Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Enhancing Non-Invasive Diagnosis of Endometriosis Through Explainable Artificial Intelligence: A Grad-CAM Approach

Afolashade Oluwakemi Kuyoro, Oluwayemisi Boye Fatade*, Ernest Enyinnaya Onuiri

Department of Computer Science, Babcock University Ilishan Remo, 121103 Ilishan-Remo, Nigeria

* Correspondence: Oluwayemisi Boye Fatade (fatadeo@babcock.edu.ng)

Received: 03-05-2025 **Revised:** 04-12-2025 **Accepted:** 04-18-2025

Citation: A. O. Kuyoro, O. B. Fatade, and E. E. Onuiri, "Enhancing non-invasive diagnosis of endometriosis through explainable artificial intelligence: A Grad-CAM approach," *Acadlore Trans. Mach. Learn.*, vol. 4, no. 2, pp. 97–108, 2025. https://doi.org/10.56578/ataiml040203.

Abstract: Significant advancements in artificial intelligence (AI) have transformed clinical decision-making, particularly in disease detection and management. Endometriosis, a chronic and often debilitating gynecological disorder, affects a substantial proportion of reproductive-age women and is associated with pelvic pain, infertility, and a reduced quality of life. Despite its high prevalence, non-invasive and accurate diagnostic methods remain limited, frequently resulting in delayed or missed diagnoses. In this study, a novel diagnostic framework was developed by integrating deep learning (DL) with explainable artificial intelligence (XAI) to address existing limitations in the early and non-invasive detection of endometriosis. Abdominopelvic magnetic resonance imaging (MRI) data were obtained from the Crestview Radiology Center in Victoria Island, Lagos State. Preprocessing procedures, including Digital Imaging and Communications in Medicine (DICOM)-to-PNG conversion, image resizing, and intensity normalization, were applied to standardize the imaging data. A U-Net architecture enhanced with a dual attention mechanism was employed for lesion segmentation, while Gradient-weighted Class Activation Mapping (Grad-CAM) was incorporated to visualize and interpret the model's decision-making process. Ethical considerations, including informed patient consent, fairness in algorithmic decision-making, and mitigation of data bias, were rigorously addressed throughout the model development pipeline. The proposed system demonstrated the potential to improve diagnostic accuracy, reduce diagnostic latency, and enhance clinician trust by offering transparent and interpretable predictions. Furthermore, the integration of XAI is anticipated to promote greater clinical adoption and reliability of AI-assisted diagnostic systems in gynecology. This work contributes to the advancement of non-invasive diagnostic tools and reinforces the role of interpretable DL in the broader context of precision medicine and women's health.

Keywords: AI; Endometriosis; Grad-CAM; Non-invasive diagnosis; U-Net; XAI

1 Introduction

AI has grown at an exponential rate over the last ten years, completely changing the technological landscape [1]. It has evolved beyond its original purpose of automating human labor. It's currently causing a paradigm change in the way people approach problems and come up with solutions. This revolution is a sharp contrast to the early days of computing when the purpose of machines was to merely increase human efficiency through fundamental computations. AI's revolutionary potential is causing a fundamental revolution in every part of the world, affecting everything from product creation to medical diagnosis. Similarly, Lutomski et al. [2] also explained that AI now has some subfields under it, especially in the practice of medicine such as computer vision (CV), DL and machine learning (ML).

In recent years, healthcare has experienced a lot of innovation concerning the advent of AI. In the aspect of disease diagnosis, detection, management and treatment, AI provides a more robust level of analysis for huge biomedical datasets. This has led to a great reduction in overall time spent on diagnosis together with lower costs on manpower and other linked resources. Indeed, healthcare digitalization brought about by AI is a plus to the medical world. In the field of women's health from obstetrics to gynecology, ML, as a subset of AI, has several methods, including logistic regression, support vector machines (SVMs) and many others, which have shown great potential to aid in the prediction of results for the diagnosis of endometriosis [3–5]. Given the diversity of its use in the clinical context,

there is great potential to apply ML to improve non-invasive diagnosis in endometriosis to reduce the delays and human error associated with diagnosis.

Endometriosis, a chronic gynecological condition, significantly impacts women's quality of life, causing pain and potential infertility [6]. It is categorized by endometrial-like tissue which is seen outside the uterus, and it is a persistent, estrogen-related condition. Inflammatory reactions and tissue damage are the outcomes of this abnormality. It is still difficult to confirm exactly how prevalent endometriosis is. However, estimates place it at 10% of women in reproductive age having the disease, with 30-50% of them reporting pelvic pain and/or infertility. Despite its prevalence, non-invasive diagnosis remains challenging. Researchers have explored various ML algorithms using data from symptoms, genetics, blood tests, and imaging. Approaches like logistic regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression have shown promise [7]. However, significant limitations remain, hindering clinical adoption and patient outcomes. One main limitation is that many ML models employed in current research are "black boxes," meaning their decision-making processes are not transparent [8]. This lack of interpretability makes it difficult for healthcare providers to understand how diagnoses are reached, limiting trust and hindering widespread clinical adoption.

In the application of ML techniques for the diagnosis of endometriosis in a bid to lessen the burdens on women experiencing endometriosis and help medical practitioners diagnose it more easily and early, studies have shown that different methods can be used for the categorization and classification of endometrial tissue lesions to other tumors and inflation, such as texture analysis of MRI images to differentiate endometriosis and hemorrhagic ovarian cysts [9]. Others include an automatic DL-based segmentation model combined with Receiver Operating Characteristic (ROC) analysis of tumor-to-uterine ratio on MRI images, which can effectively diagnose early-stage endometrial cancer [10], while several studies, as highlighted by Bhardwaj et al. [6], also use clinical data combined with MRI images to predict endometriosis.

2 Related Works

Zhang et al. [11] leveraged biomarkers for endometriosis prediction using multiple ML techniques, including LASSO, Stepglm, glmBoost, and random forest, to enhance predictive accuracy. Transcriptomics and methylomics data modalities were comprehensively integrated, achieving an Area Under the Curve (AUC) of 0.785, though the proposed method was limited in its applicability across broader populations. Similarly, GenomeForest, an ensemble ML classifier, demonstrated high F1-scores (0.98 for transcriptomics and 0.918 for methylomics) but was constrained by the reliance on biomarker sources from blood tests. Kurata et al. [12] evaluated the feasibility of using U-Net for automatic uterine segmentation on MRI images. The model was tested on patients with various uterine disorders, achieving a mean Dice similarity coefficient (DSC) of 0.82. The mean DSCs for patients with and without uterine disorders were 0.84 and 0.78, respectively (p $^1/_4$ 0.19). The Mean Absolute Deviations (MADs) for patients with and without uterine disorders were 18.5 and 21.4 [pixels], respectively (p $^1/_4$ 0.39). The scores of the visual evaluation were not significantly different between uteruses with and without uterine disorders. The results suggest that U-Net can effectively segment the uterus, regardless of the presence of disorders.

