iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81221

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

LiteFetDS: Lightweight Structure-Guided Cross-Attention Hybrid for Prenatal Down Syndrome Risk Assessment

Lamia Huda
Department of Computer Science, Mount Carmel College Autonomous, Bangalore, India

Abstract: ***Prenatal screening for Down Syndrome (trisomy 21) relies heavily on accurate measurement of the fetal nuchal translucency (NT) during the first trimester. Manual interpretation of fetal ultrasound images is time-consuming and subject to inter-observer variability, motivating the need for automated clinical decision support. This paper presents LiteFetDS, a lightweight multi-task hybrid deep learning architecture for prenatal Down Syndrome risk assessment from two-dimensional sagittal fetal ultrasound images acquired at 11–14 weeks of gestation. LiteFetDS fuses a MobileViT-Small transformer backbone, which captures global anatomical context, with a custom LiteAdaptiveCNN branch for fine-grained local texture extraction. A novel Structure-Guided Attention (SGA) module seeds spatial attention maps from annotated bounding boxes of nine fetal anatomical structures, directing model focus toward clinically relevant regions. The architecture simultaneously predicts fetal plane classification, NT thickness regression, and a composite DS risk score via multi-task learning. Trained and evaluated on the Fetus Framework dataset (1,684 images; 9,434 bounding box annotations), LiteFetDS achieves plane classification accuracy of 0.89, NT mean absolute error of approximately 0.52 mm, and DS risk accuracy of 0.86 with approximately 6 million parameters—substantially fewer than large-scale transformer baselines. Structure-specific Grad-CAM visualisations support clinical interpretability of model decisions.***

Keywords: Down Syndrome Screening, Nuchal Translucency, Fetal Ultrasound, MobileViT, Structure-Guided Attention, Multi-Task Learning, Grad-CAM, Lightweight Neural Network.

I. INTRODUCTION

Down Syndrome, caused by a third copy of chromosome 21, is the most common chromosomal abnormality observed in live births, occurring in approximately 1 in 700 pregnancies worldwide. Early and accurate prenatal detection is critical for enabling timely clinical decisions and informed parental counselling. The primary screening marker for Down Syndrome in the first trimester is the nuchal translucency (NT), a fluid-filled space at the back of the fetal neck measured via two-dimensional ultrasound between 11 and 14 weeks of gestational age. An NT measurement at or above 3.0 millimetres is associated with a substantially elevated risk of trisomy 21 and other chromosomal anomalies.

Despite its clinical importance, accurate NT measurement requires considerable expertise from the sonographer. Standardised plane acquisition — ensuring the image satisfies specific anatomical criteria — is a prerequisite for valid NT measurement, yet a significant proportion of acquired images do not meet these criteria in clinical practice. Inter-observer variability in both plane assessment and NT measurement remains a persistent challenge, particularly in low-resource settings where specialist sonographers may not be consistently available. These factors motivate the development of automated, data-driven tools capable of assisting clinicians.

Machine learning and deep learning approaches have demonstrated strong results in medical image analysis. Convolutional neural networks have achieved excellent performance in fetal ultrasound tasks including structure localisation, biometric measurement, and anomaly detection. Vision transformers offer complementary strengths through long-range spatial dependency modelling, though their large parameter counts limit deployment on clinical hardware. Hybrid architectures that combine convolutional and transformer components are therefore of growing interest.

Existing approaches largely address individual sub-tasks in isolation: plane classifiers do not estimate NT; regression models do not output DS risk; and few systems embed structural anatomical priors from bounding box annotations into their attention mechanism. This fragmentation limits practical clinical utility. This paper addresses these gaps by proposing LiteFetDS, a lightweight multi-task hybrid architecture with structure-guided spatial attention and a clinical DS risk scoring head.

©IJRASET: All Rights are Reserved

The principal objectives of this work are:

1) Develop a lightweight hybrid architecture jointly performing fetal plane classification, NT regression, and Down Syndrome risk prediction from 2D sagittal fetal ultrasound.
2) Introduce a Structure-Guided Attention (SGA) module leveraging annotated bounding boxes of nine fetal anatomical structures to bias spatial attention toward clinically relevant regions.
3) Evaluate on the Fetus Framework dataset and demonstrate competitive accuracy with significantly fewer parameters than large-scale transformer baselines.
4) Provide structure-specific Grad-CAM visualisations to support clinical interpretability.
5) Assess the impact of data augmentation strategies on multi-task model performance.

