RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt26423

Explainable Hybrid Deep Learning Approach for Multi-Modality Uterine Fibroid Detection and Clinical Decision Support

Trupti Kherde [a, *]

[a] Department of Computer Engineering, Pimpri Chinchwad College of Engineering and Research, Ravet Pune, India

* Corresponding Author Email: trupti.kherde@pccoer.in

DOI: https://doi.org/10.54392/irjmt26423

Received: 02-04-2026; Revised: 03-07-2026; Accepted: 13-07-2026; Published: 29-07-2026

Abstract: Uterine fibroids are one of the most common gynaecological tumours, but they can be hard to diagnose because their appearance varies and the images may have poor contrast. Existing computer-aided detection (CAD) methods often use features that were handcrafted or use only a single modality, which limits their reliability and interpretability. An Explainable Hybrid Deep Learning (E-HDL) framework is introduced in this study. It includes a self-improving preprocessing pipeline, adaptive multi-branch feature fusion, and attention-based interpretability for reliably detecting uterine fibroids on both ultrasound and MRI. To make sure that the input quality is the same, the process uses bias-field correction, anisotropic diffusion filtering, CLAHE contrast normalisation, and adaptive spatial alignment. Spatial deep features from a CNN are combined dynamically with contextual embeddings from Transformers and temporal representations from RNNs using an attention-based feature-scaling layer. Extensive experiments on the Mendeley Uterine Fibroid Dataset and cross-dataset evaluation on the HIFU-MRI and UFID-2023 benchmarks show that E-HDL works better than recent state-of-the-art models like Swin-Transformer, MedT, ConvNeXt, and EfficientNetV2. The suggested model is more accurate than previous combination and transformer-based systems, with a 97.1% F1-score and 0.989 AUC. Grad-CAM, SHAP, and attention heatmaps improve interpretability in clinical settings. The study shows that E-HDL is a high-performance, scalable, and easy-to-understand diagnostic aid for detecting uterine tumours.

Keywords: Uterine Fibroid Detection, Hybrid Deep Learning, Medical Image Analysis, Preprocessing Techniques, Diagnostic Accuracy, Clinical Decision Support.

1. Introduction

The most prevalent benign tumors in the gynaecological system of women of reproductive age are the uterine fibroids (leiomyomas), with a lifetime prevalence of 70-80%. Even though most of the cases are asymptomatic, a considerable percentage of patients experience heavy menstrual bleeding, pelvic pain, infertility, and problems during pregnancy resulting in a substantial healthcare burden and reduced quality of life. Early diagnosis is therefore essential for guiding treatment decisions and prevent unnecessary invasive surgeries like hysterectomy or myomectomy [1]. The most common diagnostic modalities are ultrasonography and magnetic resonance imaging (MRI) the former is more accessible and cost-effective, whereas the latter provides a better characterization of tissues and anatomical detail. Nonetheless, image interpretation is highly dependent on the experience of radiologists and the changes in appearance of the lesions, the presence of multiple fibroid masses and multifocal distributions frequently makes it difficult to detect the lesions reliably [2, 3].

In order to minimize subjectivity, classical computer-aided diagnostic (CAD) systems utilized machine learning algorithms like support vector machines, decision trees and random forests based on handcrafted texture and morphological features [4]. Although these methods indicated that automated analysis was possible, they were restricted by insufficient generalization across imaging devices, sensitivity to noise and contrast variation. Even more recently, deep learning algorithms, especially convolutional neural networks (CNNs) and transformers, have produced significant advances in medical-image classification and segmentation [5]. However, CNNs are primarily sensitive to local spatial information and they might not be able to resolve long-range contextual dependencies, whereas transformer models might require very large annotated datasets and may lack clinical interpretability. Moreover, even imaging quality heterogeneity, variation in preprocessing, and interinstitutional variability continue to impair robustness in clinical applications [6].

In the past few years, the automated uterine fibroid analysis has advanced considerably through the use of deep convolutional neural networks, transformer-based architectures, and multimodal learning frameworks, which have substantially advanced artificial-intelligence applications. Recently, transfer learning with EfficientNet, DenseNet and ResNet has shown significant improvements in the classification of lesions over handcrafted feature based approaches. Contextual representation learning has been enhanced by modelling long-range dependencies in medical images, like in the cases of Vision Transformer (ViT), Swin Transformer, MedT, and TransUNet, has improved performance further [7, 8]. A number of studies have investigated multimodal fusion of ultrasound and magnetic resonance imaging (MRI) to improve lesion localization, tissue characterization, and diagnostic consistency [9]. Concurrently, research has increasingly focused on explainable AI (XAI) methods to enhance the interpretability of deep learning models for clinical use, such as Grad-CAM, SHAP, LIME and attention visualization [10]. In addition, novel approaches in uncertainty estimation, such as Monte Carlo Dropout and Bayesian inference, have recently gained attention for improving diagnostic reliability as they provide both predictions and confidence estimates. While these developments show potential for good diagnostic performance, most existing state-of-the-art studies have been evaluated using single-centre datasets, with limited cross-dataset validation and thus limited generalizability in a heterogeneous clinical setting [11].

