spatial structure of the input images, but transformers process input data as a sequence. To address this, position embeddings are added to the input sequence to inject details about the position of each patch into the model. Class embedding is a unique token added to the input sequence in transformer models, particularly in the context of classification tasks. In ViT, a class embedding token (often called the [CLS] token) is prepended to the sequence of image patches. In the proposed H-ViT model, position embeddings are added to each image patch to provide spatial information, and a class embedding token is prepended to the sequence. These embeddings allow the transformer model to understand the spatial relationships between patches and gather a holistic representation for classification purposes.

Transformer encoder: unveiling complex relationships: Operationally, the encoder initiates by taking the enriched patch-based representations. Through self-attention, the encoder perceptively estimates the significance of each patch in relation to its adjacent spatial values, thereby capturing intricate interdependencies that might elude conventional convolutional constructs. The intrinsic capability empowers the model to discern nuanced spatial hierarchies and complex cross-patch relationships, fostering a granular understanding of image content.

Furthermore, the self-attention process is perpetuated through the multi-head mechanism. The model adeptly encapsulates the fetal US image feature relationships and patterns by executing parallel attention computations, each capturing distinct aspects of inter-patch relationships. Subsequently, the outputs of these multi-faceted computations are harmoniously concatenated and subject to linear transformations, yielding a synthesis that matches diverse attentional perspectives.

The concatenated multi-head attention outputs are then channeled through position-wise feedforward neural networks, infusing the process with non-linearity and intricate processing. The iterative procedure unfolds across multiple encoder layers, thereby fostering successive refinement of patch relationships and representations.

Multi-layer perceptron The final fetal US image classification is estimated using the multi-layer perceptron (MLP) head. The MLP is a robust, interconnected structure comprised of various dense layers aimed at high-level feature refinement. MLP head subjects the amalgamated information to iterative transformations, fostering intricate non-linear abstractions. Combined with non-linearity, the sequential application of operations empowers the model to extract salient high-level semantic facets encoded within the input fetal US data.

2.6 Speckle Noise

Speckle noise is inherent to US imaging and can significantly impact the quality of acquired images [31]. By introducing synthetic speckle-noise at different levels, we simulate real-world US imaging conditions. Including speckle noise in the training dataset gives the model a more realistic representation of imaging scenarios. This helps the model adapt to noise patterns and improves its robustness against noisy inputs during inference. By training the model on a diverse dataset that includes speckle noise, we expect the model to generalize better to unseen real-world data, which typically exhibit speckle artifacts. The Rayleigh

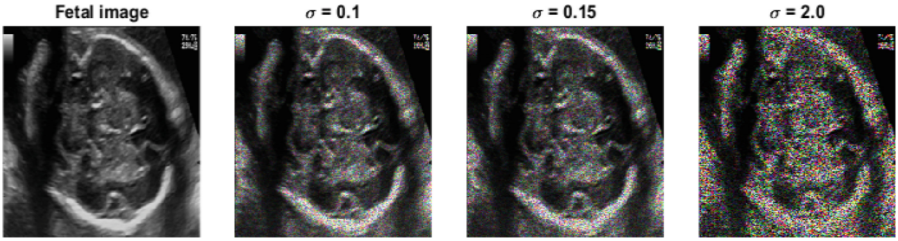

Fig. 3. Speckled noise fetal US images with different levels.

distribution is often appropriate for describing the amplitude of the received US signals, as speckle noise tends to exhibit a Rayleigh distribution when the signals are coherent [32]. The probability density function representing the Rayleigh distribution is given as:

$$\mathbb{P}\left(z;\sigma\right) = \frac{z}{\sigma^2}\exp\left(-\frac{z^2}{2\sigma^2}\right), \tag{1}$$

where z represents the row vectored input US image and σ denotes the spread of the speckle noise. In this paper, the generalization of the proposed model is evaluated by considering the different values of σ as mentioned in Figure 3.