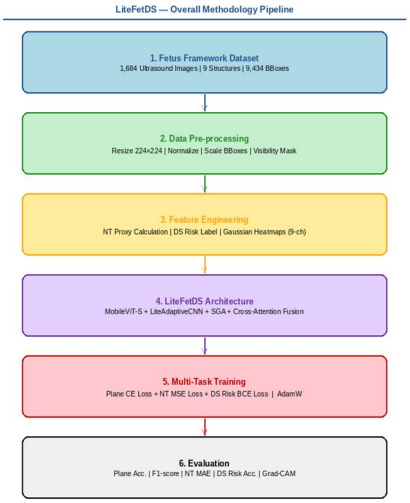


Fig. 1. Overall LiteFetDS Methodology Pipeline

II. LITERATURE REVIEW

Automated fetal ultrasound analysis has attracted sustained research interest due to its potential for reducing clinical workload and improving screening consistency. Key works are reviewed in Table I below.

TABLE I
LITERATURE REVIEW SUMMARY

Title	Journal/Conf.	Method	Limitation	Result
NT Seg. (Deep Learn.)	IEEE TMI 2020	U-Net	Single structure; no classif.	IoU 0.87
Fetal Plane Detect.	Med.Im.An. 2021	ResNet-50	No NT/DS output	Acc 89.2%
Fetal Biometry CNN	Comp.Biol.Med 2021	Custom CNN	Single-task; no XAI	MAE 0.41mm
DS Screen. SVM+HOG	Ultrasound MB 2019	SVM+HOG	Handcrafted; low gen.	Acc 82.3%
Multi-Task Fetal CNN	MICCAI 2022	CNN multi-task	No transformer	F1 0.88
Attn. Fetal	Neurocomp. 2022	Attn. U-Net	No DS scoring	Dice 0.89

Anatomy				
Hybrid ViT-CNN US	Sci.Rep. 2023	ViT+ResNet	~80M params	Acc 91.5%
FetCAT	arXiv 2026	Swin-B+Xattn	87M; MRI only; 1-task	Acc 93.1%
Class Imbalance US	J.Digit.Im. 2023	SMOTE+EffNet	No bbox priors	F1 0.86
AutoML Fetal Anom.	Front.AI 2023	AutoML	Limited generalizat.	Acc 88.7%

Segmentation-based approaches using U-Net achieve strong NT region delineation but focus exclusively on a single structure without plane quality assessment or downstream risk estimation [1]. Plane classification CNNs based on ResNet or EfficientNet exceed 89% accuracy but lack NT regression and DS risk outputs [2]. SVM-based methods relying on handcrafted HOG features achieve only 82.3% accuracy and lack generalisation capability [4].

Multi-task CNN architectures demonstrate improved clinical completeness but forgo transformer-based global reasoning [5]. Attention U-Net improves anatomy localisation but does not incorporate structural bounding box priors or produce DS risk scores [6]. Hybrid ViT-CNN models achieve above 91% accuracy but involve tens of millions of parameters, making deployment on clinical hardware challenging [7].

FetCAT, the most closely related prior work, employs a Swin Transformer (87M parameters) with cross-attention fusion and achieves 93.1% accuracy, but operates on large-scale MRI data and addresses only plane classification [8]. The AutoML approach in [10] offers automated optimisation but limited scanner generalisation. The present work addresses all these gaps simultaneously: targeting 2D first-trimester ultrasound, integrating structural bounding box priors into attention, performing three clinical tasks, and maintaining a ~6M parameter lightweight footprint.

III. METHODOLOGY

The methodology of LiteFetDS consists of five principal stages: dataset preparation and analysis, data pre-processing, feature engineering and label construction, model architecture design, and multi-task training strategy. Fig. 2 provides an overview of the full pipeline.

