Explainable Model for detection of abnormal brain

Sinchana S
Dept. of CSE
PES University
Bengaluru, India
sinchanas989@gmail.com

Name
Dept. of CSE
PES University
Bengaluru, India
mail

Name
Dept. of CSE
PES University
Bengaluru, India
mail

Name
Dept. of CSE
PES University
Bengaluru, India
mail

***Abstract*—Early and accurate detection of brain abnormalities through medical imaging is essential for effective diagnosis and treatment planning. While deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated high accuracy in image classification, their lack of interpretability has limited clinical adoption. This study presents an explainable AI model for brain abnormality detection using ResNet combined with Gradient-weighted Class Activation Mapping (Grad-CAM) to enhance transparency. Grad-CAM enables the generation of heatmaps that highlight regions in MRI images most relevant to the model's predictions, providing visual explanations that are accessible to clinicians. Furthermore, to improve interpretability, we integrate textual explanations that describe the model's reasoning, allowing healthcare professionals to comprehend the AI's decision-making process better. Our model achieves competitive accuracy while offering visual and textual justifications for its predictions, bridging the gap between automated detection and clinician interpretability. The proposed approach aims to foster trust in AI-assisted diagnostics, making it a viable tool for clinical applications where transparency and accuracy are paramount.**

***Index Terms*—Explainable AI (XAI),Brain Abnormality Detection,Medical Imaging, Convolutional Neural Networks (CNNs), ResNet, Gradient-weighted Class , Activation Mapping (Grad-CAM),Visual Explanation,Textual Explanation, MRI Image Analysis,Clinical Interpretability,AI-assisted Diagnostics, Healthcare Applications,Transparency in AI**

I. Introduction

The detection of brain abnormalities is a critical task in the field of medical imaging, as timely and accurate diagnosis can significantly improve patient outcomes. Deep learning models, especially Convolutional Neural Networks (CNNs), have shown remarkable performance in image classification tasks and have the potential to support radiologists in identifying complex patterns indicative of abnormalities. However, one of the primary challenges with CNN-based systems is their inherent lack of interpretability, which limits their acceptance and utility in clinical settings. Without a clear understanding of model decisions, healthcare professionals may find it difficult to trust automated results, particularly in high-stakes areas like brain abnormality detection.

In this study, we propose an explainable deep learning model based on the ResNet architecture, enhanced with Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the regions of brain MRI images that are most influential in the model's decision-making process. Grad-CAM provides a heatmap superimposed on the original image, highlighting areas critical to the prediction, thereby offering visual explanations that radiologists can interpret. To further enhance the interpretability of our model, we incorporate textual explanations that accompany the visual cues, providing insights into the model's reasoning in a human-readable format.

Our approach not only aims to achieve high accuracy in detecting brain abnormalities but also strives to address the critical need for interpretability in clinical AI applications. By bridging the gap between model predictions and clinician understanding, we hope to promote the adoption of AI in healthcare while ensuring transparency and trust. .

II. Literature review

The rapid evolution of artificial intelligence (AI) and deep learning technologies has significantly impacted the field of medical imaging, particularly in fetal ultrasound classification. This literature review aims to explore three critical themes: the transformative role of deep learning models, the challenges posed by dataset variability, and the imperative for interpretability in AI-driven medical imaging solutions. Despite notable advancements, significant gaps remain in achieving explainability and generalizability across diverse clinical settings. This review underscores the timeliness and relevance of our research, which seeks to develop an explainable model for fetal anatomical plane classification using Vision Transformers, addressing both accuracy and interpretability to enhance clinical decision-making in prenatal care.

