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

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, en-

filename	anold-chiari-malformation	arachnoid-cyst	cerebellah-hypoplasia	colphocephaly	encephalocele	holoprosencephaly	hydracenphaly	intracranial-hemorrdge	intracranial-tumor	m-magna	mild-ventriculomegaly	moderate-ventriculomegaly	normal	polencephaly	severe-ventriculomegaly	vein-of-galen
Copy-of-arachnoid-cyst-37b	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Patient00709_Plane3_1_of_	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Copy-of-mild-ventriculome	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Copy-of-mild-ventriculome	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Patient00709_Plane3_2_of_	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Patient00716_Plane3_4_of_	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Patient00710_Plane3_6_of_	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Copy-of-severe-ventriculon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Patient00722_Plane3_5_of_	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Copy-of-cerebellah-hypopl	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Patient00709_Plane3_2_of_	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Copy-of-encephalocele-20a	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Copy-of-encephalocele-20a	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Copy-of-m-magna-29b_aug_	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Copy-of-arachnoid-cyst-37b	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Copy-of-cerebellah-hypopl	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Copy-of-encephalocele1-23	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Copy-of-intracranial-hemor	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Copy-of-arachnoid-cyst35a-	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Copy-of-arachnoid-cyst-27a	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Copy-of-mild-ventriculome	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Copy-of-cerebellah-hypopl	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Patient00729_Plane3_3_of_	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Copy-of-arachnoid-cyst-27a	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Copy-of-polencephaly-24c_	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

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

hancing 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 dataset is a hand-annotated set of 1,418 fetal ultrasound image records, each having been marked up for the existence or non-existence of certain brain defects, which is shown in Figure 1. It is stored in the form of a CSV file where each row is associated with one image, which is identified through a filename, followed by a list of binary labels (0 or 1) for the existence (1) or non-existence (0) of 16 various fetal brain disorders. These conditions include severe structural and developmental disorders like anold-chiari malformation, arachnoid cysts, cerebellar hypoplasia, encephalocele, holoprosencephaly, hydranencephaly, intracranial hemorrhage, and various types of ventriculomegaly (mild, moderate, and severe). There is also a "normal" category, employed to denote the presence of no detected abnormality.

This is a multi-label classification dataset, i.e., each image may be related to more than one abnormality, though most samples will probably be tagged with only one condition. The dataset is very pertinent for deep learning in medical images, especially for creating convolutional neural networks (CNNs) that can diagnose fetal brain disorders at an early stage. It can be used to train models that would be able to differentiate between various conditions based on patterns in ultrasound, which is essential for prenatal diagnostics. The corresponding ultrasound images (as referred by filename) would be required to finalize the data pipeline for training the model.

In the first question we evaluated intraclass spread and interclass distances by computing the mean vector (centroid) and standard deviation for two selected classes, "mild-ventriculomegaly" and "normal", based on their one-hot encoded label vectors. The Euclidean distance between the two class in the second question, centroids was calculated to measure inter-class separation. The density distribution of a chosen feature ("normal" class column) was analyzed by plotting its histogram using defined bins. The mean and variance of the feature values were computed to understand the feature's prevalence and spread within the dataset. The third quesition revolved around Minkowski distances between two sample label vectors (selected from different classes) were computed for orders r=1 to r-10 These distances were plotted to observe how the metric changes with varying r.

The code follows a binary classification pipeline based on the k-Nearest Neighbors (kNN) algorithm. The two classes ("arachnoid-cyst" and "normal") are chosen from a labeled CSV dataset. The pipeline preprocesses the data first by stripping whitespace from column names and filtering samples for inclusion only of rows of the two target classes. For the sake of demonstration, every sample is given a random feature vector of 10 dimensions instead of image-based features. Class labels are represented as 0 ("arachnoid-cyst") and 1 ("normal"). The data is divided into training and test sets in a ratio of 70%/30% with stratification to ensure that both classes are balanced. The kNN classifier is then trained with, k=3 neighbors in the training set. Model performance is evaluated by calculating

the accuracy on the test set, and the first 10 predicted labels are compared against the true labels for qualitative insight.