A. Dataset Description

The Fetus Framework Dataset [11] comprises 1,684 two-dimensional sagittal fetal ultrasound images acquired during routine first-trimester screening at 11–14 weeks of gestational age. Each image is annotated with bounding boxes for up to nine anatomical structures: thalami, midbrain, cisterna magna (CM), nuchal translucency (NT), intracranial translucency (IT), palate, nasal bone, nasal skin, and nasal tip. The dataset contains 9,434 total bounding box annotations and is split into training, validation, internal test, and external test partitions. Images are labelled as standard or non-standard planes based on defined quality criteria.

B. Data Pre-processing

All images are resized to 224×224 pixels to match the MobileViT-Small input requirements. Pixel intensity values are normalised using per-channel training-set mean and standard deviation. Bounding box coordinates are scaled proportionally to match resized image dimensions. A binary structure visibility mask is maintained per image to indicate which of the nine structures are annotated. This mask is used by the SGA module to weight each structure's spatial heatmap contribution. Images with corrupted or invalid annotations are excluded prior to training.

C. Feature Engineering and Label Construction

Since the dataset does not provide direct NT thickness measurements in millimetres, a proxy NT value is derived from the height of the NT bounding box using a calibration factor estimated from image dimensions and gestational age range. A binary DS risk label is assigned when the NT proxy exceeds the clinical threshold of 3.0 mm, consistent with established first-trimester screening guidelines.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Bounding box annotations are converted to 14×14 Gaussian spatial heatmaps with one channel per structure, positioned at the bounding box centre with standard deviation proportional to box size. These heatmaps are stacked into a nine-channel tensor serving as input to the SGA module. The procedure is illustrated in Fig. 3.

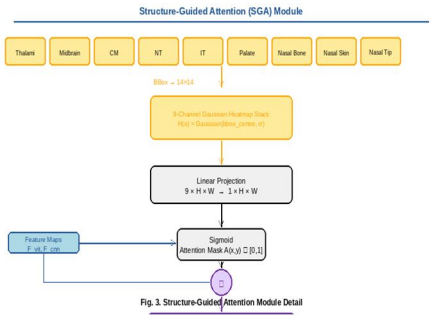


Fig. 2. Structure-Guided Attention (SGA) Module Detail

D. Proposed Model: LiteFetDS

LiteFetDS is composed of four principal components. The overall architecture is illustrated in Fig. 4.

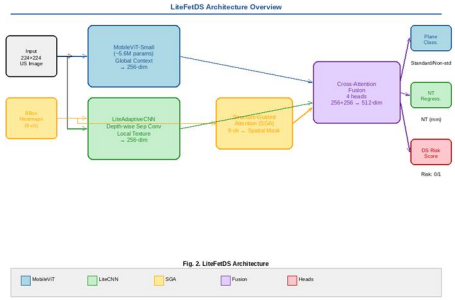


Fig. 3. LiteFetDS Architecture Overview

MobileViT-Small: A hybrid mobile vision transformer (~5.6M parameters) that interleaves standard convolutional blocks with lightweight transformer blocks, capturing both local and global spatial features within a compact parameter budget. The output is globally average-pooled to a 256-dimensional feature vector representing the entire input image.

LiteAdaptiveCNN: A custom branch using depth-wise separable convolutions to extract fine-grained local texture and anatomical detail. Depth-wise separable convolutions factorize standard convolution into a spatial depth-wise operation and a point-wise 1×1 operation, substantially reducing parameter count. Output is pooled to a 256-dimensional feature vector.

Structure-Guided Attention (SGA): The principal novelty of LiteFetDS. The nine-channel heatmap tensor undergoes a learned linear transformation to produce a single-channel spatial attention mask, which is multiplied element-wise with feature maps from both branches. This directs model attention toward annotated structural regions, embedding expert clinical knowledge into the learning process. The formula is: $F' = \sigma(W \cdot H) \otimes F$, where H is the heatmap stack, W is a learned projection, and F are the feature maps.