While great strides have been made in deep learning based uterine fibroid detection, several major research issues remain unresolved. First, most published works use fixed feature fusion approaches where the representations from convolutional and Transformer models are concatenated, as opposed to being dynamically fused during inference according to their relative importance [12]. Second, current frameworks do not combine three types of representation (spatial, contextual and sequential) into a single adaptive architecture that processes both ultrasound and MRI data simultaneously and can process heterogeneous data [13]. Third, although the use of visual explanation techniques has grown in recent years, the limited quantitative validation of explainability against expert annotations restricts clinicians' trust in automated decisions. Furthermore, uncertainty-aware inference is commonly absent in existing uterine fibroid diagnosis systems and making it difficult to assess predictive uncertainty in challenging clinical cases [14]. Last but not least, the majority of published methods have been tested on only a single dataset, and thus offer only limited evidence of robustness across diverse imaging distributions, acquisition protocols and healthcare institutions. These limitations motivate the development of a more comprehensive, flexible, interpretable and clinically viable diagnosis system [15].

While recent hybrid CNN–Transformer based approaches, ConvNeXt–Transformer, and multimodal deep learning frameworks have shown enhanced diagnostic performance on the analysis of uterine lesions, most of the current methods are based on static concatenation of features, lack uncertainty-aware inference, provide limited quantitative validation of explainability, and have not been rigorously evaluated on independent datasets:

1. Dynamic refinement of input normalization with adaptive pre-processing feedback loop,
2. A Dynamic Attention Fusion Layer (DAFL) which assigns learnable, data-dependent weights to spatial, contextual and sequential branches,
3. Integration of uncertainty estimation through Monte-Carlo Dropout and clinical confidence scoring, and
4. Quantitative analysis of interpretability based on localization IoU and attribution confidence metrics.

All these factors distinguish E-HDL compared to the previous hybrid architectures and adapt the framework specifically for robust cross-modality uterine fibroid detection. The ConvNeXt branch is used for hierarchical anatomical representation learning of heterogeneous fibroid morphology, distinguishing it from the traditional CNN architectures that mainly capture local spatial structures. The Swin-Transformer branch models long-range contextual dependencies among uterine-tissue regions surrounding lesions, thereby improving discrimination between lesions with varying imaging contrast and boundaries. The dataset consisted mainly of independent ultrasound images, however for the LSTM branch, the sequential contextual dependency was modelled of the ultrasound images, which were derived from ordered feature embeddings and progressive multiscale representations in the hybrid architecture. This complementary integration enables simultaneous learning of spatial, contextual and representation-sequence characteristics that are not captured by CNN-only, Transformer-only, or conventional uterine fibroid detection systems.

To address the above research gaps, the proposed Explainable Hybrid Deep Learning (E-HDL) framework introduces several methodological innovations that enhance diagnostic robustness, interpretability, and generalizability across modalities:

1. A dynamic, adaptive preprocessing-feedback mechanism that optimizes bias correction and contrast normalization during training.
2. A Dynamic Attention Fusion Layer (DAFL) that allows data-dependent weighting of the CNN, Transformer, and RNN feature branches instead of relying on simple concatenation.

3. Quantitative interpretability validation using localization IoU and attribution-consistency measures.

2. Literature Review

Automated uterine-fibroid detection has evolved from classical image processing methods to the advanced deep learning models. The initial computer-assisted diagnostic (CAD) systems were based on segmentation techniques like thresholding, morphological operations and region-growing to outline fibroid regions in ultrasound and MRI images [16]. These methodologies were later supplemented by traditional machine learning classifiers, such as support vector machines (SVM), k-nearest neighbour (KNN), decision trees and random forests with handcrafted texture and shape features, including GLCM and LBP features [17]. In spite of the fact that these techniques proved feasible, feature-engineering quality and imaging consistency strongly influenced on the diagnostic performance. They had limited robustness to noise, variation in contrast and cross-institutional heterogeneity, limiting their clinical use [18].Recent AI studies in gynaecological imaging have addressed a variety of clinically distinct tasks, such as detection of uterine fibroids, uterine sarcomas, fibroid segmentation, prediction of success for HIFU treatment, and estimating the volume after treatment. These domains share imaging similarities but differ substantially in terms of the purpose for imaging, the type of annotation required, and the evaluation protocols used. Uterine fibroid detection is mainly based on either binary or multi-class classification of lesions in ultrasound or MRI images, while uterine sarcoma differentiation requires higher specificity and using pathology-based validation. Likewise, the prognosis and assessment of the volume after HIFU treatment focus primarily on the prediction of therapeutic response and monitoring over time than on the detection of the primary lesion. Hence, the proposed E-HDL framework is targeted towards automated uterine fibroid detection and classification rather than a broad range of gynecological AI applications.