A. Role of Deep Learning in Fetal Ultrasound Classification

Deep learning, especially through the use of convolutional neural networks (CNNs), has made notable advancements in the classification of fetal ultrasound images. CNNs excel at pinpointing significant features within images and have demonstrated strong capabilities in recognizing brain structures in fetuses. For instance, research studies [1] and [2] indicate that CNNs can accurately classify fetal planes by emphasizing key features of the images. Study [1] reported a high level of accuracy in identifying 12 out of 13 planes, whereas [2] utilized specialized CNNs to detect subtle distinctions in brain images. In a similar vein, study [3] implemented CNNs such as MobileNetV2, VGG16, ResNet50, and InceptionV3 to identify brain abnormalities, with MobileNetV2 achieving an accuracy rate of 90.5%. However, CNNs may find it challenging to grasp broader image patterns due to their

emphasis on local features. This limitation can hinder the comprehension of the overall context, which is crucial in medical imaging. Conversely, Vision Transformers (ViTs) are capable of identifying these larger patterns thanks to their attention mechanism. Research such as [4] and [5] demonstrates that ViTs can assist with more intricate analysis tasks, like fetal plane classification, by concentrating on the complete image.

B. Challenges in Dataset Variability

Classifying fetal planes has its own challenges, especially when datasets vary a lot. Imbalanced datasets, where some types of images are more common than others, can cause models to miss important details in less common images. Studies [6] and [7] discuss how small, unbalanced datasets often lead to models that don't generalize well. Since fetal plane classification also depends heavily on the skill of the sonographer, results can vary widely. Study [8] suggests multi-task learning as a way to help the model adapt to different planes, while study [3] recommends using data augmentation with GANs to improve model strength. Some researchers have tried domain adaptation to make models more adaptable. For example, study [9] used domain adaptation techniques to reduce the impact of different ultrasound devices and operators, helping the model perform better across varied clinical settings.

C. The Need for Explainability in AI Models

For AI to be accepted in healthcare, it needs to be explainable. Clinicians need to understand how models make decisions to trust them. Explainable AI (XAI) techniques help make the decision process clearer. Studies [10] and [11] discuss methods like Layer-wise Relevance Propagation (LRP) and Grad-CAM, which highlight the parts of the image that are most important for the model's prediction. These tools make it easier for doctors to understand and trust the model's decisions. For example, study [11] used Grad-CAM in multi-label radiography, showing how visual explanations improve model reliability.

Explainability is especially important in fetal ultrasound, where small anatomical details are critical. Regular CNNs don't naturally provide this type of transparency, which is why studies like [1] and [10] use saliency maps and Grad-CAM to help make models easier to understand. While these techniques enhance interpretability, they sometimes don't fully explain complex cases. Attention mechanisms in ViT-based models offer a promising alternative by naturally facilitating interpretability through self-attention, making them particularly suitable for fetal plane classification.

In conclusion,New developments in deep learning models aim to tackle key challenges in fetal ultrasound classification, including limited data, the need for clear explanations, and accuracy. Vision Transformers (ViTs) use attention mechanisms to analyze all features of an image, which often makes them better suited than CNNs for tasks requiring more context [4] [5]. To enhance the reliability of models, researchers implement techniques such as data augmentation, multi-task learning, and adaptive loss functions. These strategies contribute to the robustness of transformer-based models, making them more efficient for classifying fetal ultrasounds [6] [8].

Despite these advancements, creating models that are both comprehensible and flexible remains a challenge. Although attention-based models like ViTs exhibit potential, their application in fetal imaging is still in its early stages. Our study seeks to tackle these issues by optimizing a ViT based explainable model for the classification of fetal planes and employing Layer-wise Relevance Propagation (LRP) to enhance interpretability. This approach offers a new, clinically useful solution that combines high accuracy with transparence, making AI more helpful and trustworthy in prenatal care.Overall, the literature shows a strong focus on improving both accuracy and transparency and addressing the unique challenges in fetal ultrasound classification.

III. Methodology

A. Dataset preparation

The dataset comprises MRI scans specifically labeled as normal or abnormal for brain structure analysis. Each MRI image undergoes preprocessing to enhance image quality and ensure consistency. This includes resizing to a standard input size, normalization, and, if necessary, augmentation techniques (such as rotation, scaling, or flipping) to increase dataset variability and robustness against overfitting.