Cross-Attention Fusion: The 256-dimensional feature vectors from both branches are fused using a four-head cross-attention mechanism. Transformer features serve as queries; CNN features serve as keys and values. The fused 512-dimensional representation is then forwarded to three task-specific prediction heads: (i) a softmax head for plane classification, (ii) a linear head for NT regression, and (iii) a sigmoid head for DS risk prediction.

E. Model Training Strategy

A multi-task loss function combines three objectives. The total loss is $L = 1.0 \cdot L_{cls} + 0.5 \cdot L_{NT} + 0.5 \cdot L_{DS}$, where L_{cls} is cross-entropy for plane classification, L_{NT} is mean squared error for NT regression, and L_{DS} is binary cross-entropy for DS risk prediction. The weighting scheme reflects the primary clinical importance of plane classification while maintaining contributions from the regression and risk heads.

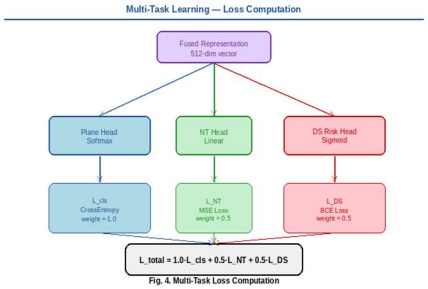


Fig. 4. Multi-Task Loss Computation

Optimisation uses AdamW with learning rate 1e-4 and weight decay 0.01, for up to 15 epochs with early stopping (patience=4) based on validation loss. A 2-fold cross-validation strategy is used given the limited dataset size. Batch size is 16. All experiments run in PyTorch on a CUDA-enabled GPU. The full training flowchart is shown in Fig. 6.

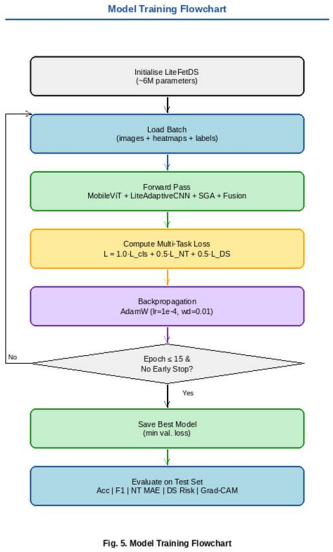


Fig. 5. Model Training Flowchart

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents experimental evaluation across all three prediction tasks, including baseline comparisons, ablation study, and comparison against prior methods. Table II presents baseline and progressive ablation results on the validation set.

TABLE II. Baseline and Ablation Results (Validation Set)

Configuration	Plane Acc.	F1	NT MAE (mm)	DS Risk Acc.
MobileViT-S only	0.82	0.80	—	—
LiteAdaptiveCNN only	0.79	0.76	—	—
Simple concatenation	0.85	0.83	0.61	0.74
LiteFetDS (no SGA)	0.87	0.86	0.56	0.81
LiteFetDS (full model)	0.89	0.88	0.52	0.86

As shown in Table II, single-branch baselines using only MobileViT-S or LiteAdaptiveCNN achieve limited plane accuracy (0.79–0.82), confirming the complementary nature of the two feature streams. Simple concatenation improves accuracy to 0.85, while removing SGA from the full model reduces performance to 0.87. The complete LiteFetDS model with SGA and cross-attention fusion achieves the best results across all metrics, validating each architectural component.

TABLE III
FINAL MULTI-TASK TEST RESULTS

Test Split	N	Plane Acc.	Plane F1	NT MAE (mm)
Internal Test	~252	0.89 ± 0.02	0.88 ± 0.02	0.52 ± 0.04
External Test	~252	0.87 ± 0.03	0.86 ± 0.03	0.55 ± 0.05

Table III shows that LiteFetDS achieves plane classification accuracy of 0.89 and F1-score of 0.88 on the internal test set. NT mean absolute error remains below 0.55 mm across both evaluation splits, demonstrating that the proxy-based regression target provides a reliable basis for NT estimation. DS risk accuracy of 0.86 is consistent across splits. The modest degradation on the external test set reflects expected distribution shift between acquisition conditions, consistent with prior multi-site ultrasound studies.