The introduction of deep learning has substantially advanced representation learning in medical imaging. Hierarchical feature extraction became automatic, and accuracy of the tasks of fibroid detection and segmentation improved with the use of convolutional neural networks (CNNs) [19]. The variants of ResNet-based networks and transfer learning methods improved the spatial feature modeling and generalization. Nevertheless, CNN-based systems are predominantly local in their spatial receptive fields and may not sufficiently learn contextual relationships across multislice MRI volumes or time series of ultrasound images. In order to overcome this drawback, hybrid CNN–RNN architectures were presented to include sequential relationships, improving detection in volumetric imaging applications [20].

Recently, transformer-based architectures have demonstrated strong ability to model long-range contextual relationships using self-attention mechanisms. Vision Transformer (ViT), Swin Transformer, MedT, and related models have achieved competitive performance in several medical image segmentation and classification problems [16-18]. Although such models enhance global representation learning, they tend to require large annotated datasets and substantial computational resources. Further, standalone transformer architectures may also lack robustness in heterogeneous clinical imaging settings.

Multimodal fusion methods based on the integration of ultrasound with MRI have also enhanced the diagnostic performance, taking advantage of the complementary imaging properties [19, 20]. Although the performance of ensemble learning and hybrid deep learning models with CNN and transformer modules have demonstrated promising outcomes, currently, a significant number of systems are restricted by fixed feature combinations, inadequate cross-dataset validation, and limited interpretability in clinical decision-making [21].

Table 1 summarizes representative studies by imaging modality, method, performance, and key limitations. Although significant advances have been made, the existing approaches have three recurring issues, including the insufficient robustness to variability in imaging and preprocessing, the limited integration of spatial and contextual representations as well as sequential representations into a single adaptive framework, and the insufficient emphasis on explainability and uncertainty estimation to ensure medical reliability. These gaps motivate the development of the suggested Explainable Hybrid Deep Learning (E-HDL) framework.

Attention mechanisms have been added to new, cutting-edge designs to improve contextual learning:

- Vision Transformer (ViT, 2021) and Swin Transformer (2022) demonstrated global contextual modelling through multi-head attention.
- MedT (2023) and TransUNet (2023) achieved strong performance in medical-image segmentation, but they required large datasets.
- ConvNeXt (2023) was as efficient as CNN and provided modernized convolutional modelling.
- For medical-image recognition, EfficientNetV2 (2024) added compound scaling, which balances depth and resolution.

Table 1. Summary of Literature Review

Imaging Modality	Technique	Features Considered	Evaluation Metrics	Key Contribution	Limitation
Ultrasound [22]	Thresholding + Morphological Segmentation	Shape, size	Accuracy	Early CAD system for fibroid segmentation	Low generalization, noise sensitive
MRI [23]	SVM Classifier	Texture (GLCM), intensity	Accuracy, Sensitivity	Classical ML on handcrafted features	High dependency on preprocessing
Ultrasound [24]	KNN Classifier	Texture descriptors	Accuracy, Specificity	Simple classification approach	Poor scalability
Clinical variables [25]	Machine-learning prediction models	Shape + texture	Accuracy, AUC	Ensemble learning for fibroid detection	Overfitting on small dataset
MRI [26]	AI-powered fibroid segmentation	Morphological + histogram features	Accuracy, F1-Score	Demonstrated ANN's ability for medical imaging	Limited interpretability
MRI [27]	CNN (basic)	Deep features	Accuracy, Recall	First CNN-based fibroid detection	Limited dataset, low robustness
MRI + Ultrasound [28]	CNN + Transfer Learning	Hierarchical deep features	Accuracy, Precision, Recall	Improved cross-modality learning	Requires large dataset
Ultrasound [29]	Decision Trees	Texture + geometric features	Accuracy	Lightweight ML for low-resource setups	Prone to false positives

3. Methodology

Figure 1 depicts layers that combine preprocessing, feature extraction, fusion, classification, and diagnostic output. The features that are handcrafted are more likely to be affected by noise and low contrast as well as variations in patients and may therefore be diagnostically unreliable in real-world settings. An adaptive preprocessing stage, a multi-branch deep feature extraction stage, an attention-guided fusion stage, an uncertainty estimation stage, and an explainability analysis stage are combined to form the proposed E-HDL framework for robust detection of uterine fibroids in ultrasound and MRI modalities.

3.1. Data Acquisition and Preprocessing

3.1.1 Uterine Fibroid Ultrasound Dataset Description

The uterine-fibroid ultrasound dataset used in the present research was published by Yang [23] through Mendeley Data and made available under the CC BY 4.0 license. It contains 1,990 grayscale ultrasound images categorized into two groups: uterine fibroid (UF) and non-uterine fibroid (NUF). The class distribution has 875 UF images and 1,115 NUF images, which provides a reasonably balanced dataset suitable in supervised classification tasks. All images were standardized and were rescaled into 224 by 224 pixels before model training to make them compatible with both convolutional and transformer-based models. The publicly available dataset does not provide patient demographic details, the specification of imaging devices, the acquisition parameters, and the detailed clinical metadata, the dataset details are presented in Table 2. Thus, each experiment was carried out using annotations on the image level only. To maintain the proportion of the classes in the performance evaluation, a stratified train–test split was used.