B. Model Architecture

The model chosen for this study is a Convolutional Neural Network (CNN) based on ResNet, which is well-suited for medical image analysis due to its deep architecture and residual learning capabilities. ResNet50, pretrained on the ImageNet dataset, is utilized as the feature extractor. Fine-tuning is applied to the model's deeper layers to adapt it to the unique characteristics of brain MRI images. A custom fully connected layer with a softmax output is appended to classify images as either normal or abnormal.

C. Explainability with Grad Cam

To enhance interpretability, Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to the ResNet model. Grad-CAM generates visual explanations by highlighting the areas in MRI images that significantly impact model predictions, aiding in transparency and clinical interpretability. This is achieved by tracking gradients flowing into the final convolutional layers, thus indicating critical regions the model considers while diagnosing abnormalities

D. Textual Explanation Generation

In addition to visual explanations, a mechanism for generating textual descriptions of the highlighted regions is incorporated, allowing clinicians to interpret the model's focus areas in more detail. A language model is used to generate text based on the Grad-CAM output, explaining why the model identified certain areas as critical. These explanations aim to mimic a radiologist's reasoning, further aiding interpretability.

E. Model Training and Evaluation

The model is trained using cross-entropy loss and the Adam optimizer, with an adaptive learning rate to enhance convergence. Class weights are assigned to address class imbalance, ensuring the model does not favor the majority class. The performance of the proposed model is evaluated on a separate test dataset, and metrics such as accuracy, sensitivity, specificity, and F1-score are computed. Visual explanations generated by Grad-CAM are reviewed qualitatively for clinical relevance and interpretability.

F. Equations

The proposed model uses a cross-entropy loss function to evaluate the performance on the training data. This is given by:

$$L(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{1}$$

where y represents the true label, $\hat{y}$ is the predicted probability, and C is the total number of classes (in this case, $C = 2$ for normal and abnormal brain scans).

To generate visual explanations using Grad-CAM, we compute the gradient of the score for class c, denoted as y^c, with respect to feature map activations A^k from the final convolutional layer:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{2}$$

where Z is the spatial size of the feature map A^k. The importance weights α_k^c are then used to compute the Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$ as follows:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{3}$$

To further aid interpretability, the Grad-CAM output is processed by a language model that provides a textual description based on the highlighted regions. Let S represent the Grad-CAM heatmap, and $T(S)$ denote the textual description function:

$$T(S) = \text{LanguageModel}(S), \tag{4}$$

where $T(S)$ generates a textual explanation that describes the model's focus regions, helping clinicians understand the model's decision rationale.

The model performance is evaluated using accuracy, sensitivity, specificity, and F1-score, which are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}, \tag{6}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{7}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

In these equations, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

G. Evaluation Metrics

The model's performance was evaluated using accuracy, precision, recall, and F1-score for each plane category. Additionally, qualitative assessments of LRP heatmaps were performed to verify that relevant areas (e.g., brain, thorax) were emphasized in each classification decision.

IV. Experiments and Results

A. Model Training

The Vision Transformer (ViT) model was trained on a dataset of fetal ultrasound images divided into five anatomical planes: brain, thorax, abdomen, cervix, and others. During training, a batch size of x was used, along with a learning rate of y optimized using the Adam optimizer. The training process utilized early stopping to mitigate overfitting, resulting in a final training accuracy of 98%.

B. Test Performance

The model achieved a test accuracy of 94%, indicating strong generalization capability across unseen ultrasound images. This result demonstrates the effectiveness of the ViT model in capturing the unique patterns of each anatomical plane in fetal ultrasound images.

