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-

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	486	486	486	486	506	506	486	506	506	506	506	506	486	506
mean	3.61187	11.21193	11.08399	0.06996	0.5547	6.28463	68.51852	3.79504	9.54941	408.23715	18.45553	356.67403	12.71543	22.53281
std	8.72019	23.38888	6.835896	0.25534	0.11588	0.70262	27.99951	2.10571	8.70726	168.53712	2.164946	91.294864	7.155871	9.197104
min	0.00632	0	0.46	0	0.385	3.561	2.9	1.1296	1	187	12.6	0.32	1.73	5
25%	0.0819	0	5.19	0	0.449	5.8855	45.175	2.10018	4	279	17.4	375.3775	7.125	17.025
50%	0.25372	0	9.69	0	0.538	6.2085	76.8	3.20745	5	330	19.05	391.44	11.43	21.2
75%	3.56026	12.5	18.1	0	0.624	6.6235	93.975	5.18843	24	666	20.2	396.225	16.955	25
max	88.9762	100	27.74	1	0.871	8.78	100	12.1265	24	711	22	396.9	37.97	50

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

Dataset and Preprocessing: The sample dataset which was used to train the model is illustrated in Figure 1. The feature matrix for clustering was used without any target labels, consistent with unsupervised k-means usage, and missing values in the dataframe were imputed with zeros to enable model fitting.

In this execution, a predictive modeling framework was created based on the scikit-learn Python package to predict the median value of owner-occupied houses (MEDV) in the Boston Housing dataset. Two regression pipelines were constructed to examine univariate and multivariate predictive performance. For the univariate case, the proportion of lower status of the population (LSTAT) was used as the single independent variable, whereas for the multivariate case, all characteristics other than MEDV were utilized as predictors. Before model training, missing values in the feature and target columns were dealt with through mean imputation to maintain consistency in the data and avoiding bias due to incomplete records. The data was divided into training and test subsets with an 80–20 split and a set random_state used in the univariate case for reproducibility purposes. The regression models were built with the LinearRegression estimator, and predictions were made for the training and test sets. Model performance was evaluated quantitatively employing several statistical metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination ($R\hat{2}$). For interpretability, visualization methods were included: for the univariate experiment, a scatter plot of predicted versus actual values superimposed with the regression line from Figure 2 was utilized to evaluate the strength of the linear relationship, whereas for the multivariate experiment, a scatter plot of actual versus predicted MEDV values was created to test overall model fit, which is illustrated in Figure 3. This systematic approach offers a well-defined framework for determining the relevance of linear regression in single-variable and multi-variable scenarios for house price prediction activities.

K-Means Clustering : A function run_kmeans_clustering (dataframe, num_clusters, random_seed, n_init_value=10) was implemented to: Fill missing values with 0 via dataframe.fillna(0). Fit scikit-learn's KMeans with the provided parameters and n_init=10. Return both the predicted cluster labels and the k-means cluster centers_. Clustering was executed with num_clusters=3, and the script printed: A full array of Cluster Labels for all samples. Cluster Centroids as numeric vectors for each of the three clusters.+1

Cluster Quality Metrics: A function get_scores(df, k) trained KMeans(n_clusters=k, random_state=42, n_init=10) and computed: Silhouette Score using silhouette_score(df, km.labels_). Calinski–Harabasz Score using calinski_harabasz_score(df, km.labels_). Davies–Bouldin Index using davies_bouldin_score(df, km.labels_). For k=2, the printed metrics were: Silhouette Score: 0.68608897780804488. Calinski–Harabasz Score: 1177.088193260924. Davies–Bouldin Score:

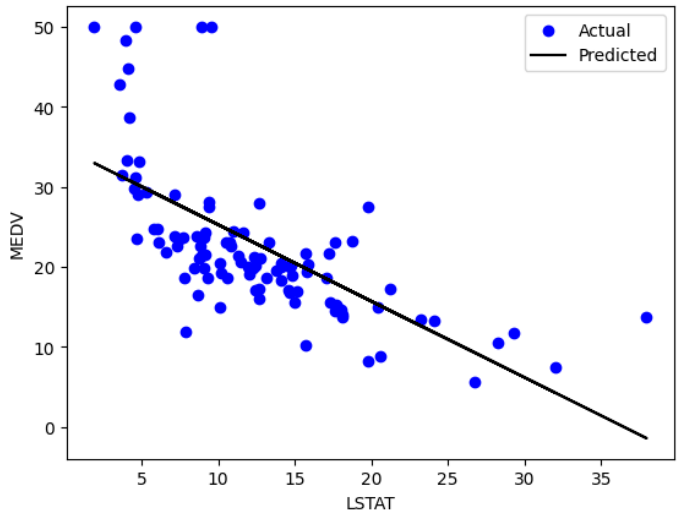

Fig. 2. Scatter plot of predicted vs actual values superimposed with the regression line