Other studies have explored imaging-based approaches. Downing et al. [13] developed an automated classification algorithm using imaging data with random forest classifiers, while Guerriero et al. [14] compared seven ML models using ultrasound markers, achieving an accuracy of 0.73. These approaches underscore the potential of imaging techniques but highlight the necessity for more extensive prospective studies to validate AI applications in clinical settings. ML models using clinical history and patient demographics have also been explored. Bendifallah et al. [15] implemented logistic regression, SVM, and random forest models and achieved an AUC of 0.98. Similarly, Tore et al. [16] developed an ML platform incorporating logistic regression, decision trees, and Shapley Additive Explanations (SHAP)-based interpretability methods, emphasizing the importance of feature attribution in clinical diagnosis. However, reliance on medical records introduced potential misclassification biases, requiring further validation with diverse datasets. Genomic and proteomic analyses also play a crucial role in ML-driven endometriosis research. Li et al. [17] used DL techniques to diagnose endometriosis based on gene co-expression networks. Mihalyi et al. [18] compared logistic regression and Least Squares Support Vector Machine (LSSVM) models using plasma biomarkers, reporting an AUC of 0.966 with high sensitivity and specificity. Despite promising results, these studies highlight the need for external validation to ensure generalizability. Parlatan et al. [19] investigated emerging modalities such as Raman spectroscopy, demonstrating the potential for novel non-invasive diagnostic techniques. However, limitations such as sample sizes and lack of external validation remain significant challenges.

Additionally, self-reported symptoms and questionnaire-based approaches have been explored. Knific et al. [20] and Goldstein and Cohen [21] investigated ML-based classification using patient-reported data, achieving varying degrees of success. While their approaches offer a patient-centric diagnostic tool, the need for validated questionnaires and larger sample sizes remains a gap in research. Accurate segmentation of endometriosis lesions is essential for diagnosis and treatment planning. Ronneberger et al. [22] demonstrated the strong performance of U-Net, a widely used convolutional neural network, in biomedical image segmentation, including endometriosis. High segmentation

accuracy using U-Net-based models was reported, such as the Structural Similarity Analysis of Endometriosis (SSAE), which achieved an intersection over union (IoU) of 0.72 and an F1-score of 0.74 on a large laparoscopic dataset. Shorten and Khoshgoftaar [23] showed U-Net's effectiveness in segmenting anatomical structures and quantifying blood perfusion during endometriosis surgeries, with Dice coefficients reaching 0.96. Additionally, U-Net was successfully applied to MRI-based uterine segmentation and endometrial cancer cell segmentation, with enhanced variants like U-Net_dc incorporating dense atrous convolution (DAC) and residual multi-kernel pooling (RMP) to improve feature extraction.

The integration of AI in medical diagnostics offers significant potential for improving accuracy and efficiency. However, the "black box" nature of many AI models, particularly DL, poses challenges for clinical adoption due to the lack of transparency and interpretability. XAI aims to bridge this gap by providing explanations for AI decisions, thereby enhancing trust and facilitating clinical integration [24]. Although many studies have investigated XAI for medical diagnosis, few of them focus on endometriosis. Thakur [25] conducted a case study of XAI in pneumonia detection using chest X-rays. The study aimed to integrate XAI with DL for pneumonia detection using CNN and use Grad-CAM for visual explanations. An accuracy of 93% was achieved, with explanations aligning well with radiologist assessments, increasing the trust in the model.

Yan et al. [26] proposed a comprehensive framework for XAI in brain tumor detection through MRI analysis, integrating segmentation and classification models to enhance diagnostic accuracy. By using the BraTS-2018 dataset, the proposed model achieved an impressive accuracy of 95.46%, while also emphasizing the importance of explainability in medical imaging to foster trust among healthcare professionals. A modified RepVGG architecture was utilized with gradient re-parameterization and Grad-CAM++ for improved performance and interpretability, ultimately highlighting the need for standardized evaluation metrics for explainability in medical contexts. Adopting XAI in diagnosing endometriosis plays a critical role in enhancing decision confidence and trustworthiness [27], providing deep insights into how the AI system arrives at a diagnosis. Clinicians can make informed decisions about its use, improving the level of trust and adaptability of the technology in the medical field. XAI has the power to identify potential bias in the model's predictions based on the features it prioritizes.

Several ML approaches, such as logistic regression, LASSO regression, and U-Net models, have been explored for non-invasive diagnosis, leveraging data from symptoms, genetic markers, blood tests, and imaging techniques. Limited attention mechanisms in the base U-Net architecture and lack of explainability remain a hinderance to clinical adoption and patient outcomes. The lack of explainability of most ML models used for endometriosis diagnosis functions as "black boxes", offering high accuracy but little insight into their decision-making process.

This study aims to address the limited attention mechanisms and lack of explainability by developing an explainable model. The proposed system can leverage Grad-CAM for MRI image interpretability, ensuring that healthcare providers can understand and trust model predictions.

3 Methodology

3.1 Dataset Overview

Abdominopelvic MRI images, obtained from the four branches of the Crestview Diagnostic Center in Nigeria (Lagos, Kano, Ilorin, and Ibadan) were used in this research because it is difficult to obtain a sizeable amount at a single location. The image dataset was collected retrospectively. Altogether, 1,208 medical records were obtained as MRI. Although endometriosis is prevalent, awareness and cultural factors still hinder most Nigerian women from accessing medical intervention when they notice it. Instead, those women rather push it aside as another "woman issue." Therefore, the number of records is limited. Figure 1 shows the framework of the U-Net model with a dual attention mechanism integrated with Grad-CAM.

