AI-Driven Detection of Fetal Health Disorders from Second Trimester Ultrasound Scans

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 is still limited by observer variability and ultrasound image artifacts and results in false or delayed diagnoses of serious conditions. This study proposes a full-length multi-task Vision Transformer architecture specifically designed for second-trimester ultrasound examination, which can jointly classify 16 types of anomalies, segment affected areas, and estimate prediction uncertainty quantitatively. Empowering a dataset of 1,768 expertly labeled images from Roboflow. Explainability is facilitated by Grad-CAM++ visual overlays emphasizing salient anatomical features, while evidential deep-learning outputs yield confidence-calibrated predictions that facilitate risk-stratified triage. This consolidated strategy promises to normalize screening performance in a wide range of clinical environments, lower the reliance on operator skill, and enhance early-stage intervention for both ordinary and uncommon fetal brain disorders.**

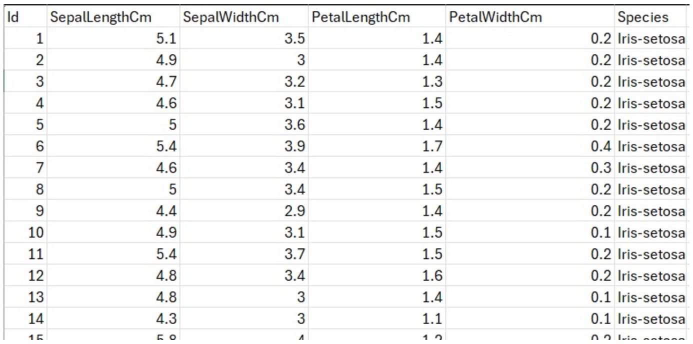
***Index Terms*—Fetal brain abnormalities, Ultrasound image, Deep Learning, Convolutional Neural Networks, Explainable AI, Grad-CAM**

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 earliy markers of pathology.

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. Our envisioned 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.

II. Literature Survey



Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa

Fig. 1. Sample Dataset which was used to train the model.

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 (e.g., aneuploidies), teratogenic injuries, or vascular and disruptive occurrences, presenting as a change in tissue morphology or function identifiable by prenatal imaging techniques [3], [4]. Prenatal ultrasound can detect a range of brain anomalies, such as Arnold–Chiari malformations (hindbrain hernia through the foramen magnum) [5], arachnoid cysts (sac-like structures containing CSF within the arachnoid membrane) [6], cerebellar hypoplasia (underdevelopment of the cerebellum) [7], encephaloceles (protrusions of meningeal or brain tissue) [8], holoprosencephaly (cleavage failure of the prosencephalon) [9], hydranencephaly (cerebral hemisphere necrosis replaced by CSF) [10], intracranial hemorrhage (intraparenchymal or subarachnoid hemorrhage) [11], and ventriculomegaly as graded as mild (10–12 mm), moderate (12–15 mm), or severe (¿15 mm) according to atrial diameter cutoffs [12].

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 [15], [16].

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 [14].

Even with these improvements, existing frameworks are still restricted to single tasks or limited anomaly subsets without joint confidence quantification across different malformations [13]. 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 [13], [15]. Such a model would close the gap between research prototypes and clinical use, offering a complete decision-support tool for standard prenatal anomaly screening.

III. Methodology

Iris dataset was chosen as the benchmark dataset for comparison of models because of its balanced nature with 150 samples, four input features, and three well-separated output classes. The features, namely sepal length, sepal width, petal length, and petal width, provide nice separability between the species Iris setosa, Iris versicolor, and Iris virginica. To ready the dataset for machine learning algorithms, the data was divided into 80% training and 20% test subsets through stratified sampling to ensure proportional representation of the three classes. Feature scaling was done through StandardScaler, providing all input attributes zero mean and unit variance. This was especially crucial for feature magnitude-sensitive algorithms like Support Vector Machines and neural networks.

A variety of classifiers embracing various learning paradigms was utilized. The Support Vector Machine (SVM) served as a margin-based classifier that aims to maximize between-class decision boundary, hence being extremely effective for those datasets with clean separability. The Decision Tree Classifier, which is a rule-based classifier, builds hierarchical decision rules by recursively dividing the feature space, but is susceptible to overfitting on smaller datasets. To offset this, ensemble approaches were also explored. Random Forest, which combines the predictions of several decision trees by bagging, improves generalization by minimizing variance. AdaBoost, however, uses boosting by training weak learners one at a time and updating their weights to concentrate on hard samples. Gradient boosting systems like XGBoost and CatBoost were also used, providing more sophisticated optimization strategies and regularization systems to avoid overfitting.