3 Experimental results

In this section, we present the empirical results of our proposed methodology, which leverages a combination of DenseNet-121 and ViT for maternal fetal US image classification task. We conducted a comprehensive experimental analysis, including training and evaluation, to assess the effectiveness and robustness of our approach.

3.1 Experimental Setup

To compare the performance of the H-ViT model, we selected four prominent DL architectures: VGG19, Xception, InceptionV3, and DenseNet-121. After training on the fetal dataset, we observed that DenseNet-121 consistently outperformed the other architectures across a range of evaluation metrics. As a result, we selected DenseNet-121 as the backbone for our proposed approach. Further, to enhance the model's generalization capabilities and resilience to noise, we added different levels of speckle noise to the fetal US images. The incorporation of speckle noise not only improved the model's robustness but also contributed to achieving enhanced classification accuracy.

3.2 Evaluation Metrics

We assessed the performance of H-ViT and existing models using a comprehensive set of evaluation metrics, such as accuracy, precision, recall, and F1 score. Accuracy delivers insights into the ratio of correctly predicted instances to the total number of instances in the testing dataset, providing an overall measure of the model's correctness. The proportion of true positive predictions out of the total positive predictions made by the model gives the measure of precision. It measures the model's ability to avoid false positives. Recall indicates the proportion of true positive predictions from the actual positive instances in the testing dataset, which describes the model's capacity to capture positive instances. F1 score is the harmonic mean of precision and recall. It provides a balanced assessment of both metrics and is especially useful when dealing with imbalanced classes.

Table 2. Performance comparison of the proposed method (in %) with the backbone CNN and ViT model

Model	Accuracy	Precision	Recall	F1 score
DenseNet-201	93.00	93.07	93.00	93.01
ViT-base16	89.30	90.81	88.20	89.23
H-ViT	**96.33**	**96.77**	**96.36**	**96.56**

Table 3. Performance comparison of proposed model with state-of-the-art deep CNN models.

Model	Train Acc	Test Acc	Test Loss	Precision	Recall	F1-score	Test Time
VGG19	0.911	0.889	0.331	0.892	0.889	0.889	119s
InceptionV3	0.957	0.931	0.231	0.931	0.931	0.931	215s
Xception	0.969	0.927	0.222	0.927	0.927	0.927	150s
DenseNet-121	0.974	**0.940**	0.208	0.930	0.93	0.930	79s
ViT	0.997	0.893	0.311	0.908	0.882	0.892	50s
Proposed (H-ViT)	0.981	**0.963**	0.188	0.967	0.963	0.965	**37s**

3.3 Performance of the proposed H-ViT model

We have evaluated the performance of the backbone CNN model (DenseNet-121) and vision transformer (ViT) individually as an ablation study, and the corresponding results are tabulated in Table 2. The experimental results show

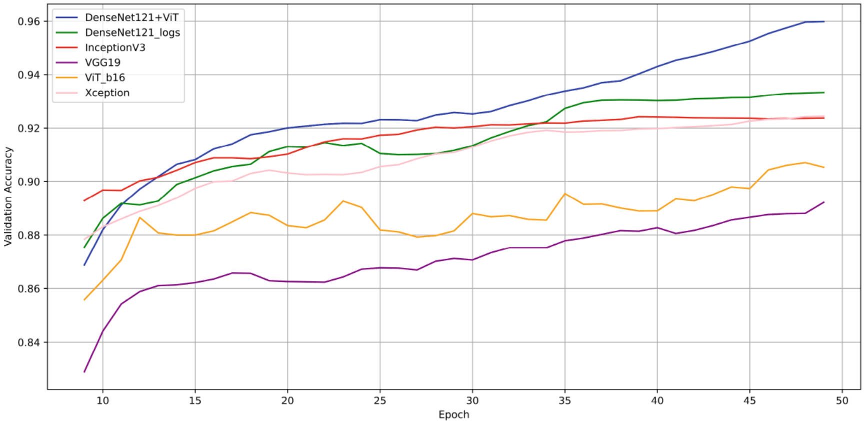

Fig. 4. Accuracy values of the state-of-the-art CNN models for different number of epochs.