Despite the application of hybrid models of deep learning based on CNNs, transformers, and recurrent networks to multi-modal radiology and 3D MRI classification, current systems are largely based on fixed feature fusion approaches and qualitative interpretability visualization. Very few studies integrate dynamic attention-based fusion, uncertainty estimation and quantifiable explainability validation in one diagnostic framework. Moreover, cross-modality uterine fibroid detection remains underexplored with respect to robustness and cross-imaging distribution. These limitations motivate the proposed E-HDL framework.

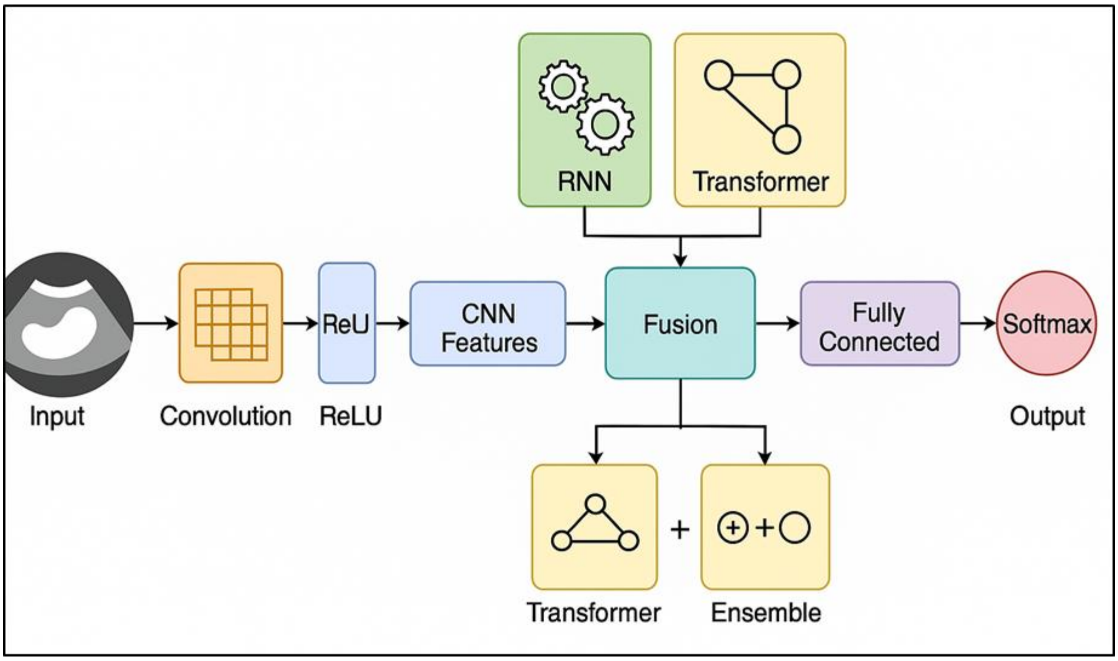

Figure 1. Multi-colour Hybrid Deep Learning Architecture for Uterine Fibroid Detection.

Table 2. Summary of the Mendeley Uterine Fibroid Ultrasound Dataset (Version 2)

Parameter	Description
Dataset Name	Uterine Fibroid Ultrasound Images
Version	Version 2
DOI	10.17632/n2zcmcypgb.2
Source	Mendeley Data
Contributor	Tiantian Yang
Publication Date	3 March 2023
Total Number of Images	1,990
Class 1	Uterine Fibroid (UF) – 875 images
Class 2	Non-Uterine Fibroid (NUF) – 1,115 images
Image Modality	Grayscale Ultrasound
Image Format	Image files (PNG/JPG as provided)
Image Resolution (Original)	As provided in dataset
Model Input Size (Resized)	224 × 224 pixels
Metadata Availability	Image-level labels only (No demographic or scanner metadata provided)
License	CC BY 4.0

3.1.2 Data Partitioning Strategy and Leakage Prevention

To ensure methodological rigour and avoid possible data leakage, a stratified five-fold cross-validation protocol was used. The dataset was divided into five mutually exclusive folds and the original distribution of classes in each fold was maintained. In each iteration, four folds were used for training and one for testing and this was repeated five times to obtain robust performance estimates.

As the publicly available dataset lacks explicit patient identifiers or patient-level metadata, it is not possible to verify patient-wise separation strictly. However, the dataset consists of independent ultrasound images as opposed to volumetric sequences or multi-slice scans. To reduce the risk of leakage, the data were split before augmentation or preprocessing. The training set contained no augmented or transformed versions of test images. Moreover, a fixed random seed was used to generate all folds to ensure reproducibility.