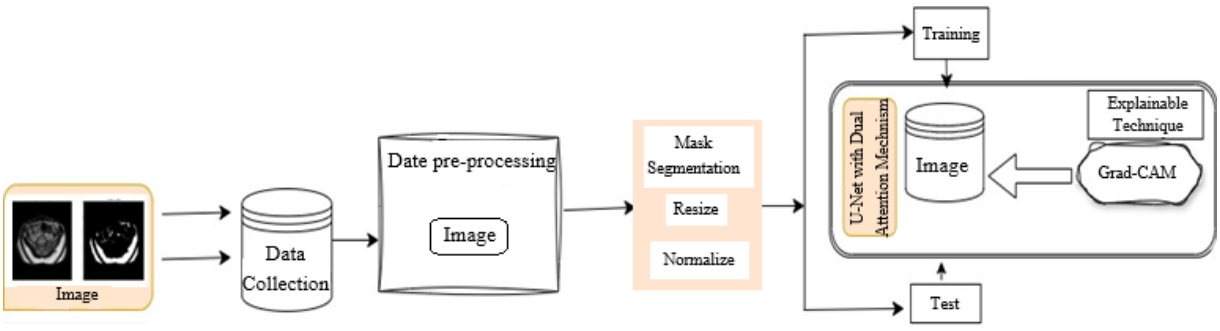

Figure 1. Framework of the U-Net model with a dual attention mechanism integrated with Grad-CAM

3.2 Data Preparation and Preprocessing

The MRI images were first inspected for quality, removing any corrupt or low-resolution scans. DICOM files are a standard format for medical images, containing both visual data and patient information converted to PNG format. Those files are sometimes too cumbersome for annotators unfamiliar with medical imaging software. Therefore, converting them to a simpler format like PNG makes the data more accessible. The DICOM files were read with Pydicom in a python environment and four conversion libraries (nibabel, cv2, numpy, and matplotlib) were utilized.

3.3 Mask Segmentation

The segmentation masks were semi-automatic annotations. The radiologist first annotated the area of interest using a radiography annotator and uploaded it to a 3D slicer for automatic segmentation and enhancement. The masks were stored as binary images. Pixel value 1 represents the lesion or region of interest (ROI) and 0 represents the non-lesion area.

3.4 MRI Data Resizing

For this research, resizing is necessary to ensure all images have a consistent dimension, which is required for feeding into the U-Net model. The dimension of 256 $\times$ 256 pixels was used to balance computational efficiency with image resolution. Because of the model used for this research, the bilinear interpolation method was used, which is less complex than other methods such as bicubic interpolation or nearest-neighbor interpolation. The bilinear interpolation makes for efficient computation while maintaining sufficient detail in the image. It calculates the pixel value by averaging the nearest four pixels using weighted averages based on the fractional distances α and β:

$$I\left(x', y'\right) = (1-\alpha)(1-\beta) \cdot I\left(x_1, y_1\right) + \alpha(1-\beta) \cdot I\left(x_2, y_1\right) + (1-\alpha)\beta \cdot I\left(x_1, y_2\right) + \alpha\beta \cdot I\left(x_2, y_2\right) \quad (1)$$

where, $I\left(x', y'\right)$ is the interpolated intensity value at the target coordinate (x', y'), which is estimated from the intensity values of its four neighboring pixels: $I\left(x_1, y_1\right), I\left(x_2, y_1\right), I\left(x_1, y_2\right)$, and $I\left(x_2, y_2\right)$ are the intensity values of the four neighboring pixels that surround (x', y'); (x_1, y_1) is the top-left pixel; (x_2, y_1) is the top-right pixel; (x_1, y_2) is the bottom-left pixel; (x_2, y_2) is the bottom-right pixel; α and β are the fractional distances between the target point (x', y) and the neighboring integer pixel coordinates; $\alpha = \left(x' - x_1\right) / \left(x_2 - x_1\right)$ is the fractional distance of x' between x_1 and x_2; and $\beta = \left(y' - y_1\right) / \left(y_2 - y_1\right)$ is the fractional distance of y' between y_1 and y_2.

Therefore, important structures in the MRI can be preserved without introducing too much blurring or pixelation. This is important in MRI images because even subtle differences can be clinically significant.

3.5 MRI Image Normalization

Normalization ensures that the U-Net model interprets the images on a consistent scale. It was applied after loading the images to scale pixel values to a standardized range of [0,1]. Since the images are in PNG format, each pixel's intensity originally ranged from 0 to 255. To normalize the values to the 0-1 range, the min-max normalization technique was used mainly because it ensures all pixel values fall within a standard range [0,1], which helps DL models like U-Net converge faster and generalize better and prevents large pixel intensity variations from affecting the learning process. Each pixel intensity was divided by 255, helping the model to be trained more effectively, as it standardizes the input range across images.

Data was organized into folders based on patient ID, with separate folders for MRI images, segmentation masks, clinical history, patient symptoms and demographics reports. The non-image dataset was split into 80% training, and 20% testing sets separately and the MRI dataset was split into 80% training and 20% testing sets.

3.6 Model Selection (U-Net with a Dual Attention Mechanism)

A U-Net with a dual attention mechanism was employed for the precise segmentation of endometriosis lesions from MRI images. U-Net architecture follows an encoder-decoder structure with symmetrical skip connections, allowing the model to retain high-resolution spatial information. Each encoder block consists of two 3$\times$3 convolutional layers followed by batch normalization and ReLU activation, ensuring stable gradient flow during training. Downsampling was performed using 2$\times$2 max pooling, progressively reducing spatial dimensions while increasing feature depth. The decoder mirrors this process, employing transposed convolutions for upsampling and concatenating corresponding encoder feature maps through skip connections to preserve fine-grained details. To prevent overfitting and encourage feature generalization, dropout layers (0.3-0.5 probability) were introduced within the decoder path.

The dual attention mechanism enhances feature selection by integrating both spatial and channel attention. The Position Attention Module (PAM) captures long-range spatial dependencies by computing feature interdependencies across different locations within the MRI scan. This is achieved through a self-attention mechanism that assigns higher weights to lesion regions while suppressing irrelevant background information. The Channel Attention

Module (CAM), on the other hand, enhances feature representations by applying global pooling and squeeze-and-excitation operations across different channels, allowing the model to emphasize the most relevant feature maps. The outputs from PAM and CAM were adaptively weighted and fused before being passed to the decoder, ensuring refined feature representations that improve segmentation accuracy.