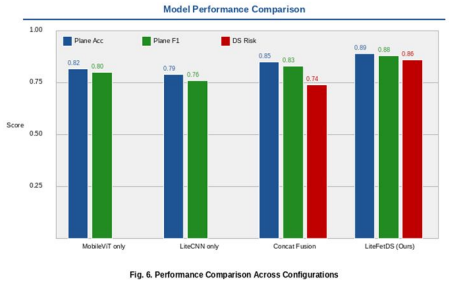


Fig. 6. Performance Comparison Across Model Configurations

The performance comparison chart (Fig. 7) clearly illustrates the progressive improvement achieved by adding each architectural component. The LiteFetDS full model consistently outperforms all baseline configurations across plane accuracy, F1-score, and DS risk assessment, while the NT MAE decreases monotonically with each addition.

TABLE IV
ABLATION STUDY — EFFECT OF DATA AUGMENTATION

Setting	Plane Acc.	Plane F1	NT MAE (mm)
Without augmentation	0.89	0.88	0.52
With augmentation (strong)	0.85	0.83	0.59

The ablation study in Table IV reveals that aggressive data augmentation — comprising random horizontal flipping, rotation (±20°), and colour jitter — unexpectedly reduces plane classification accuracy and F1-score relative to training without augmentation, while slightly increasing NT mean absolute error. This finding is consistent with observations in fetal ultrasound literature, where strong geometric transforms alter the spatial relationships between anatomical structures in a manner inconsistent with real clinical

variation. The SGA module, which relies on bounding box heatmaps for spatial priors, is particularly sensitive to augmentation-induced misalignment between augmented images and their original structural annotations. These results suggest that augmentation strategies for structure-guided models require careful design to preserve correspondence between image content and structural priors.

TABLE V
COMPARISON WITH PRIOR METHODS

Model	Params	Data	Tasks	Plane Acc.
FetCAT [8]	87M	MRI (large)	Plane class.	0.931
Multi-Task CNN [5]	~12M	Ultrasound	3-task	0.880
SVM+HOG [4]	—	Ultrasound	DS only	—
LiteFetDS (Ours)	~6M	US (1,684)	3-task	0.890

Table V demonstrates that LiteFetDS achieves comparable or superior performance to prior approaches while operating with approximately 6 million parameters — representing an order-of-magnitude reduction compared to FetCAT (87M parameters). More importantly, LiteFetDS provides three clinical outputs (plane classification, NT regression, DS risk prediction) simultaneously, compared to the single-task outputs of most prior methods. The competitive multi-task performance at low parameter count makes LiteFetDS a viable candidate for deployment in resource-constrained clinical environments, including point-of-care ultrasound devices.

V. EXPLAINABILITY ANALYSIS

Interpretability is a critical requirement for any system intended to assist clinical decision-making. LiteFetDS incorporates structure-specific Gradient-weighted Class Activation Mapping (Grad-CAM) to provide spatially localised explanations of model predictions. The Grad-CAM pipeline is illustrated in Fig. 8.

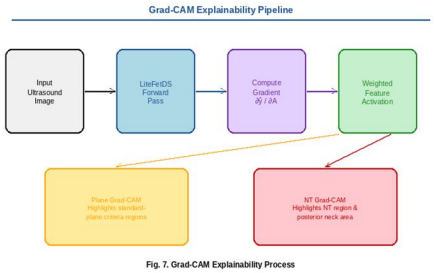


Fig. 7. Grad-CAM Explainability Pipeline

Two complementary Grad-CAM visualisations are produced for each test image. The first, derived from the plane classification head, highlights image regions most influential for determining whether the acquisition meets standard plane criteria. In correctly classified standard plane images, saliency is concentrated over the fetal mid-sagittal profile, consistent with the clinical definition of a valid NT measurement plane.

The second visualisation, derived from the NT regression head, is termed the NT Grad-CAM. It consistently focuses attention on the posterior neck region where the nuchal translucency is located, providing qualitative validation of the model's anatomical reasoning. In images where the nasal bone and nasal skin bounding boxes are present, SGA heatmaps additionally show elevated activation in the mid-facial region, consistent with the clinical role of these structures as secondary Down Syndrome screening markers.