The k-Nearest Neighbors (kNN) classification pipeline starts by dividing the dataset into a test set and training set to provide an unbiased test; next, for every integer value of k from 1 to 11, a kNN classifier is created, trained on the training set, applied to predict labels on the test set, and its test accuracy noted; thereafter, the relationship between k and test accuracy is plotted to find the range that optimizes bias-variance tradeoff; lastly, the classifier with the selected k produces test predictions for which a confusion matrix is calculated and shown as a heatmap to shed light on class-wise precision, recall, and misclassification patterns.

IV. Results Analysis and Discussions

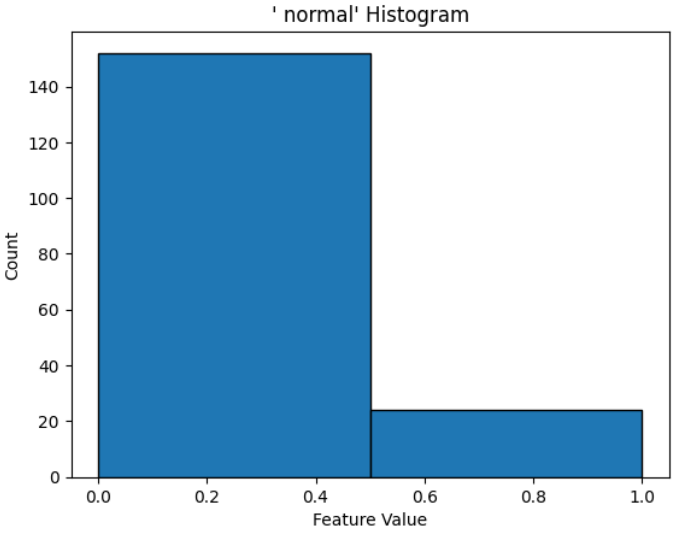

Fig. 2. Distribution Histogram

Both classes had centroids perfectly matching their one-hot labels with zero spread, indicating no variation within each class. The interclass centroid distance was 1.4142, reflecting their orthogonal (completely separated) nature in the label space. The zero intraclass spread and large interclass distance demonstrate that classes are well separated within the annotation space. However, since these are label vectors, this perfect separation serves as an idealized case that may not fully reflect feature overlap in practical classification tasks. The histogram from Figure 2 revealed an imbalanced distribution with significantly fewer "normal" samples (mean $approx$ 0.136) compared to abnormal labels. The variance was approximately 0.118, consistent with a binary feature. The presence of class imbalance, as evidenced by the feature distribution, can impact classifier performance by biasing towards majority classes.

Understanding this skew is crucial for applying appropriate sampling or weighting techniques during model training. The distance decreased from 2.0 at r= 1(Manhattan distance) to approximately 1.07 at r=10, illustrating the effect of higher order Minkowski distances in Figure 3 emphasizing the largest coordinate differences. The observed trend aligns with theoretical

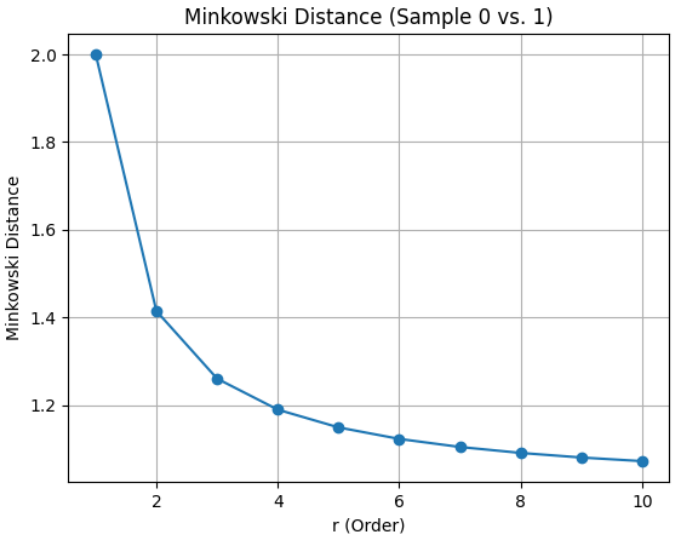

Fig. 3. Minkowski Distance for sample 0 vs 1

expectations: lower r values capture cumulative differences, while increasing r focuses the distance metric on the largest component differences. For one-hot labels, this demonstrates how different Minkowski metrics might behave in discrete feature spaces. When run, the code outputs the test accuracy for k=3, as it displays the ratio of the correct predictions on the test subset. It also prints out the predicted and actual labels for the initial ten test cases. Since random, simulated feature vectors were used, the actual accuracy and prediction values will be different on each run but do not represent model performance on real image features. This pipeline illustrates