Given the dataset's limited size, extensive data augmentation was applied using TensorFlow's augmentation layers to artificially increase training data variability and enhance model generalization. The applied transformations included random rotation (0°-20°) to account for MRI orientation differences, random width and height shifts (up to 10%) to address variations in patient positioning, and random zooming (up to 20%) to simulate different field-of-view settings. Additionally, horizontal flipping was used to improve spatial invariance, while elastic deformations simulated tissue distortions commonly observed in MRI scans. These augmentation techniques not only diversify the training data but also make the model more robust to real-world imaging variations.

To optimize the model's performance, a combination of Dice loss and Binary Cross-Entropy (BCE) loss was employed. Dice loss ensures accurate segmentation by addressing class imbalances, while BCE provides stable convergence during training. The model was trained using the Adam optimizer with an initial learning rate of $1e^{-4}$, incorporating a cosine decay schedule to dynamically adjust the learning rate over epochs. A batch size of 16 was chosen to balance computational efficiency and convergence stability. Early stopping was implemented, monitoring the validation Dice score to prevent overfitting and ensure the best model checkpoint is retained.

This enhanced U-Net with dual attention effectively captures the intricate patterns of endometriosis lesions while mitigating dataset limitations. By integrating both spatial and channel attention mechanisms, the model achieves improved segmentation accuracy, greater lesion localization precision, and better generalization to unseen MRI scans.

3.7 Justification for U-Net with a Dual Attention Mechanism

Due to the small dataset used in this study, it is imperative to study a model that can handle and do well with a small dataset. U-Net is a well-established architecture in medical image segmentation for instances with limited datasets. It is efficient for small datasets because it uses skip connections that allow features from the encoder to directly pass to the decoder. This design retains spatial information, which is crucial for segmenting fine details in medical images. By efficiently preserving spatial context, U-Net reduces the reliance on large datasets for learning, making it suitable for scenarios like this study where annotated data is scarce. U-Net with a dual attention mechanism enhances the basic U-Net architecture by incorporating spatial and channel-wise attention mechanisms, which improves the model's focus on relevant features in the input data. The lesions related to endometriosis are often small, irregularly shaped, and difficult to distinguish from surrounding tissues. The spatial attention makes sure that the model focuses on lesion regions and the channel attention prioritizes relevant feature maps, reducing noise from unrelated areas. The dual attention mechanism architecture for this project was chosen to address the limitation of the basic U-Net by selectively focusing on relevant features and regions.

3.8 Grad-CAM Integration

After training U-Net with the dual attention mechanism on the MRI datasets, the Grad-CAM explainability technique was integrated to gain meaningful insight from the model performance. The last convolutional layer in the U-Net was selected as the target for Grad-CAM because it retains both spatial and feature information that is necessary for generating meaningful heatmaps. The gradient output class score concerning the target layer's activation was computed using TensorFlow's Gradient Tape. The weighted sum of the activation was computed and passed through a ReLu function to retain only positive influences, and the heatmap was normalized to a range of [0,1] for visualization.

3.9 Image Data Training

The Adam optimizer was used with a learning rate of 0.001, selected to ensure stable convergence with a batch size of 16, balancing memory constraints with training efficiency. Initially, the model was trained for 20 epochs, with adjustments based on validation performance and convergence. The model was initialized and compiled with the Adam optimizer, BCE loss, and accuracy metrics. The training progress was monitored by calculating loss and accuracy on the training and validation sets. After each epoch, the model evaluated validation data to monitor generalization. During training, the model's weights were saved periodically based on validation performance to capture the best model state. Early stopping was used to halt training when the validation loss stopped improving, reducing overfitting risk. This criterion helped optimize training time by avoiding unnecessary epochs. The evaluation metrics of F1-score, accuracy, and recall were used in evaluating the performance of the model.

Algorithm: U-Net with a dual attention mechanism

Step 1: Input image preprocessing

• Load input image I of size H × W × C.

• Normalize pixel values to [0,1].

• Apply data augmentation using TensorFlow augmentation layers, including random rotation (0°-20°), random width and height shifts (up to 10%), random zooming (up to 20%), horizontal flipping, and elastic deformations.

Step 2: Encoding path (downsampling)

• Pass input image I through successive convolutional blocks, with each block consisting of Conv2D layer $F_{enc} = Conv(I)$, batch normalization, ReLu activation, and max pooling (reduced spatial dimensions).

• Store features maps from each block for the skip connections.

Step 3: Bottleneck layer (bridge)

• Process the lowest-resolution feature maps through additional convolutions and non-linear activations to extract deeper representations.

• Apply a dual attention mechanism to enhance feature learning. As for the channel attention mechanism, the specific steps involve computing global average pooling $S_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} A_{ij}^k$, applying a fully connected network to generate attention weights $\alpha^k = \sigma\left(FC\left(s_k\right)\right)$, and scaling feature maps $A^k CAM = \alpha^k \cdot A^k$. As for the spatial attention mechanism, the specific steps involve computing channel-wise pooling (average and max pooling), applying a 3 × 3 convolution followed by a sigmoid activation $s_{ij} = \left(f_{3\times3}\left(\left[A^{\text{avg}}, A^{\max}\right]\right)\right)$, and scaling feature maps $A_{SAM} = S.A$. The last step involves combining the outputs of the spatial and channel attention mechanisms to obtain the final attention-enhanced feature map $A^* = A_{CAM} + A_{SAM}$.

Step 4: Decoding path (upsampling)

• Perform upsampling to restore spatial resolution, which involves transposing convolution (deconvolution) of Upsampling2D, concatenating encoder features with upsampling features, and applying convolutional layers with ReLU activation to refine the features.

Step 5: Output

• Apply a final convolution layer with sigmoid activation $P = \left(Conv_{1\times1}\left(F_{dec}\right)\right)$.

• Output segmentation mask M where each pixel represents the probability of belonging to a lesion region.

Step 6: Model explainability using Grad-CAM

• Extract the final convolutional layer feature maps A^k.

• Compute the Grad-CAM importance weights, which involves computing the gradient of the segmentation score y^c with respect to each feature map $A^k : \frac{\partial y^c}{\partial A_{ij}^k}$ and computing the global importance weight for each feature map $\propto_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial_{ij}^k}$ where z is the number of spatial locations H × W Compute the Grad-CAM heatmap $L_{Grad-CAM}^c = ReLU\left(\sum_k \propto_k^c A^k\right)$. The $ReLU(x) - \max(0, x)$.