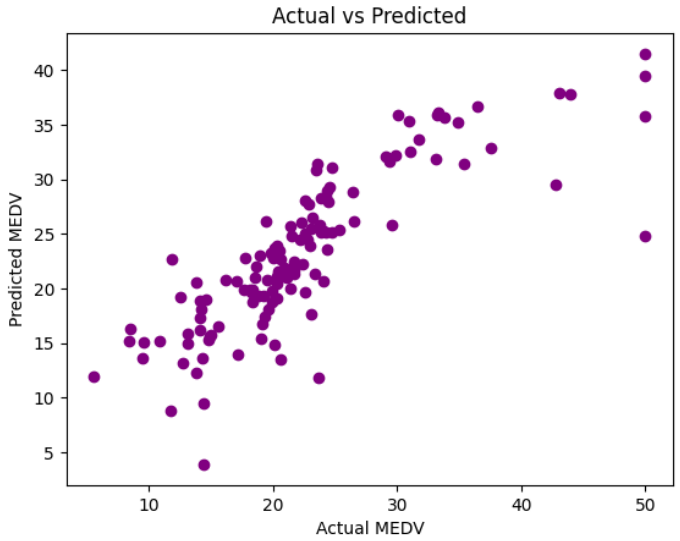

Fig. 3. Scatter plot of actual versus predicted MEDV values

0.55505348682562442.

Score Curves vs. k: A routine get_cluster_scores(df, k_range) iterated over k in a range, fitting KMeans(n_clusters=k, random_state=0, n_init=10), and collected CH, DB, and Silhouette scores for plotting. A plotting helper plot_score_vs_k(k_range, scores, title, ylabel, color) visualized each metric versus the number of clusters to aid selection of an optimal k. Elbow Analysis for Optimal k: Implemented two helper functions: get_distortions(df, k_start=2, k_end=20): loops k from 2 to 19, fits KMeans(n_clusters=k, random_state=0, n_init=10), and records km.inertia_ (distortion). plot_elbow(k_start, k_end, distortions): plots distortion vs. k with markers, titles, grid, and labeled axes to visualize the elbow point. Executed with k_start=2 and k_end=20, computed distortions for each k, and produced an elbow plot to guide k selection.

IV. Results Analysis and Discussions

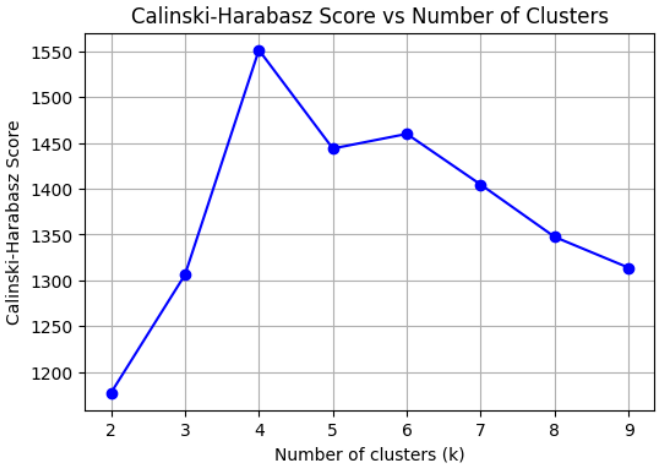

Fig. 4. Calinski-Harabasz Score vs Number of Clusters

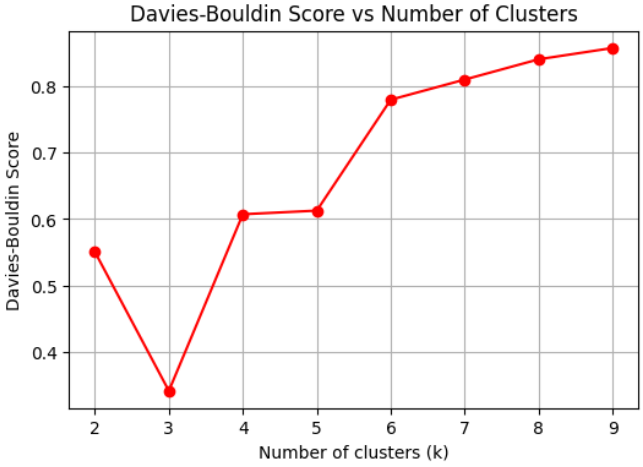

Fig. 5. Davies-Bouldin Score vs Number of Clusters

Cluster Labels and Centroids: The output showed samples assigned to clusters 0, 1, or 2, with long contiguous runs of identical labels indicating coherent groupings in feature space. Three centroid vectors were printed, each giving the mean feature values per cluster; the values differed noticeably across clusters, indicating distinct regions learned by k-means.