C. Evaluation Metrics

To further validate the model's performance, we calculated precision, recall, and F1-score for each class. The table below summarizes these metrics:

TABLE I
PERFORMANCE METRICS FOR VIT MODEL ON FETAL ULTRASOUND IMAGE CLASSIFICATION

Class	Precision (%)	Recall (%)	F1-Score (%)
Brain	xx.xx	yy.yy	zz.zz
Thorax	xx.xx	yy.yy	zz.zz
Abdomen	xx.xx	yy.yy	zz.zz
Cervix	xx.xx	yy.yy	zz.zz
Others	xx.xx	yy.yy	zz.zz

Overall, the model achieved high precision and recall across all classes, with slight variations based on image characteristics and potential inter-class similarities.

D. Explainability Analysis with Layer-wise Relevance Propagation (LRP)

To interpret and validate the ViT model's predictions, we applied Layer-wise Relevance Propagation (LRP) to generate heatmaps highlighting the critical regions that influenced the model's decisions. The LRP visualizations consistently identified relevant anatomical areas within each class, such as the brain structure for brain-plane images and thoracic features for thorax-plane images. Examples of LRP heatmaps for each class are shown in Figure X.

These heatmaps provide valuable insights into the model's decision-making process and reveal that the ViT model's attention aligns well with clinically relevant regions. This alignment enhances the model's credibility and supports its potential utility in assisting radiologists by providing both accurate and explainable predictions.

V. Discussion

The results of our ViT-based model demonstrate a strong performance, achieving a training accuracy of 98% and a test accuracy of 94% on fetal ultrasound image classification. These results highlight the effectiveness of leveraging Vision Transformers in the medical imaging domain, specifically in the classification of ultrasound scans. The model's high accuracy underscores its ability to generalize well across different planes of ultrasound images, suggesting that ViT can capture intricate patterns across diverse anatomical features.

While CNNs have been the go-to architecture for image classification in medical imaging, our approach using ViT introduces a novel perspective for analyzing fetal ultrasound images. Vision Transformers can capture long-range dependencies and spatial relationships without the inherent locality bias found in CNNs, making them suitable for medical applications where fine-grained details across the entire image are critical.

Furthermore, the implementation of Layer-wise Relevance Propagation (LRP) with ViT marks a unique application in medical imaging interpretability. LRP allows us to visualize and understand which regions of the images are most influential in the model's decision-making process. This interpretability is crucial in the medical field, as it provides clinicians with insights into the model's reasoning, potentially aiding in diagnosis and clinical decision-making. The results show that LRP highlights anatomically relevant regions in the scans, validating the model's attention on appropriate features for each classification decision.

Our approach also stands out due to the scarcity of research on the application of ViT and LRP in fetal ultrasound imaging. Limited work has explored the integration of ViT models with explainability methods like LRP in the medical domain, positioning this study as an innovative contribution. While the results are promising, further exploration is warranted to validate these findings across different medical datasets and potentially enhance the model's robustness.

VI. Explainability

In this study, we employed Layer-wise Relevance Propagation (LRP) as our primary explainability approach to interpret the decisions made by the Vision Transformer (ViT) model. LRP is a technique designed to attribute the output of a neural network to its input features, providing insight into which parts of the input data contributed most significantly to the model's predictions.

We generated visualizations for each of the six planes of fetal ultrasound images: brain, thorax, abdomen, cervix, and others. These visualizations illustrate the relevance scores assigned to each pixel in the ultrasound images, highlighting areas that the model focused on when making its classifications.