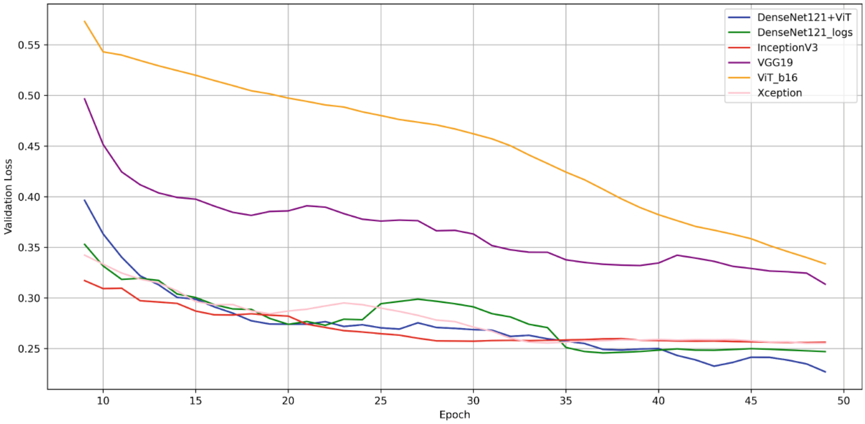

Fig. 5. Validation loss of the state-of-the-art CNN models for different number of epochs

that the CNN model has achieved better than the ViT base model. However, the proposed H-ViT model achieves superior results by leveraging the refined feature maps extracted from the backbone CNN model fed into the ViT. A comprehensive overview of the performance metrics for various existing models, including VGG19, InceptionV3, Xception, DenseNet-121, ViT, and the proposed hybrid model (H-ViT) is provided in Table 3. The metrics include training accuracy, testing accuracy, loss, precision, recall, F1-score, and testing time. From Table 3, one may observe that the proposed H-ViT model achieved remarkable improvements in terms of performance metrics showcasing its superior capabilities in capturing intricate visual patterns of fetal US planes. Figure 4 visualizes the accuracy values for different numbers of epochs on different state-of-the-art CNN

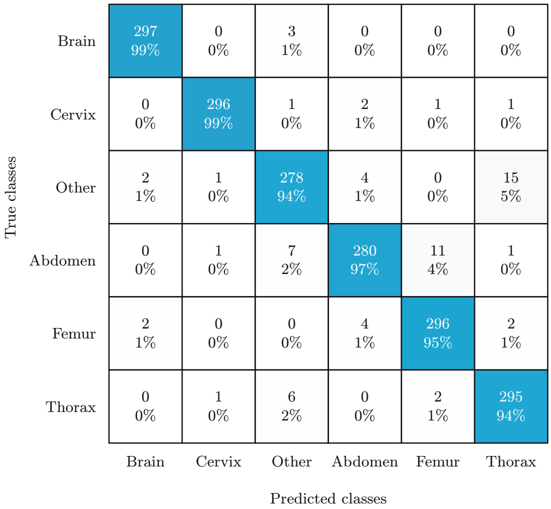

Fig. 6. Confusion matrix of the proposed H-ViT model for different classes of the fetal US dataset.

Table 4. Performance results of the DenseNet and the proposed H-ViT architecture for different levels of speckle noise.