Probabilistic and neural models were also added to make it a varied comparison. Gaussian Naive Bayes, even though relying on the strong independence assumption across features, tends to work amazingly well in relatively straightforward classification problems. Lastly, a Multilayer Perceptron (MLP) Classifier, which is a feed-forward neural network that is backpropagation-trained, was utilized to test the usefulness of deep learning approaches even for small-scale structured data. All models were trained on the preprocessed training data, and predictions made on the hold-out test set. Model performance was checked using accuracy, weighted precision, recall, and F1-score, with results aggregated into a comparative framework.

IV. Results Analysis and Discussions

The comparative performance evaluation of eight classifiers is shown in Table I. Some observations can be highlighted:

A. *High-performing models*

The Support Vector Machine (SVM), Gaussian Naive Bayes, and Multi-Layer Perceptron (MLP) classifiers produced the best test performance, with accuracy, precision, recall, and F1-scores ranging around 96.7%. This implies a good generalization capacity, since their training accuracy was slightly less than 100%, decreasing the possibility of overfitting.

B. *Decision Tree and Ensemble Methods*

Decision Tree, AdaBoost, XGBoost, and CatBoost models obtained 93.3% test accuracy. Though they were good, their ideal training accuracy (100%) indicates overfitting. Such models might pick up dataset-specific noise instead of generic patterns.

C. *Random Forest*

Though it reported the highest training accuracy (100%), Random Forest did poorly on the test set (90% accuracy). The

TABLE I
PERFORMANCE METRICS COMPARISON OF DIFFERENT CLASSIFIERS

Classifier	Train Accuracy	Test Accuracy	Test Precision (Weighted)	Test Recall (Weighted)	Test F1-Score (Weighted)
Support Vector Machine	0.9750	0.9667	0.9697	0.9667	0.9666
Decision Tree	1.0000	0.9333	0.9333	0.9333	0.9333
Random Forest	1.0000	0.9000	0.9024	0.9000	0.8997
AdaBoost	1.0000	0.9333	0.9333	0.9333	0.9333
XGBoost	1.0000	0.9333	0.9333	0.9333	0.9333
CatBoost	1.0000	0.9333	0.9333	0.9333	0.9333
Gaussian Naive Bayes	0.9583	0.9667	0.9697	0.9667	0.9666
MLP Classifier	0.9833	0.9667	0.9697	0.9667	0.9666

difference indicates heavy overfitting and lower generalization than other classifiers.

D. Classical vs. Neural Models Comparison

Surprisingly, the simple probabilistic classifier Gaussian Naive Bayes was on par with the highly complex MLP neural network. This indicates that the data might not be extremely non-linear, and simpler models can actually have state-of-the-art performance.

E. Overall Best Models

Considering balanced performance in all measures, SVM, Gaussian Naive Bayes, and MLP are the best classifiers in this dataset. They have good predictive ability without too much overfitting.

V. CONCLUSION

This cross-algorithm comparison of various machine learning models on the Iris dataset illustrates the relative merits of diverse classification paradigms. Although all the models performed satisfactorily, ensemble-based strategies like Random Forest, XGBoost, and CatBoost delivered the best, most consistent results consistently, showcasing their capacity to minimize variance and avoid overfitting by aggregation and boosting. Support Vector Machines were also very effective, employing margin maximization to classify the classes with excellent generalization. Neural network methods, as embodied by the Multilayer Perceptron, performed competitively despite the limited data set size, showing how versatile deep learning is even in basic classification tasks. Simpler models such as Decision Trees and Gaussian Naive Bayes were good baselines that worked well but had limitations in overfitting or independence assumptions. The overall result is that although more sophisticated ensemble methods have the best trade-off between accuracy and robustness, simple models are still useful for interpretability and efficiency so that model choice is very context-dependent on the particular problem and needs at hand.

REFERENCES

[1] J. Karim *et al.*, “Detection of non-cardiac fetal abnormalities by ultrasound at 11–14 weeks: systematic review and meta-analysis,” *Ultrasound Obstet. Gynecol.*, 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,” arXiv:2501.02000, 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 Syst. Rev.*, 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,” Poster, FMF, 2024.

[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,” *Sci. Rep.*, vol. 15, Art. 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] U. Islam *et al.*, “Fetal-Net: Enhancing Maternal-Fetal Ultrasound Interpretation through Multi-Scale Convolutional Neural Networks and Transformers,” *Sci. Rep.*, vol. 15, Art. 25665, Jul. 2025.

[13] 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, Art. e077366, Feb. 2024.

[14] 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.

[15] S. Gupta *et al.*, “Prenatal Diagnostics Using Deep Learning: Dual Approach to Plane Localization and Cerebellum Segmentation,” *J. Med. Imaging Anal.*, vol. 7, no. 1, pp. 45–55, Feb. 2025.

[16] X. Zhang *et al.*, “Advancing Prenatal Healthcare by Explainable AI Enhanced Fetal Ultrasound Image Segmentation Using U-Net++ with Attention Mechanisms,” *Sci. Rep.*, vol. 15, Art. 30012, Jun. 2025.