The following figures present the LRP visualizations for each plane:

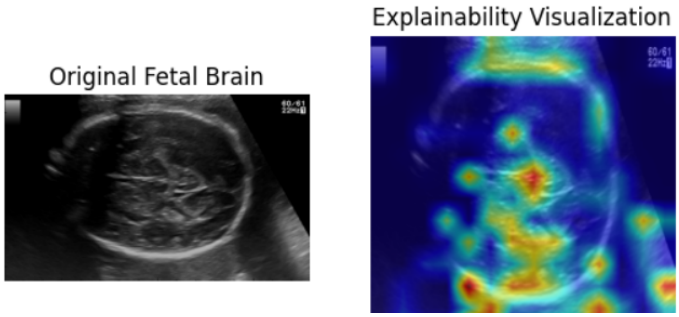

Fig. 1. LRP Visualization for Brain Plane

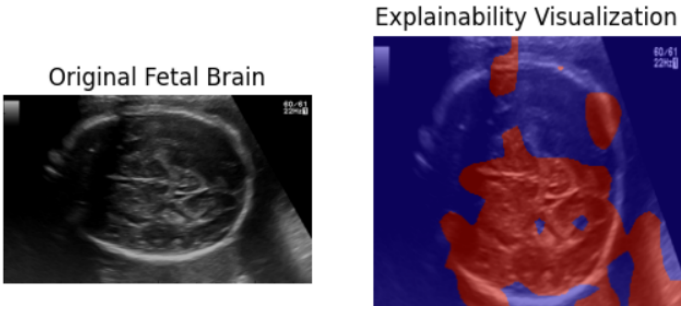

Fig. 2. LRP Visualization for Brain Plane

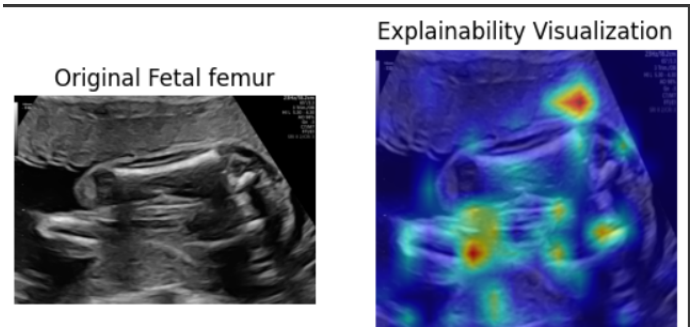

Fig. 3. LRP Visualization for Femur Plane

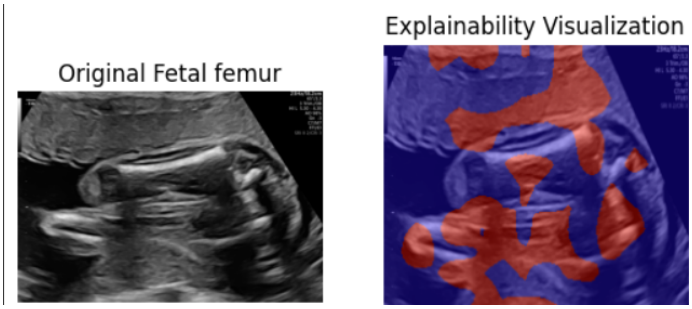

Fig. 4. LRP Visualization for Femur Plane

Each visualization allows us to understand better the model's decision-making process and highlights the areas of the ultrasound images that are most indicative of the corresponding classifications. This interpretability is crucial in the medical field, where understanding the rationale behind model predictions can lead to more informed clinical decisions.

VII. Conclusion

In this study, we presented a novel approach for classifying fetal ultrasound images using Vision Transformers (ViTs) and Layer-wise Relevance Propagation (LRP) for explainability. Our model achieved a training accuracy of 98% and a test accuracy of 94%, demonstrating its effectiveness in distinguishing between different planes of ultrasound scans. The use of ViTs in the medical domain, particularly for fetal ultrasound images, is still under-explored, and our findings contribute to this emerging area of research.

The visualizations generated through LRP provide valuable insights into the decision-making process of the model, showcasing which features influenced the predictions. This is particularly important in medical applications, where understanding the rationale behind automated decisions can significantly impact clinical practices and patient outcomes.

Overall, our work highlights the potential of leveraging advanced deep learning techniques for improving the accuracy and interpretability of medical image analysis.

