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.**

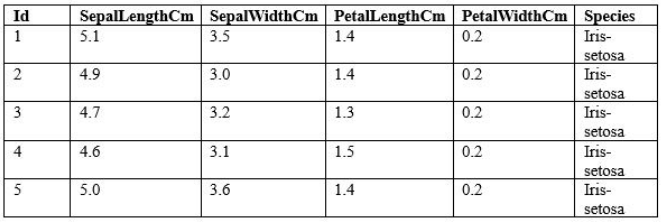
***Index Terms*—Fetal brain abnormalities, Ultrasound image, Deep Learning, Convolutional Neural Networks, Explainable AI, Grad-CAM**

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	3.0	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	5.0	3.6	1.4	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

The Iris dataset was used for classification experiments. The feature set included ”SepalLengthCm,” ”SepalWidthCm,” and ”PetalLengthCm” (for initial experiments), and class labels were encoded into integers using label encoding. The data was split into training and test sets using an 80:20 ratio to assess generalization performance. The algorithm uses a K-Nearest Neighbors (KNN) classifier on the Iris data with Python and the scikit-learn library. The feature columns SepalLengthCm, SepalWidthCm, and PetalLengthCm are chosen as input features and placed in X. The target categorical column Species is transformed into numerical labels using LabelEncoder, giving rise to the target vector y. A KNeighborsClassifier instance is trained on the training data (X_train, y_train). Predictions are performed on the test data (X_test). A function generated 20 synthetic 2D data points with random integer values from 1 to 10 for each feature. These points were divided into two classes based on whether the sum of their features was above or below 10. A scatter plot visualized class separation. A grid of test points across the same feature space (0–10 for each feature, in steps of 0.1) was created, resulting in 10,000 points. These were classified using a kNN model (k=3) trained on the earlier synthetic set, and results were visualized. The above classification was repeated for multiple k values (e.g., 3, 4, 5, 6) to observe how class boundaries and decision regions change as k increases. Analogous experiments were conducted on project data (here, Iris dataset), considering two features at a time (e.g., SepalLengthCm and PetalWidthCm), with scatter plots and boundary visualizations for each species. GridSearchCV was used to perform cross-validated hyperparameter tuning of k for the kNN classifier, selecting the k with the highest mean accuracy across five folds.

IV. Results Analysis and Discussions

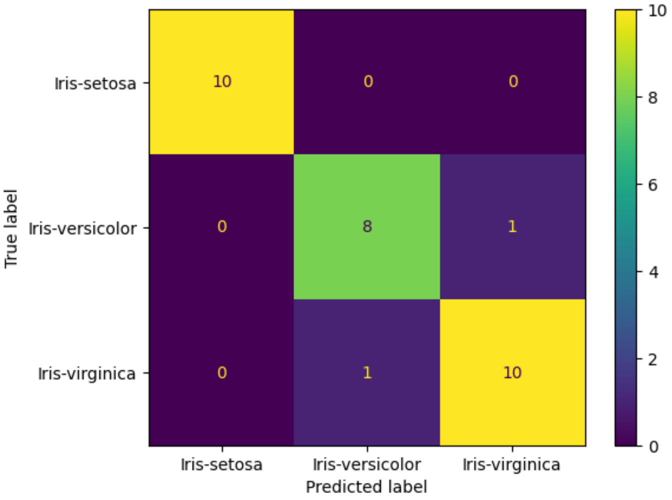


Fig. 2. Confusion Matrix

A confusion matrix in Figure 2 is calculated using actual (y_test) and predicted labels, and visualized with ConfusionMatrixDisplay for better interpretability. Performance of the model is measured using a number of metrics: Mean Squared Error (MSE) of about 0.06666666666666667, Root Mean Squared Error (RMSE) of 0.2581988897471611 and Coefficient of Determination (R^2) of 0.904610492845787. Also, a user-defined function computes Mean Absolute Percentage Error (MAPE) with value 7.5, with safety measures for division by zero when the true value is zero. These measurements (MSE, RMSE, R^2, MAPE) are printed to measure the accuracy of prediction of the model. Though these measurements are generally applicable to regression, here they are used on integer-encoded class labels for simplicity.