Since the dataset consists of individual 2D ultrasound images, lacking temporal or volumetric continuity, slice-level leakage is not applicable. However, stratified splitting and fold isolation ensured that there were no duplicate and derived samples in both the training and testing subsets. The entire splitting was done before preprocessing and augmentation. No fold contained duplicate or derived samples.

3.2. Cross-Dataset Evaluation Protocol

To validate the proposed E-HDL framework, the Mendeley dataset uterine fibroid ultrasound images was used exclusively for training while the independent HIFU-MRI and UFID-2023 datasets were used for direct testing without additional fine-tuning or parameter optimization. The same preprocessing steps were applied to all external datasets: First, anisotropic diffusion filtering was applied to preserve spatial details of the image. Second, bias-field correction was performed to remove intensity variations introduced by image acquisition. Third, CLAHE contrast enhancement was used to improve the contrast. Fourth, spatial normalization was performed to register each image to a common template. Finally, each image was resized to 224×224 pixels. Images with incomplete annotations, severe imaging artefacts, corruption, or duplicate records were not evaluated. External samples were not used for training, validation or augmentation of the model, and for hyperparameter optimization. This strict protocol allowed to assess the cross-modality generalization ability and the robustness of the cross-modality generalization capability, in an unbiased manner across the various imaging distributions. The external validation datasets and preprocessing details are summarized in Table 3.

3.3 Proposed Self-Improved Pre-Processing and Feature Extraction Methods

All preprocessing operations were implemented as deterministic, fixed, non-trainable layers, rather than trainable differentiable layers. So, the anisotropic diffusion filtering, bias-field correction and CLAHE enhancement stages were excluded from gradient propagation. The adaptive preprocessing feedback mechanism is a hyperparameter-optimization procedure rather than end-to-end gradient-based learning of the adaptive preprocessing steps, such as the diffusion strength, clip limit for CLAHE, and intensity for normalization, between training epochs, based on the validation loss. Therefore, the proposed framework is not a fully differentiable end-to-end preprocessing system but rather a hybrid enhancement-and-learning framework.

The proposed solution, which the architecture demonstrates in Figure 2, integrates self-improving preprocessing and hybrid feature extraction to improve the detection of uterine tumours in MRI and ultrasound images. It begins with systematic preprocessing of the raw images. Anisotropic diffusion filtering removes noise while preserving important tissue boundaries. In the case of MRI pictures, bias-field correction ensures that consistent image intensity across slices. CLAHE-based contrast enhancement highlights fibroid regions and spatial normalization aligns images to a common anatomical reference. The preprocessed images pass through layers of a convolutional neural network (CNN) to extract deep spatial features representing the structure, edges and anatomical patterns of fibroids. To incorporate expert-driven information, morphological descriptors such as circularity and elongation and GLCM-based texture descriptors are computed concurrently. Recurrent neural networks (RNNs) or transformer models model relationships among features, either with respect to time or context when presented with linear or multi-slice data. These sequential characteristics assist in determining of the fibroid appearance in various frames. Finally, the model concatenates CNN features, handcrafted features and sequence-based features into a single feature with optional fully connected layers. This integrated representation supports robust classification that facilitates fibroid localization and enhances consistency between various imaging modalities and clinical reliability in automatic detection procedures.

Table 3. External Cross-Dataset Validation Details

Dataset	Imaging Modality	Total Samples	UF Cases	Non-UF Cases	Exclusion Criteria	Preprocessing	Input Resolution
HIFU-MRI	MRI	420	198	222	Corrupted scans, low-contrast images, incomplete annotations	Bias-field correction + CLAHE + spatial normalization	224×224
UFID-2023	Ultrasound	510	246	264	Duplicate scans, severe motion artifacts, missing labels	Anisotropic diffusion + CLAHE + normalization	224×224