Noise level	Test Accuracy	Test Loss	Precision	Recall	F1-score
DenseNet-121+Noise					
σ=0.1	0.878	0.366	0.887	0.875	0.876
σ=0.15	0.806	0.639	0.848	0.808	0.814
σ=0.2	0.725	0.978	0.814	0.727	0.735
DenseNet-121+ViT+Noise					
σ=0.1	0.928	0.210	0.929	0.928	0.929
σ=0.15	0.898	0.3000	0.900	0.893	0.894
σ=0.2	0.873	0.368	0.881	0.871	0.873

models. From Figure 4, it is clear that H-ViT (DenseNet-121+ViT) achieved the best accuracy values, demarcating its efficacy in identifying the key anatomical attributes of the fetal US planes. Similarly, Figure 5 depicts the validation loss of different CNN models across different numbers of epochs. The plot clearly shows that the H-ViT model's loss cure slowly converges toward zero as the number of epochs increases, indicating its suitability for precision-driven tasks. Figure 6 illustrates the confusion chart of the proposed fetal US image classification model, depicting the class-wise performance in terms of accuracy. Additionally, the confusion plot provides insights into the potential misinterpretations made by the proposed model for each class. It is concluded that the majority of fetal US images are accurately classified, except for the other class, which exhibits structural similarities with the remaining classes, leading to misclassification. Also, we performed an ablation study by introducing speckle noise to the fetal US planes to check the generalization capability of the proposed H-ViT model. The models trained with speckle noise augmentation exhibited enhanced generalization, effectively mitigating the adverse effects of noise and variability in real-world scenarios. Table 4 illustrates the performance results of the DenseNet-121 model and the proposed H-ViT modeling in a noisy environment. The table depicts performance metrics that are slightly declined by increasing the level of speckle-noise across three different levels of speckle noise on DenseNet-121 architecture. However, embedding ViT into the DenseNet-121 model increases the performance values, and the performance results are slightly increased compared to the DenseNet-121 model. This indicates that the proposed H-ViT model performs consistently across different levels of speckle noise, highlighting its efficacy in the noisy environment.

3.4 Comparison with the existing works

The fetal US image classification performance of the proposed method is further compared with the competing works [12,16,17,21,23,33,34] to show the effectiveness. The quantitative performance measures are presented in Table 5. The results presented in the table show that the H-ViT model achieved superior outcomes than the competing methods.To compare the proposed and existing works fairly, all competing methods are trained and tested on the fetal US dataset used to evaluate the H-ViT model.

Table 5. Performance comparison of the proposed model (in %) with the competing methods.

Method	Accuracy	Precision	Recall	F1 score
Xavier et al. [12]	93.73	91.95	93.08	92.5
Zhen Yu et al [16]	94.12	92.87	94.25	93.48
Baumgartner et al. [17]	94.48	93.56	93.06	93.28
Krishna et al. [21]	95.10	93.83	95.00	94.38
Krishna et al. [23]	95.33	93.58	95.84	94.64
HaifaGhabri et al [33]	93.63	91.50	93.59	92.48
Sendra et al. [34]	94.94	93.17	95.36	94.19
Proposed method	**96.33**	**96.77**	**96.36**	**96.56**

3.5 Ablation study

The effectiveness of the proposed method is also evaluated through an ablation study using different dataset sizes and splits, including five fold cross-validation.

We acknowledge the importance of 5-fold cross-validation in evaluating machine learning models. To address this concern, we have conducted additional experiments using 5-fold cross-validation. The cross-fold validation results are provided in Table 6, and the experimental results provide a more comprehensive evaluation of the proposed model performance. Additionally, we evaluated the performance of the proposed model using different data splits: 60%, 70%, and 80% for training (train), 20%, 15%, and 10% for validation (vali), and 20%, 15%, and 10% for testing (test), respectively. The qualitative results for these experiments are provided in Table 7. From the table, one may observe that the 70:15:15 split yields superior performance due to more balanced class distribution across training, validation, and testing sets, enhancing the model's ability to learn and classify unseen instances accurately. The 15% validation set size provides sufficient evaluation data, improving performance while mitigating overfitting. This balanced approach ensures effective training and evaluation. These results are highlighted with brown color text in the revised manuscript.