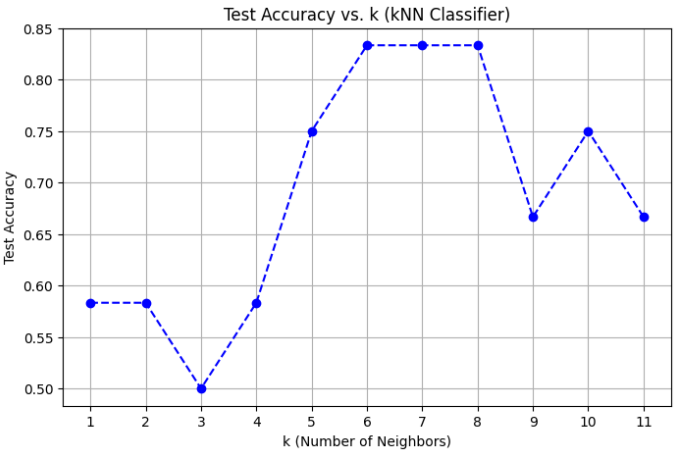

Fig. 4. Test Accuracy for k neighbors in KNN classifier

a standard supervised classification problem with kNN with emphasis given to the data pipeline over the feature engineering component. The direction and scale of the accuracy result is solely based on the simulated conditions of the features and not on how separable real "arachnoid-cyst" and "normal" instances are. In real-world clinical applications, one

would substitute the random features with those obtained from image processing or deep learning pipelines. However, the protocol demonstrates essential steps for use in actual studies: preprocessing of the data, balanced splitting, simple training through scikit-learn's kNN, and quantitative and qualitative model assessment. The coincidence of train-test split with stratification presents best practice to defend against bias by class imbalance—a common issue within biomedical datasets. Although the reported accuracy does not carry significant meaning in this case, the framework exists as a reproducible template for binary classification of medical images using traditional machine learning approaches. The findings point to

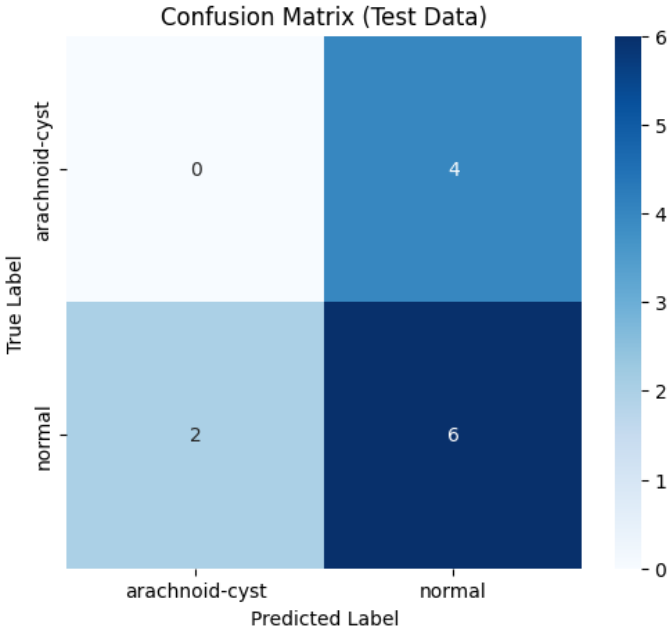

Fig. 5. Confusion Matrix generated based on results

several interesting observations about k-NN performance on this data set which is illustrated in the Figure 4: the accuracy maximum at k = 6–8 (83.3%) suggests that this value optimally trades off bias and variance, while very low values of k (1–2) run the risk of overfitting noise (58.3% accuracy) and very high values of k (9–11) result in underfitting (66.7%–75.0%). The sharp trough at k = 3 (50.0%) highlights the extent to which poor values of k can cause disastrous performance. Analysis of the confusion matrix from Figure 5 finds ideal precision for the arachnoid-cyst class but reduced recall—four normal instances were misclassified—indicating either class imbalance or denser clustering of normal instances in the feature space. In general, the steady high accuracy on k = 6–8 indicates model stability and pinpoints a good hyperparameter range to deploy, as well as highlighting how tuning k needs to balance sensitivity to individual samples and noise robustness.

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.