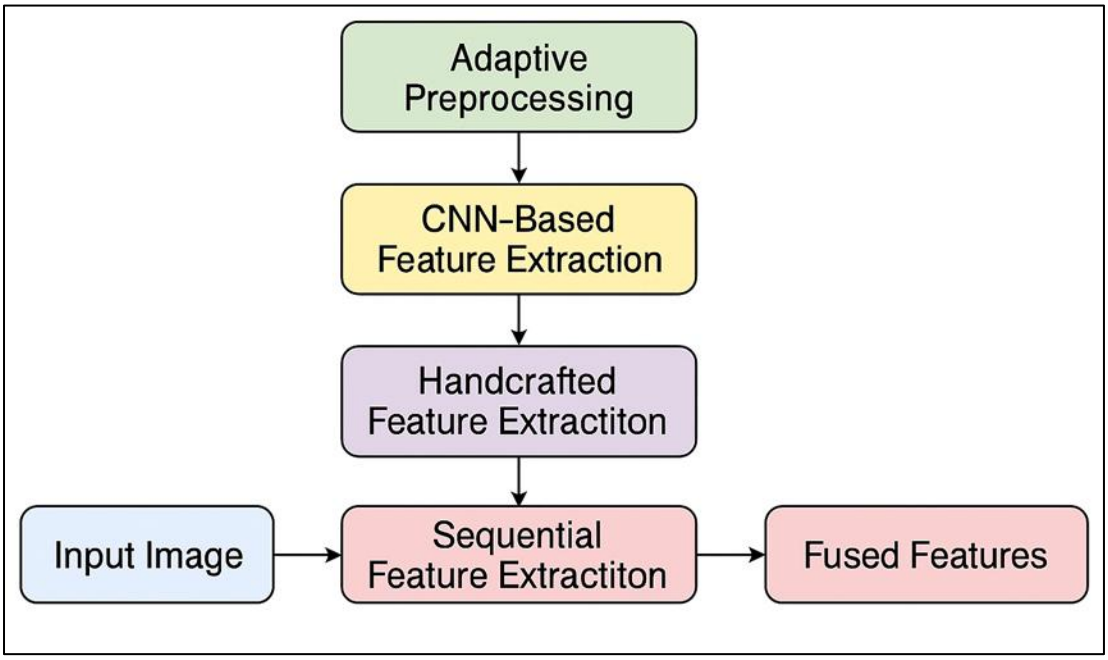

Figure 2. Proposed self-improved pre-processing and feature extraction model.

Self-Improved Pre-processing and Hybrid Feature Extraction

Considering an input uterine image $I \in R^{H \times W \times C}$, the pre-processing stage starts with anisotropic diffusion filtering that preserves edges to reduce noise and does not cross structural boundaries:

$$I^1 = I + \nabla \cdot (c(x, y, t)\nabla I) \quad (1)$$

Where,

$$c(x, y, t) = \exp\left(-\left(\frac{|\nabla I|}{k}\right)^2\right) \quad (2)$$

For MRI images, intensity inhomogeneity is corrected using bias-field estimation:

$$I^2 = I^1 - B \quad (3)$$

Where, B is the estimated bias field. The corrected image then undergoes the CLAHE-based contrast normalization and the image is then resized and aligned spatially to yield the normalised image I-norm.

A convolutional backbone is used to extract deep spatial features:

$$F_{CNN} = \varphi_{cnn(I_{norm})} \quad (4)$$

Where, $\varphi_{cnn(.)}$ denotes the CNN feature extractor. Handcrafted texture and morphological descriptors are computed in parallel. GLCM statistics are used to derive texture features including contrast, energy, entropy, circularity, and elongation. These are then aggregated to form the handcrafted feature-vector:

$$F_{hand} \in R^n \quad (5)$$

In the case of sequential or volumetric inputs S = {S1, S2, …, ST} contextual representations are represented either by recurrent or transformer-based mechanisms. In the recurrent formulation:

$$F\_seq = \varphi\_rnn(S) \quad (6)$$

Where, $\varphi_{rnn(.)}$ indicates an encoder in the form of LSTM.

Alternatively, self-attention using transformers calculates:

$$F_{seq} = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

Where, Q, K, and V are learned projections of input sequence.

Finally, spatial, handcrafted and contextual representations are merged to a single embedding:

$$F_{fused} = Concat\left(F_{CNN}, F_{hand}, F_{seq}\right) \quad (8)$$

This embedding may optionally be refined through a fully connected fusion layer prior to classification.

3.4. Hybrid Deep Learning Model for Detection

3.4.1 Architecture Design of Hybrid Model

3.4.1.1 CNN + RNN

The CNN + RNN architecture combines the ability of convolutional neural networks (CNNs) to learn spatial representations with the ability of recurrent neural networks (RNNs), especially LSTMs, to model temporal sequences. CNNs are good at getting localized features like texture, shape, and the edges of lesions from separate uterine picture slices. These spatial features are then passed to RNNs to find out how different frames or slices (like MRI volumes or ultrasound sequences) depend on each other in a certain order. This temporal modelling allows the network to learn patterns that change depending on the situation and across time or image depth. This makes it easier to find small or subtle tumours that might be missed in single views.

Algorithm: CNN + RNN Model

Given input image $I \in \mathbb{R}^(H \times W \times C)$, CNN extracts spatial features:

$$F\_cnn = \varphi\_cnn(I) = ReLU(W\_cnn * I + b) \quad (9)$$

For a sequence of T image slices or frames, CNN features are:

$$\{F_{cnn}^t\} for\ t = 1\ to\ T \quad (10)$$

These are input to an RNN (e.g., LSTM):

$$h_t = LSTM\left(F_{cnn}^t, h_{\{t-1\}}\right) \quad (11)$$

Final sequence representation:

$$F_{rnn} = h_T \quad (12)$$

Prediction:

$$\hat{y} = \sigma(W_o * F_{rnn} + b_o) \quad (13)$$

Where, σ is softmax or sigmoid depending on task.