Table 6. Five fold Cross-validation results

Model	Accuracy	Precision	Recall	F1 score
Fold 1	96.33	96.77	96.36	96.56
Fold 2	92.66	93.27	92.41	92.84
Fold 3	95.58	95.97	95.41	95.69
Fold 4	95.16	95.24	95.08	95.16
Fold 5	94.66	94.97	94.41	94.69

Table 7. Performance comparison of the proposed method under various dataset splits

Split ratio (train:vali:test)	Accuracy	Precision	Recall	F1 score
60:20:20	85.20	86.79	84.08	85.41
70:15:15	96.33	96.77	96.36	96.56
80:10:10	95.49	95.57	95.49	95.60

4 Conclusion

This article introduces an H-ViT that integrates a DenseNet-121 backbone with transformer architecture to enhance the automatic classification of maternal-fetal US planes during prenatal screening. To assess the effectiveness of the proposed approach, we utilized a publicly available fetal US image dataset obtained from high-resource settings. Analysis of experimental results, conducted with various noise levels, demonstrates the classification model's generalization capability across diverse fetal US image qualities. The proposed method accurately identifies frequently investigated fetal structures, offering valuable support to sonographers and obstetricians in monitoring fetal development and early detection of complications.

In the future, it would be interesting to incorporate more advanced preprocessing techniques, such as noise reduction and artifact removal, which can further enhance the quality of input images. Consequently, the performance of the model might be improved.

References

1. Wells, P.N.: Ultrasound imaging. Physics in Medicine & Biology **51**(13), R83 (2006)
2. Levine, D.: Ultrasound versus magnetic resonance imaging in fetal evaluation. Top. Magn. Reson. Imaging **12**(1), 25–38 (2001)
3. Huang, Q., Zeng, Z., et al.: A review on real-time 3D ultrasound imaging technology. BioMed Research International 2017 (2017)
4. Meng, L., Zhao, D., Yang, Z., Wang, B.: Automatic display of fetal brain planes and automatic measurements of fetal brain parameters by transabdominal three-dimensional ultrasound. J. Clin. Ultrasound **48**(2), 82–88 (2020)
5. Hadlock, F.P., Harrist, R., Sharman, R.S., Deter, R.L., Park, S.K.: Estimation of fetal weight with the use of head, body, and femur measurements–a prospective study. Am. J. Obstet. Gynecol. **151**(3), 333–337 (1985)
6. Turan, S., Miller, J., Baschat, A.A.: Integrated testing and management in fetal growth restriction. In: Seminars in Perinatology. vol. 32, pp. 194–200. Elsevier (2008)
7. Nicolaides, K.H., Syngelaki, A., Ashoor, G., Birdir, C., Touzet, G.: Noninvasive prenatal testing for fetal trisomies in a routinely screened first-trimester population. Am. J. Obstet. Gynecol. **207**(5), 374-e1 (2012)
8. Burgos-Artizzu, X.P., Coronado-Gutierrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: FETAL_PLANES_DB: Common maternal-fetal ultrasound images (Jun 2020), https://doi.org/10.5281/zenodo.3904280
9. Krishna, T.B., Kokil, P.: Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. Expert Syst. Appl. **238**, 122153 (2024)
10. Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S.: A review on deep-learning algorithms for fetal ultrasound-image analysis. Med. Image Anal. **83**, 102629 (2023)
11. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T.: Deep learning in medical ultrasound analysis: A review. Engineering **5**(2), 261–275 (2019)
12. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Sci. Rep. **10**(1), 10200 (2020)
13. Sridar, P., Kumar, A., Quinton, A., Nanan, R., Kim, J., Krishnakumar, R.: Decision fusion-based fetal ultrasound image plane classification using convolutional neural networks. Ultrasound in Medicine & Biology **45**(5), 1259–1273 (2019)
14. Rasheed, K., Junejo, F., Malik, A., Saqib, M.: Automated fetal head classification and segmentation using ultrasound video. IEEE Access **9**, 160249–160267 (2021)