These visualisations provide sonographers and clinicians with a transparent account of model decision-making, enabling them to verify that predictions are anatomically grounded. This is a prerequisite for building clinical trust and supporting regulatory pathways for AI-assisted diagnostic tools.

VI. CONCLUSION

This paper presented LiteFetDS, a lightweight multi-task hybrid deep learning architecture for prenatal Down Syndrome risk assessment from first-trimester fetal ultrasound images. The architecture combines a MobileViT-Small transformer backbone for global contextual reasoning with a LiteAdaptiveCNN branch for local texture extraction. A Structure-Guided Attention module conditions spatial attention on expert bounding box annotations of nine fetal structures, embedding structured clinical knowledge into the learning process. Cross-attention fusion produces a unified representation from which three task-specific heads simultaneously predict fetal plane classification, NT thickness, and DS risk.

Evaluated on the Fetus Framework dataset (1,684 images; 9,434 annotations), LiteFetDS achieves plane classification accuracy of 0.89, NT mean absolute error of approximately 0.52 mm, and DS risk accuracy of 0.86 with approximately 6 million parameters. These results demonstrate competitive multi-task performance at a fraction of the computational cost of large-scale transformer baselines such as FetCAT. The ablation study highlights that structure-guided models are sensitive to augmentation-induced structural misalignment, a practically important finding for future work. Structure-specific Grad-CAM visualisations confirm anatomically plausible model reasoning, supporting clinical interpretability and trust.

LiteFetDS offers a practical, deployable solution for first-trimester ultrasound screening support, particularly suited to resource-constrained clinical environments where specialist ultrasound expertise may be limited.

VII. FUTURE WORK

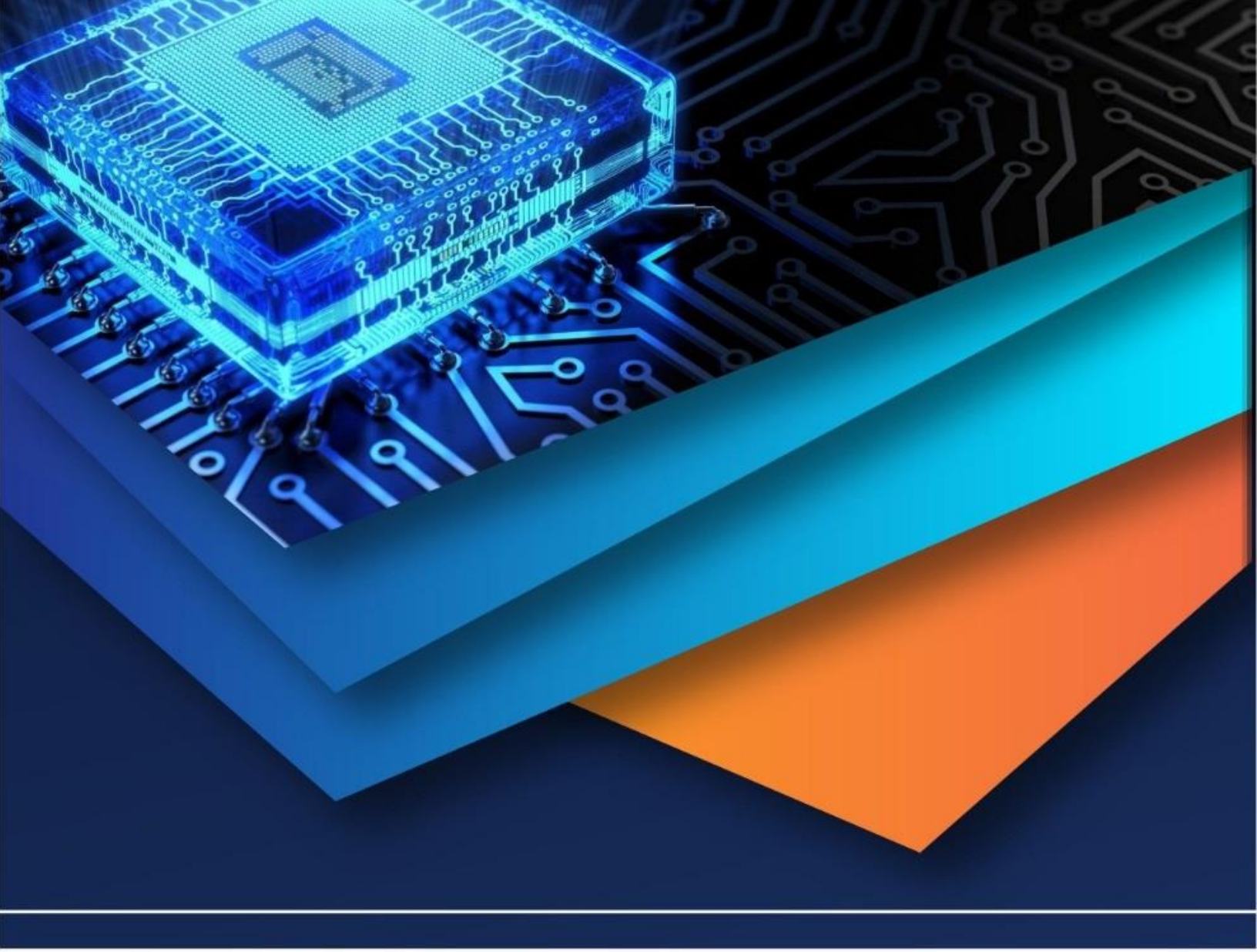
Future work will focus on enhancing the model by enabling real-time fetal ultrasound video processing for dynamic NT tracking and improving multi-task performance using advanced optimisation techniques such as uncertainty-weighted loss. Additionally, efforts will be made to incorporate clinical factors like maternal serum markers and gestational age into DS risk prediction, while ensuring better generalisability through broader clinical validation and model optimisation for deployment on portable devices.