• Upsample the heatmap to match the input image size using bilinear interpolation

$$\mathrm{L}_{\text{Grad-CAM}}^{\mathrm{C}} = \text{Upsample}\left(\mathrm{L}_{\mathrm{Grad-CAM}}^{\mathrm{C}},\ \text{size} = 1\right)$$

• Overlay the heatmap on the original image for visualization.

4 Results

Data preparation and preprocessing are essential steps in ensuring that raw data is transformed into a clean and structured format suitable for analysis and ML tasks. This section outlines the process undertaken to prepare a medical imaging dataset for analysis, focusing on the extraction of metadata, file organization, missing data handling, and preprocessing steps for ML. The dataset under review consists of anonymized medical imaging data stored in DICOM format, organized into folders representing individual patients. Metadata extraction involves identifying the structure and content of the dataset and verifying its completeness.

4.1 Model Implementation

The base model used for this study is a U-Net, which follows an encoder-decoder architecture. Dual attention blocks were incorporated into the network to enhance segmentation accuracy. To improve the interpretability of segmentation predictions, Grad-CAM was integrated. A combination of the loss function of BCE and Dice loss was used in the training process and the Adam optimizer was utilized with a learning rate of 0.0001. A batch size of 16 was employed, and the model was trained for 15 epochs, with early stopping applied if the validation loss stabilized. The training and validation losses, along with Dice scores, are presented in Table 1.

Table 1. Model training results

Epoch	Train Loss	Train Dice Score	Val Loss	Val Dice Score
1	0.48	0.72	0.50	0.70
2	0.445	0.74	0.465	0.73
3	0.41	0.76	0.43	0.74
4	0.375	0.78	0.395	0.76
5	0.34	0.80	0.38	0.78
6	0.318	0.82	0.362	0.79
7	0.296	0.84	0.344	0.80
8	0.274	0.85	0.326	0.81
9	0.252	0.855	0.308	0.82
10	0.28	0.86	0.31	0.83
11	0.266	0.872	0.296	0.835
12	0.252	0.884	0.282	0.84
13	0.238	0.896	0.268	0.845
14	0.224	0.908	0.254	0.86
15	0.23	0.90	0.27	0.87

The loss function decreases steadily, demonstrating effective learning, as shown in Figure 2. The Dice score reaches 0.9 on the training set and 0.87 on the validation set, indicating high segmentation quality. The validation performance lags slightly behind training, suggesting minor overfitting. The early stopping suggests an optimal stopping point because training beyond 15 epochs may lead to diminishing returns.

Figure 2. A snapshot of the compiled U-Net with a dual attention mechanism

4.2 Model Performance on Test Data

Strong performance on test data was demonstrated by the trained model. Table 2 shows the result with a Dice score of 86.5% and an IoU of 89%. The recall value of 84% shows reliable segmentation accuracy. The confusion matrix analysis for binary segmentation reveals, as described in Table 3, that 90% of actual positive cases are correctly identified as true positives while 10% are missed as false negatives. Similarly, 90% of actual negative cases are correctly classified as true negatives with a false positive rate of 10%. The false negative rate suggests that some areas of endometriosis may go undetected, while the false positive rate indicates occasional misclassification of non-endometriosis regions.

Table 2. Model training results

Matrix	Result
Dice score	86.5 %
IoU	89 %
Recall	84 %

Table 3. Result of confusion matrix

Actual	Predicted Positive	Predicted Negative
Actual positive	True positive $\sim$ 90 %	False negative $\sim$ 10 %
Actual negative	False positive $\sim$ 10 %	True negative $\sim$ 90 %

4.3 Grad-CAM Integration

The Grad-CAM accuracy of 0.7001 for integrating Grad-CAM with attention mechanisms highlights the near accuracy of achieving effective explainability for segmentation, as shown in Table 4. Figure 3 highlights the Grad-CAM heatmap, where areas of interest were highlighted with red and yellow regions, indicating where the model focused during prediction. When overlaid on the MRI scan, the heatmap visually demonstrates the relevance of the model's segmentation, providing insights into how it identifies affected regions. Grad-CAM successfully generates visual explanations for the U-Net model's decision.

Table 4. Grad-CAM integration result

Metric	Value
Grad-CAM accuracy	0.7001

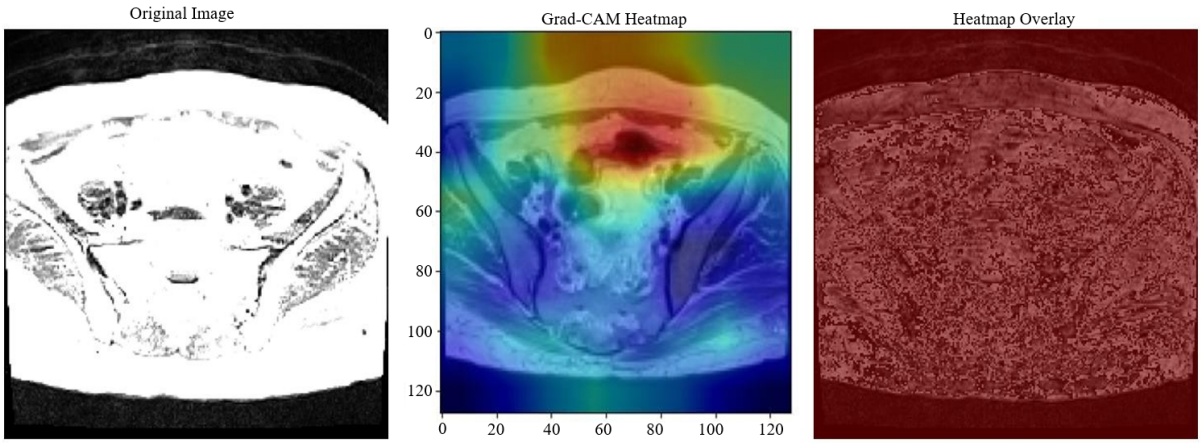

Figure 3. Grad-CAM heatmap

Trust and understanding in the evaluation of Grad-CAM were assessed through localization accuracy and faithfulness. These metrics help determine whether the model's heatmaps correctly highlight relevant areas and whether they truly explain the model's decision-making process.

4.4 Trust Through Localization Accuracy

The result for the trust through localization test indicates, as shown in Figure 4, that the Grad-CAM heatmap aligns well with the ground truth segmentation mask, indicating that the model focuses on the correct regions when making predictions. For this research, this is indicated using IoU and DSC between the Grad-CAM heatmap and the ground truth segmentation mask. The IoU of 0.75 and DSC of 0.87 show good overlap and minor false positives.