15. Krishna, T.B., Kokil, P.: Automated detection of common maternal fetal ultrasound planes using deep feature fusion. In: IEEE 19th India Council International Conference (INDICON). pp. 1–5. Kochi, India (2022)
16. Yu, Z., Tan, E.L., Ni, D., Qin, J., Chen, S., Li, S., Lei, B., Wang, T.: A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition. IEEE J. Biomed. Health Inform. **22**(3), 874–885 (2017)
17. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging **36**(11), 2204–2215 (2017)
18. Yang, X., Ni, D., Qin, J., Li, S., Wang, T., Chen, S., Heng, P.A.: Standard plane localization in ultrasound by radial component. In: 11th International Symposium on Biomedical Imaging (ISBI). pp. 1180–1183. IEEE (2014)
19. Lei, B., Zhuo, L., Chen, S., Li, S., Ni, D., Wang, T.: Automatic recognition of fetal standard plane in ultrasound image. In: 11th International Symposium on Biomedical Imaging (ISBI). pp. 85–88. IEEE (2014)
20. Pu, B., Li, K., Li, S., Zhu, N.: Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Trans. Industr. Inf. **17**(11), 7771–7780 (2021)
21. Krishna, T.B., Kokil, P.: Automated classification of common maternal fetal ultrasound planes using multi-layer perceptron with deep feature integration. Biomed. Signal Process. Control **86**, 105283 (2023)
22. Sindhu, K.G., R, A.: Ensemble-based advancements in maternal fetal plane and brain plane classification for enhanced prenatal diagnosis. International Journal of Information Technology pp. 1–17 (2024)
23. Krishna, T.B., Kokil, P.: Integration of a deep convolutional neural network with adaptive channel weight technique for automated identification of standard fetal biometry planes. IEEE Trans. Instrum. Meas. **73**, 1–11 (2024)
24. Ma'Sum, M.A., Jatmiko, W., Tawakal, M.I., Al Afif, F.: Automatic fetal organs detection and approximation in ultrasound image using boosting classifier and Hough transform. In: International Conference on Advanced Computer Science and Information System. pp. 460–467. IEEE (2014)
25. Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F.: Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610 (2022)
26. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: Experimental evaluation. Inf. Sci. **513**, 429–441 (2020)
27. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
28. Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., Mittal, A.: Pneumonia detection using CNN based feature extraction. In: International Conference on Electrical, Computer and Communication Technologies (ICECCT). pp. 1–7. IEEE (2019)
29. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 357–366 (2021)
30. Angelina, C.L., Chou, Y.K., Lee, T.C., Kongkam, P., Han, M.L., Wang, H.P., Chang, H.T.: Hybrid vision transformer for classification of pancreatic cystic lesions on confocal laser endomicroscopy videos. In: International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan). pp. 47–48. IEEE (2023)

31. Duarte-Salazar, C.A., Castro-Ospina, A.E., Becerra, M.A., Delgado-Trejos, E.: Speckle noise reduction in ultrasound images for improving the metrological evaluation of biomedical applications: An overview. IEEE Access **8**, 15983–15999 (2020)
32. Tuthill, T., Sperry, R., Parker, K.: Deviations from Rayleigh statistics in ultrasonic speckle. Ultrason. Imaging **10**(2), 81–89 (1988)
33. Ghabri, H., Alqahtani, M.S., Ben Othman, S., Al-Rasheed, A., Abbas, M., Almubarak, H.A., Sakli, H., Abdelkarim, M.N.: Transfer learning for accurate fetal organ classification from ultrasound images: A potential tool for maternal healthcare providers. Sci. Rep. **13**(1), 17904 (2023)
34. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., Ahmed, Y.A., Elattar, M., Ohene-Botwe, B., Nyangulu, P., Stones, W., Ammar, M., Benamer, L.N., et al.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries. Sci. Rep. **13**(1), 2728 (2023)