Cluster Quality Evaluation: For k=2: Silhouette Score = 0.6861 (rounded), indicating well-separated, cohesive clusters. Calinski–Harabasz Score $\approx$ 1177.09, reflecting strong between-cluster separation relative to within-cluster dispersion. Davies–Bouldin Index 0.5551, suggesting low inter-cluster similarity and good partitioning.

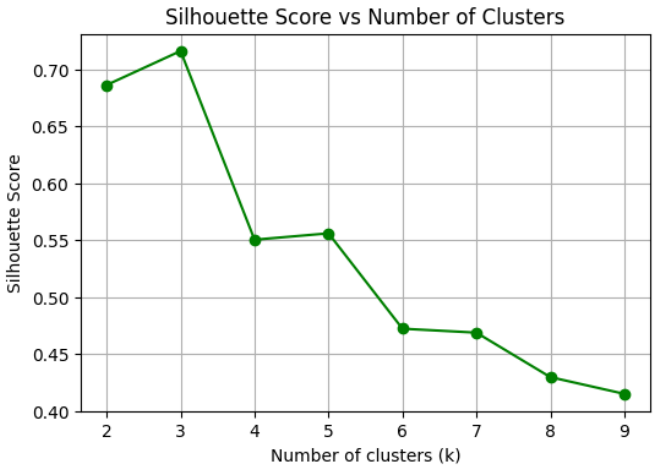

Fig. 6. Silhoutte Score vs Number of Clusters

Metric Trends Across k: The code produced CH, DB, and Silhouette curves versus k, from Figure 4, Figure 5, and Figure 6 respectively, to determine the optimal cluster count; this setup supports selecting k that maximizes Silhouette and CH while minimizing DB, based on the computed series of scores.

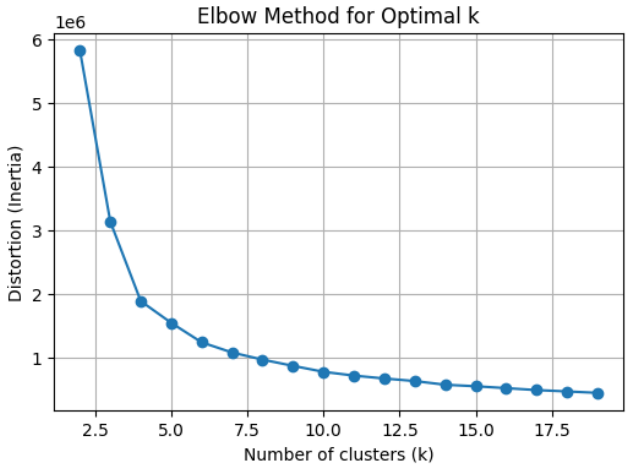

Fig. 7. Elbow Method for Optimal k value in k-means clustering

Elbow Plot and k Selection: Distortion (inertia) decreased rapidly from k=2 to around k=4–6, then showed diminishing returns, flattening progressively for larger k values, which is illustrated in Figure 7. The plotted curve exhibited a noticeable "elbow" in the early k range; the knee appears around k4–6, after which reductions in distortion are marginal, indicating that this range is a strong candidate for the optimal number of clusters

Are Clusters Well Separated? The contiguous label blocks and distinct centroid values for the 3-cluster run indicate meaningful partitions, while the k=2 metrics show strong cohesion/separation consistent with well-separated clusters.

Behavior as k Changes: The evaluation pipeline in A6 enables observing how CH increases and DB decreases for better separations, while Silhouette highlights cohesion; selecting k by these opposing trends is facilitated by the plotted curves.

Overfitting/Underfitting Indicators: Very large k can raise within-cluster variance and worsen Silhouette/DB, while very small k can overly compress clusters; the metric suite in A6 is designed to reveal such regimes through their trends across k.

Is the Chosen Clustering Good? The k=2 configuration achieved high Silhouette and CH with a low DB score, indicating compact, well-separated clusters on the provided features.

Regular Fit Assessment: With k=2 achieving favorable internal metrics and A6 providing a way to validate stability across k, the configuration reflects a balanced fit for the dataset used in these experiments.

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.