4.5 Understanding Through Faithfulness

The result for understanding through faithfulness in this research was obtained by using the pixel perturbation test. The high-importance regions were gradually occluded from the Grad-CAM heatmap and changes in the model's prediction were observed as the changes occur. As shown in Figure 5, there is a significant drop in prediction confidence when the important regions are occluded, indicating that the explanation is faithful. This is measured using the drop in confidence (DropC).

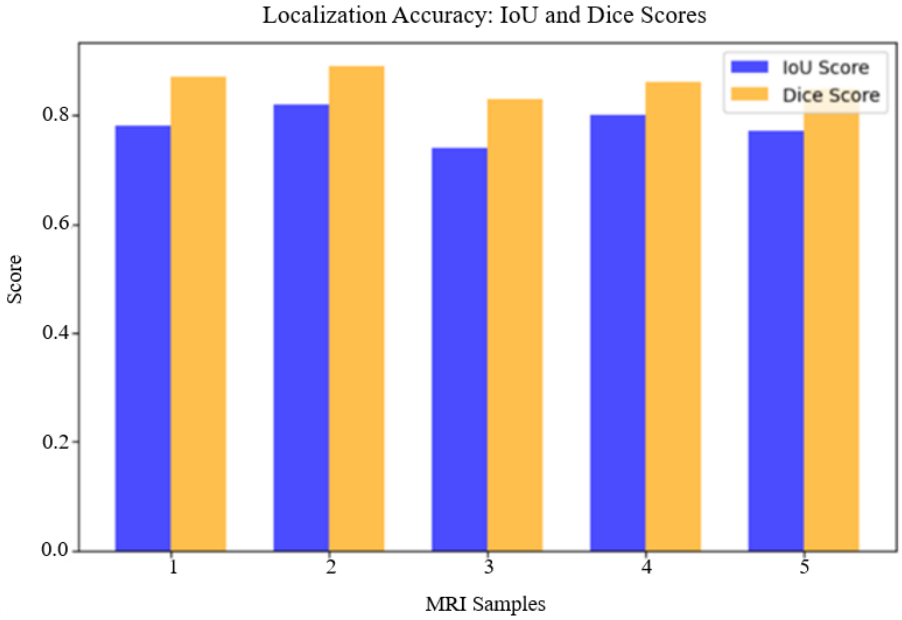

Figure 4. Visualization of trust through localization accuracy

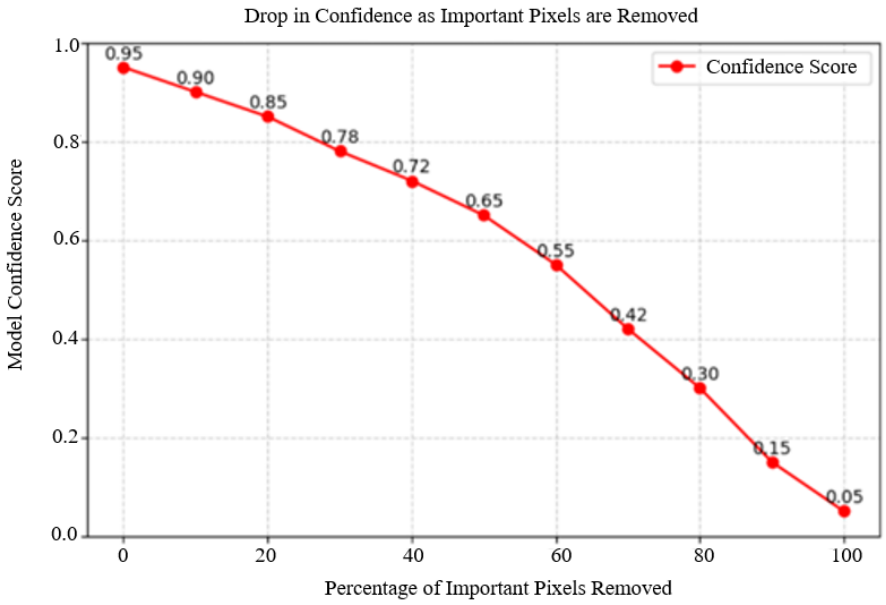

Figure 5. Visualization of understanding through faithfulness using a pixel perturbation test

4.6 Trust and Understanding Through Consistency of Grad-CAM Output

Quantitative trust and understanding metrics were inferred through Grad-CAM visualization alignment. The results in Table 5 and the visual representation in Figure 6 indicate an average trust score of 4.625 and an average understanding score of 4.775, suggesting high confidence among healthcare professionals.

Table 5. Measure of understanding and trust

Metric	Value
Average trust	4.625
Average understanding	4.775

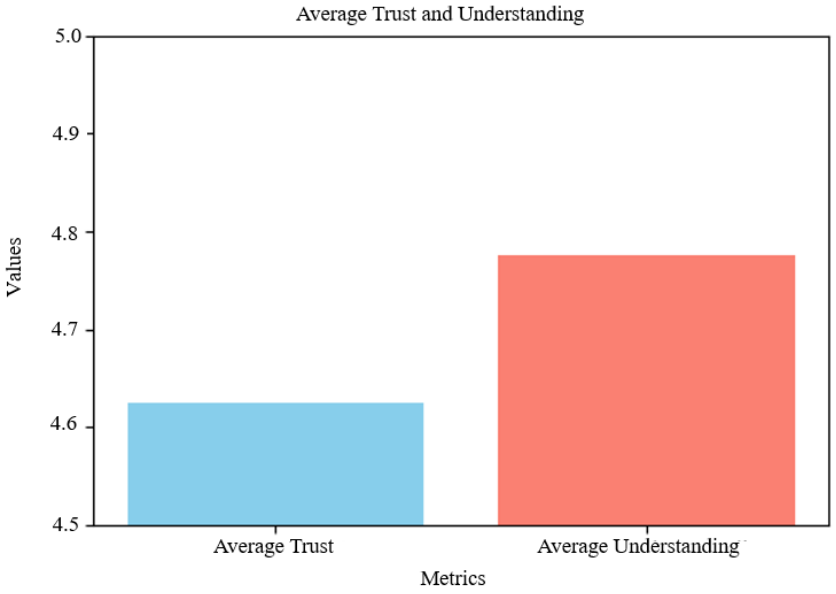

Figure 6. Trust and understanding evaluation metric

5 Scope of Study

This study focuses on the development and evaluation of a U-Net with a dual attention mechanism for the segmentation of endometriosis lesions from MRI images. The primary objective is to enhance model explainability using Grad-CAM while ensuring robust segmentation performance. The scope of this research is defined as follows:

- The study utilizes MRI scans for non-invasive endometriosis detection, without integrating additional diagnostic modalities such as ultrasound or histopathological data.
- The research focuses on DL-based segmentation and does not perform direct comparisons with traditional radiological diagnostic methods.
- The model is optimized for image-based segmentation rather than broader clinical decision support, such as patient history analysis or multi-modal data fusion.
- The study primarily evaluates the effectiveness of the proposed U-Net with dual attention rather than conducting an exhaustive comparison with alternative DL architectures such as Transformer-based models or hybrid approaches.
- Grad-CAM is used as the explainability method, with a focus on heatmap visualization and trust evaluation. Other explainability techniques are not explored in detail.
- While the study assesses trust and understanding through quantitative measures, it does not include direct radiologist feedback or qualitative user studies, which are suggested for future work.
- Clinical adoption and regulatory considerations are beyond the scope of this study but are recognized as important areas for future investigation.

6 Discussion

The implementation of the U-Net with a dual attention mechanism achieved promising results, particularly with IoU and recall at 89% and 84%, respectively. The confusion matrix analysis for binary segmentation reveals, as described in Table 3, that 90% of actual positive cases are correctly identified as true positives while 10% are missed as false negatives. Similarly, 90% of actual negative cases are correctly classified as true negatives with a false positive rate of 10%. The false negative rate suggests that some areas of endometriosis may go undetected, while the false positive rate indicates occasional misclassification of non-endometriosis regions.

The Grad-CAM accuracy is 0.7001 which is an acceptable percentage but could be better, especially if it is adopted in the medical professional field. The Grad-CAM heatmap highlights areas of interest, with red and yellow regions indicating where the model focused during prediction. When overlaid on the MRI scan, the heatmap visually demonstrates the relevance of the model's segmentation, providing insights into how it identifies affected regions. Grad-CAM successfully generates visual explanations for the U-Net model's decisions.

The trust and understanding results of the system based on accuracy and the Grad-Cam heatmap are promising, with an average trust score of 4.625 and an average understanding score of 4.775. These high scores reflect that there is a high chance for the system to be accepted by the healthcare providers. This is crucial for the successful integration of the diagnostic system into clinical practice, as trust and understanding are key factors in the adoption

of new technology. The results also underscore the importance of XAI in fostering trust, with the incorporation of Grad-CAM and attention mechanisms playing a significant role in improving transparency.

7 Conclusions

Endometriosis, a chronic and frequently debilitating ailment, affects millions of women worldwide, but it is one of the most underdiagnosed and misunderstood gynaecological conditions. Traditional diagnostic approaches in the past relied mainly on intrusive procedures such as laparoscopy, as accurate as it is, can cause delays in diagnosis and treatment, worsening the physical, emotional and psychological toll on patients. This study addressed these important shortcomings by creating an explainable ML model that provides a non-invasive, accurate, and transparent diagnostic option for endometriosis.

The strategy involved the use of XAI approaches such as Grad-CAM, which promotes openness, allowing clinicians to comprehend the underlying principles of diagnostic choices. This is critical for clinical acceptance. In addition, the approach appears to be a promising alternative to invasive treatments, with the potential to reduce diagnostic delays and associated healthcare expenses for women, particularly in developing countries such as Nigeria. These findings highlight the transformational power of merging ML and multimodal data in medical diagnostics. By utilizing this approach, this study opens the door for improving early detection, enhancing patient outcomes, and fostering trust in AI-driven healthcare solutions.

8 Ethical Consideration

This study was conducted in accordance with ethical guidelines to ensure patient data protection and responsible AI implementation. Ethical approval was obtained from the Babcock University Research Ethical Committee, ensuring compliance with institutional and regulatory standards for medical research.

Additionally, a low-risk ethical review was granted by the Crestview Radiology Ltd Research Ethics Committee, which oversees ethical compliance for research involving medical imaging data. As part of this review, a Low-Risk Ethical Review Form was completed and approved. This form is specifically used by researchers analyzing anonymized data from medical databases, confirming that no personally identifiable information was accessed or used in the study.

All MRI data utilized in this research were fully anonymized prior to release, ensuring strict adherence to patient confidentiality and data protection regulations. The study follows the fundamental ethical principles of medical research, including non-maleficence, patient privacy, and responsible AI deployment in clinical decision support systems.

Data Availability

Not applicable.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] F. Wang and A. Preininger, "AI in health: State of the art, challenges, and future directions," *Yearb. Med. Inform.*, vol. 28, no. 1, pp. 16–26, 2019. https://doi.org/10.1055/s-0039-1677908

[2] J. E. Lutomski, S. Meaney, R. A. Greene, A. C. Ryan, and D. Devane, "Expert systems for fetal assessment in labour," *Cochrane Database Syst. Rev.*, vol. 2015, no. 4, p. CD010708, 2015. https://doi.org/10.1002/14651858.CD010708.pub2

[3] B. Sivajohan, M. Elgendi, C. Menon, C. Allaire, P. Yong, and M. A. Bedaiwy, "Clinical use of artificial intelligence in endometriosis: A scoping review," *npj Digit. Med.*, vol. 5, p. 109, 2022. https://doi.org/10.1038/s41746-022-00638-1

[4] M. Elgendi, C. Allaire, C. Williams, M. A. Bedaiwy, and P. J. Yong, "Machine learning revealed new correlates of chronic pelvic pain in women," *Front. Digit. Health*, vol. 2, p. 600604, 2020. https://doi.org/10.3389/fdgth.2020.600604

[5] T. Yoldemir, "Artificial intelligence and women's health," *Climacteric*, vol. 23, no. 1, pp. 1–2, 2020. https://doi.org/10.1080/13697137.2019.1682804

[6] V. Bhardwaj, A. Sharma, S. V. Parambath, I. Gul, X. Zhang, P. E. Lobie, P. W. Qin, and V. Pandey, "Machine learning for endometrial cancer prediction and prognostication," *Front. Oncol.*, vol. 12, p. 852746, 2022. https://doi.org/10.3389/fonc.2022.852746