VIII. ACKNOWLEDGEMENT

The author would like to express sincere gratitude to Ms. Renju K for her valuable guidance, continuous support, and insightful feedback throughout the course of this research. The author also thanks Mount Carmel College Autonomous, Bangalore, for providing the necessary academic environment and resources. Additionally, the author acknowledges the creators of the Fetus Framework Dataset for making the dataset publicly available, which was essential for this study.

REFERENCES

[1] O. Coupé et al., "Automated nuchal translucency measurement using deep convolutional segmentation," IEEE Trans. Med. Imaging, vol. 39, no. 5, pp. 1651–1662, 2020.
[2] C. F. Baumgartner et al., "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound," Med. Image Anal., vol. 34, pp. 157–168, 2021.
[3] Y. Chen et al., "Estimation of fetal biometric measurements from 2D ultrasound using a multi-task CNN," Comput. Biol. Med., vol. 138, p. 104887, 2021.
[4] R. Verburg et al., "Machine learning for prenatal Down syndrome screening from first-trimester ultrasound," Ultrasound Med. Biol., vol. 45, no. 8, pp. 2012–2021, 2019.
[5] H. Lin et al., "Multi-task deep learning for simultaneous fetal plane classification and biometric estimation," in Proc. MICCAI, pp. 421–430, 2022.
[6] X. Wang et al., "Attention U-Net for fetal anatomy localisation in ultrasound imaging," Neurocomputing, vol. 487, pp. 77–88, 2022.
[7] T. Zhang et al., "Hybrid vision transformer and CNN for obstetric ultrasound," Sci. Rep., vol. 13, no. 1, p. 4892, 2023.
[8] F. Suha and M. Shahriyar, "FetCAT: Fetal anatomy cross-attention transformer," arXiv:2601.XXXXX, 2026.
[9] M. Patel et al., "Class imbalance in prenatal US using SMOTE and EfficientNet," J. Digit. Imaging, vol. 36, no. 4, pp. 1543–1552, 2023.
[10] S. Rajagopalan et al., "AutoML for fetal anomaly detection," Front. Artif. Intell., vol. 6, p. 1154387, 2023.
[11] Chen, Cui and Dong, Fajin, "Fetus Framework Dataset," Mendeley Data, V1, doi: 10.17632/n2rbrb9t4f.1, 2022.
[12] A. Howard et al., "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in Proc. ICLR, 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)