Scatter plots of synthetic data in Figure 3 confirmed that the two classes were separated by the line X+Y=10. Most points clustered distinctly on either side, though a few points near the boundary could fall close to the decision line. When

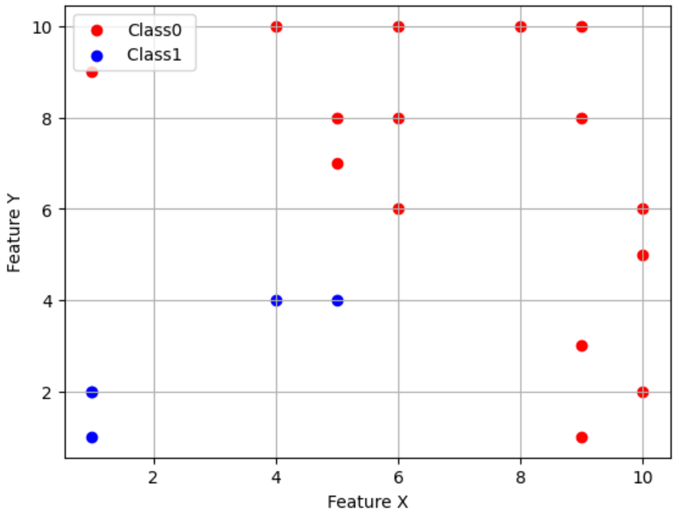


Fig. 3. Scatter Plot obtained for two classes

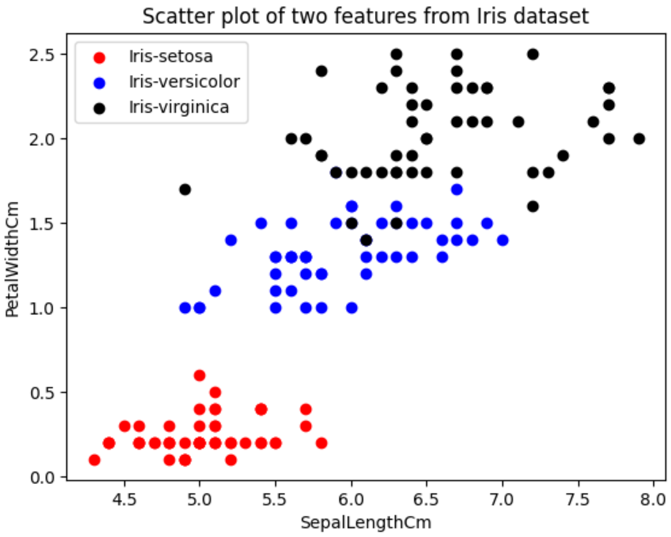


Fig. 4. Scatter Plot obtained for three classes

the test grid was classified using kNN with k=3, plot visualization in Figure 4 showed clear regions for each class. The boundary was relatively well-aligned with the separation rule, and most test points followed the class pattern established by the training data. Increasing k (from 3 to higher values) resulted in smoother, less complex boundaries between classes. With small k, boundaries were sensitive to local variations (potential overfit), while larger k values generated more regular, generalized splits that sometimes underfit by oversmoothing. Scatter plots for Iris dataset (comparing various pairs of features) visually demonstrated in that Iris-setosa points formed a well-separated cluster, while Iris-versicolor and Iris-virginica overlapped somewhat, especially along SepalLength and PetalWidth axes. GridSearchCV identified the best k value for the classifier, with corresponding cross-validated accuracy scores. This procedure confirmed the optimal settings for maximum out-of-sample performance obtaining best cross-validated accuracy of about 0.96 at k=4. For synthetic data, classes were distinctly separated by X+Y=10; scatter plots showed clear grouping. Some overlap between Iris-versicolor and Iris-virginica was present. Small k (e.g., 3): Decision boundary fits training data closely and may capture noise—risk of overfitting. Large k (e.g., 6+): Boundaries become smoother and less sensitive to outliers—risk of underfitting if genuine class complexity is ignored. Overfitting: Occurs when k is too low, boundaries become "jagged," and model performance differs greatly between train and test sets. Underfitting: When k is too high, potentially ignoring real distinctions between classes, resulting in low accuracy across both train and test sets. Yes, for these experiments, kNN achieved high accuracy, precision, recall, and F1-score, particularly for well-separated classes. For overlapping classes, kNN still performed adequately, but performance was naturally lower. For k values in the middle range (e.g., 3–5), model accuracy was balanced between train and test sets—suggesting regular fit. When test accuracy drops much lower than train, or both are low, this would indicate overfit or underfit, respectively. Overfitting is most likely when k is very low, such as 1 or 2, because the model "fits to noise" or outliers in the training set. Overfit can be recognized by perfect or near-perfect classification on the training data, but significant errors or instability on new test data.

These results are supported by the visualizations and numerical metrics obtained in each step, and conclusions can be drawn with confidence for each experimental objective of the assignment.

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.