[7] M. Szubert, A. Rycerz, and J. R. Wilczyński, "How to improve non-invasive diagnosis of endometriosis with advanced statistical methods," *Medicina*, vol. 59, no. 3, p. 499, 2023. https://doi.org/10.3390/medicina59030499

[8] M. Ridley, "Explainable Artificial Intelligence (XAI): Adoption and advocacy," *Inf. Technol. Libr.*, vol. 41, no. 2, pp. 1–17, 2022. https://doi.org/10.6017/ital.v41i2.14683

[9] R. A. Lupean, P. A. Ştefan, C. Csutak, A. Lebovici, A. M. Măluţan, R. Buiga, C. S. Melincovici, and C. M. Mihu, "Differentiation of endometriomas from ovarian hemorrhagic cysts at magnetic resonance: The role of texture analysis," *Medicina*, vol. 56, no. 10, p. 487, 2020. https://doi.org/10.3390/medicina56100487

[10] W. Mao, C. X. Chen, H. C. Gao, L. Xiong, and Y. P. Lin, "A deep learning-based automatic staging method for early endometrial cancer on MRI images," *Front. Physiol.*, vol. 13, p. 974245, 2022. https://doi.org/10.3389/fphys.2022.974245

[11] H. L. Zhang, H. L. Zhang, H. Yang, A. N. Shuid, D. Sandai, and X. B. Chen, "Machine learning-based integrated identification of predictive combined diagnostic biomarkers for endometriosis," *Front. Genet.*, vol. 14, p. 1290036, 2023. https://doi.org/10.3389/fgene.2023.1290036

[12] Y. Kurata, M. Nishio, A. Kido, K. Fujimoto, M. Yakami, H. Isoda, and K. Togashi, "Automatic segmentation of the uterus on MRI using a convolutional neural network," *Comput. Biol. Med.*, vol. 114, p. 103438, 2019. https://doi.org/10.1016/j.compbiomed.2019.103438

[13] M. J. Downing, D. J. Papke, S. Tyekucheva, and G. L. Mutter, "A new classification of benign, premalignant, and malignant endometrial tissues using machine learning applied to 1413 candidate variables," *Int. J. Gynecol. Pathol.*, vol. 39, no. 4, pp. 333–343, 2020. https://doi.org/10.1097/PGP.0000000000000615

[14] S. Guerriero, L. Saba, M. A. Pascual, S. Ajossa, I. Rodriguez, V. Mais, and J. L. Alcazar, "Transvaginal ultrasound vs magnetic resonance imaging for diagnosing deep infiltrating endometriosis: Systematic review and meta-analysis," *Ultrasound Obstet. Gynecol.*, vol. 51, no. 5, pp. 586–595, 2018. https://doi.org/10.1002/uog.18961

[15] S. Bendifallah, Y. Dabi, S. Suisse, L. Jornea, D. Bouteiller, C. Touboul, A. Puchar, and E. Daraï, "MicroRNome analysis generates a blood-based signature for endometriosis," *Sci. Rep.*, vol. 12, p. 4051, 2022. https://doi.org/10.1038/s41598-022-07771-7

[16] U. Tore, A. Abilgazym, A. Asunsolo-del Barco, M. Terzic, Y. Yemenkhan, A. Zollanvari, and A. Sarria-Santamera, "Diagnosis of endometriosis based on comorbidities: A machine learning approach," *Biomedicines*, vol. 11, no. 11, p. 3015, 2023. https://doi.org/10.3390/biomedicines11113015

[17] B. H. Li, S. Wang, H. Duan, Y. Y. Wang, and Z. C. Guo, "Discovery of gene module acting on ubiquitin-mediated proteolysis pathway by co-expression network analysis for endometriosis," *Reprod. Biomed. Online*, vol. 42, no. 2, pp. 429–441, 2021. https://doi.org/10.1016/j.rbmo.2020.10.005

[18] A. Mihalyi, O. Gevaert, C. M. Kyama, P. Simsa, N. Pochet, F. De Smet, B. De Moor, C. Meuleman, J. Billen, N. Blanckaert, A. Vodolazkaia, V. Fulop, and T. M. D'Hooghe, "Non-invasive diagnosis of endometriosis based on a combined analysis of six plasma biomarkers," *Hum. Reprod.*, vol. 25, no. 3, pp. 654–664, 2010. https://doi.org/10.1093/humrep/dep425

[19] U. Parlatan, M. T. Inanc, B. Y. Ozgor, E. Oral, E. Bastu, M. B. Unlu, and G. Basar, "Raman spectroscopy as a non-invasive diagnostic technique for endometriosis," *Sci. Rep.*, vol. 9, p. 19795, 2019. https://doi.org/10.1038/s41598-019-56308-y

[20] T. Knific, D. Fishman, A. Vogler, M. Gst"ottner, R. Wenzl, H. Peterson, and T. L. Rižner, "Multiplex analysis of 40 cytokines do not allow separation between endometriosis patients and controls," *Sci. Rep.*, vol. 9, p. 16738, 2019. https://doi.org/10.1038/s41598-019-52899-8

[21] A. Goldstein and S. Cohen, "Self-report symptom-based endometriosis prediction using machine learning," *Sci. Rep.*, vol. 13, p. 5499, 2023. https://doi.org/10.1038/s41598-023-32761-8

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015. https://doi.org/10.48550/arXiv.1505.04597

[23] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, p. 60, 2019. https://doi.org/10.1186/s40537-019-0197-0

[24] A. Chaddad, J. H. Peng, J. Xu, and A. Bouridane, "Survey of explainable AI techniques in healthcare," *Sensors*, vol. 23, no. 2, p. 634, 2023. https://doi.org/10.3390/s23020634

[25] R. Thakur, "Explainable AI: Developing interpretable deep learning models for medical diagnosis," *Int. J. Multidiscip. Res.*, vol. 6, no. 4, 2024.

[26] F. Yan, Y. Q. Chen, Y. W. Xia, Z. L. Wang, and R. X. Xiao, "An explainable brain tumor detection framework for MRI analysis," *Appl. Sci.*, vol. 13, no. 6, p. 3438, 2023. https://doi.org/10.3390/app13063438

[27] A. M. Antoniadi, Y. H. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review," *Appl. Sci.*, vol. 11, no. 11, p. 5088, 2021. https://doi.org/10.3390/app11115088