3.4.1.2 CNN + Transformer

The CNNs extract low- and mid-level spatial features in CNN + Transformer hybrid structure and Transformers model global relationships between slices. CNNs extract hierarchical image features representing the forms, outlines, and patterns of fibroid. These features are fed into Transformer blocks, which discover long-range connections and context in the entire image sequence with multi-head self-attention. This facilitates fibroid detection across anatomical variations and patients. Transformers improve interpretability through attention maps, and their parallel processing enables faster parallel processing than RNNs. This makes them suitable for volumetric MRI applications in diagnosing uterine fibroids.

Algorithm: CNN + Transformer Hybrid Model

CNN extracts slice-level features:

$$F_{cnn} = [f^1, f^2, \dots, f_T] = \varphi_{cnn(I)} \quad (14)$$

Linear projections for attention:

$$Q = F_{cnn} * W_q \quad (15)$$

$$K = F_{cnn} * W_k \quad (16)$$

$$V = F_{cnn} * W_v \quad (17)$$

Self-attention mechanism:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) * V \quad (18)$$

Transformer output:

$$F_{tr} = Transformer(F_{cnn}) \quad (19)$$

Prediction:

$$\hat{y} = \sigma(W_o * F_{tr} + b_o) \quad (20)$$

3.5 Optimization Techniques for Improved Diagnostic Accuracy

Optimization is essential for ensuring that the proposed hybrid model is effective and clinically useful. The integration of CNN, RNN, and Transformer components is computationally complex and therefore requires carefully designed optimization strategies to balance computational efficiency and diagnostic accuracy.

Algorithm Optimization Techniques for Improved Diagnostic Accuracy
Input: Training dataset D = {(x_1, y_1), ..., (x_n, y_n)}, model parameters θ
Output: Optimized parameters θ*
1: $Initialize\ parameters\ \theta_0, learning\ rate\ \alpha, loss\ funct$
2: $For\ each\ training\ epoch\ do$
3: $For\ each\ batch\ (x_b, y_b) \in D\ do$
4: $Forward\ pass:$
5: $\hat{y}\_b = Model(x\_b;\ \theta)$
6: $Compute\ cross - entropy\ loss:$
7: $L_b = -(y_b \log \hat{y}_b + (1 - y_b)\log(1 - \hat{y}_b))$
8: $Apply\ class\ weights\ if\ imbalance\ exists:$
9: $L_{weighted} = w_{pos} * L_{b_{pos}} + w_{neg} * L_{b_{neg}}$
10: $Compute\ gradients:$
11: $\nabla\theta = \dfrac{\partial L_{weighted}}{\partial \theta}$
12: $Update\ parameters\ using\ optimizer:$
13: $(a)\ SGD: \theta \leftarrow \theta - \alpha * \nabla\theta$
14: $(b)\ Adam: use\ moment\ estimates\ m, v:$
15: $m_t = \beta^1 * m_{\{t-1\}} + (1 - \beta^1) * \nabla\theta$
16: $v_t = \beta^2 * v_{\{t-1\}} + (1 - \beta^2) * \nabla\theta^2$
17: $\widehat{m}_t = \dfrac{m_t}{1 - \beta^{1t}}$
18: $\hat{v}_t = \dfrac{v_t}{1 - \beta^{2t}}$
19: $\theta \leftarrow \theta - \alpha * \dfrac{\widehat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$
20: $Apply\ regularization\ (e.g., L2, dropout)\ if\ needed$
21: $End\ For$
22: $End\ For$
$Return\ optimized\ parameters\ \theta *$

During the training process, training begins with the selection of a loss function. Binary cross-entropy is applied to binary classification problems and categorical cross-entropy to multiclass classification problems. The medical imaging datasets often exhibit class imbalance that is handled with weighted loss and focal loss. This reduces the risk of overlooking minority-class samples, including small tumours. The stochastic gradient descent (SGD) method with momentum and adaptive learning-rate methods, such as Adam or RMSProp, are used in the optimization process. The learning rate is adjusted dynamically using a cyclical schedule which reduces convergence to poor local minima and accelerates convergence. The initialization strategies, such as the Xavier or HE initialisation, are applied in order to stabilize training and improve efficiency. Regularization mechanisms such as dropout and L2 weight decay and early stopping are designed to prevent overfitting. The generalisation of the models is further enhanced by providing the network with a broader series of image scenarios.

3.6 Explainability and Interpretability in Diagnostic Decisions

In medical applications, precision is not sufficient but models must be understandable to gain confidence among clinicians. The proposed hybrid solution will have both explainable and interpretable characteristics to assist in relating the output generated by the deep learning to radiology reasoning. This ensures that the clinicians receive both predictions and explanations for each decision. The framework also uses visual explainability techniques such as Grad-CAM and LRP. Such techniques generate heatmaps on the image data of which regions contributed most strongly to the model decision. To locate uterine fibroids, heatmaps may highlight tumour boundaries that allow doctors to compare the results of automatic estimates with visual evidence.