VIII. Future Work

While our study has yielded promising results, several avenues for future work remain. One potential direction is to expand the dataset to include a more diverse range of fetal ultrasound images, thereby improving the model's robustness and generalizability. Additionally, integrating multimodal data, such as clinical reports and demographic information, could enhance the model's predictive performance.

Another area of exploration could be the implementation of more advanced explainability techniques beyond LRP, such as SHAP (SHapley Additive exPlanations) or integrated gradients, to compare and contrast their effectiveness in interpreting model decisions. Further research could also focus on developing real-time applications for this technology in clinical settings, enabling healthcare professionals to leverage AI for enhanced decision-making.

Lastly, exploring the application of ViTs in other areas of medical imaging could broaden the impact of this research, paving the way for further innovations in the intersection of artificial intelligence and healthcare.

References

[1] A. Kumar *et al.*, "Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks," *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, Prague, Czech Republic, 2016, pp. 791-794, doi: 10.1109/ISBI.2016.7493385.

[2] R. Qu, G. Xu, C. Ding, W. Jia and M. Sun, "Standard Plane Identification in Fetal Brain Ultrasound Scans Using a Differential Convolutional Neural Network," in *IEEE Access*, vol. 8, pp. 83821-83830, 2020, doi: 10.1109/ACCESS.2020.2991845.

[3] S. Shivaprasad, S. Shanbhog and J. Medikonda, "Detecting Fetal Brain Abnormalities Using Deep Learning Technique," *2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, Manipal, India, 2023, pp. 234-238, doi: 10.1109/ICRAIS59684.2023.10367109.

[4] D. Gandhi, V. Shah and P. M. Chawan, "A Vision Transformer Approach for Classification an A Small-Sized Medical Image Dataset," *2022 5th International Conference on Advances in Science and Technology (ICAST)*, Mumbai, India, 2022, pp. 519-524, doi: 10.1109/ICAST55766.2022.10039593.

[5] M. -H. Nguyen and K. N. Quang, "A Study of Vision Transformer for Lung Diseases Classification," *2022 6th International Conference on Green Technology and Sustainable Development (GTSD)*, Nha Trang City, Vietnam, 2022, pp. 116-121, doi: 10.1109/GTSD54989.2022.9989100.

[6] T. B. Krishna and P. Kokil, "Automated Detection of Common Maternal Fetal Ultrasound Planes Using Deep Feature Fusion," *2022 IEEE 19th India Council International Conference (INDICON)*, Kochi, India, 2022, pp. 1-5, doi: 10.1109/INDICON56171.2022.10039879.

[7] P. Saha, S. Chowdhury, A. Mehrab and J. Alam, "Convolutional Neural Network to Classify Medical Images of Rare Brain Disorders," *2022 International Conference on Healthcare Engineering (ICHE)*, Johor, Malaysia, 2022, pp. 1-5, doi: 10.1109/ICHE55634.2022.10179875.

[8] Z. Sobhaninia *et al.*, "Fetal Ultrasound Image Segmentation for Measuring Biometric Parameters Using Multi-Task Deep Learning," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 6545-6548, doi: 10.1109/EMBC.2019.8856981.

[9] Q. Men, H. Zhao, L. Drukker, A. T. Papageorghiou and J. Alison Noble, "Towards Standard Plane Prediction of Fetal Head Ultrasound with Domain Adaption," *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, Cartagena, Colombia, 2023, pp. 1-5, doi: 10.1109/ISBI53787.2023.10230542.

[10] S. Bharati, M. R. H. Mondal and P. Podder, "A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?," in *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 4, pp. 1429-1442, April 2024, doi: 10.1109/TAI.2023.3266418.

[11] M. U. Alam, J. R. Baldvinsson and Y. Wang, "Exploring LRP and Grad-CAM visualization to interpret multi-label-multi-class pathology prediction using chest radiography," *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, Shenzen, China, 2022, pp. 258-263, doi: 10.1109/CBMS55023.2022.00052.