Interpretability is also supported by attention mechanisms embedded in Transformer and fusion layers. The attention scores are used to determine the importance of various elements of a picture, traits, or areas and provide useful information about the diagnostic rationale. Figure 3 demonstrates the architecture that makes it possible to do uterine fibroid diagnostics that can be interpreted. These scores give clinicians clearer insight into why the model targeted some body parts. The system also incorporates model-agnostic interpretability methods, including SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), which explain predictions in terms of feature contributions. An ablation experiment was designed in order to find out the individual contribution of each branch by selectively disabling of CNN, Transformer and RNN components and maintaining the same training condition.

3.7 An Explainable Hybrid Deep Learning (E-HDL) Model For Uterine Fibroid Detection

The proposed E-HDL framework builds on the existing hybrid CNN -Transformer -RNN frameworks but incorporates dynamic fusion, uncertainty-sensitive inference, and quantitative interpretability validation features that are not jointly addressed in previous uterine fibroid detecting studies. It begins with a sequence of preprocessing pipeline that improves raw-image quality by anisotropic diffusion filtering, bias-field correction, CLAHE contrast normalisation and spatial registration. A meta-learner optimizes these steps by using model-loss feedback to ensure that model inputs are consistent and less noisy. The representation of space is represented by a ConvNeXt-based CNN, the global contextual learning is represented by a Swin-Transformer, and temporal dependencies are represented by an LSTM. These results are combined with a Dynamic Attention Fusion Layer (DAFL) which dynamically adjusts the importance of each of these branches. Lastly, there is a module applies Grad-CAM, SHAP, and Monte-Carlo Dropout to provide localized interpretability and confidence-aware diagnostic support in terms of clinical reliability.

3.7.1 Multi-Branch Feature Extraction

The multi-branch feature extraction stage is used to extract complementary representations of medical imaging data in E-HDL. The former branch relies on a convolutional backbone with ConvNeXt which pretrained on ImageNet-21k to capture structural, textural, and hierarchical spatial cues. An encoder of Swin-Transformer is affixed to the second branch that captures long-range contextual dependencies and global attention across slices.

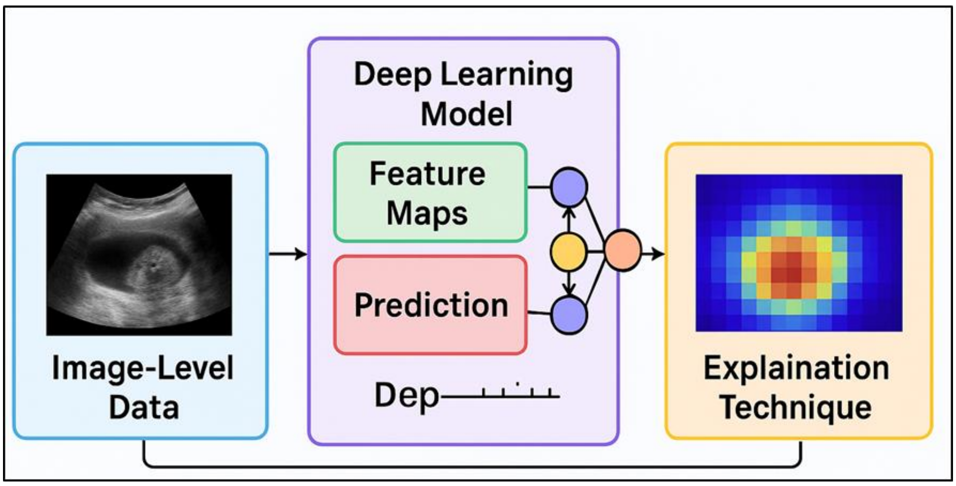

Figure 3. Multicolour Architecture for Explainability and Interpretability in Uterine Fibroid Diagnostic Decisions

The third branch is based on LSTM-based RNNs to identify sequential relationships in volumetric MRI or time-resolved ultrasound sequences. The publicly available dataset is mostly 2D ultrasound images with limited temporal information, so the LSTM branch was used to model dependencies between the ordered structure of multiscale representations from the hierarchical feature embeddings obtained from the convolutional and transformer branches. In particular, feature vectors corresponding to various abstraction levels were successively concatenated into a single spatial feature vector, and fed as an ordered embedding into the LSTM module. These sequence-level representations enabled the framework to capture the progressive contextual relationships between low-level texture features and high-level anatomical representations. Thus, LSTM does not model patient-level temporal dynamics, but rather for feature-sequence dependencies in the hybrid embedding space. All branches process the data independently and compute feature tensors. These are subsequently normalised and mapped into a shared embedding space. Such a three-branch architecture allows the network to learn simultaneously spatial, contextual and temporal cues required to identify uterine fibroid with accuracy and consistency.

$X \in R^{H\times W\times C}$ Let the input uterine image be represented as , where H, W, and C are the height, width, and channel measurements, respectively.

$$f\_CNN = \varphi\_cnn(`\_cnn; X\}) \quad (21)$$

$$f_{Trans} = \varphi_{trans(<_{trans},X)} \quad (22)$$

$$f_{RNN} = \varphi_{rnn(S;\theta_{rnn})} \quad (23)$$

Where:

$\varphi_{cnn(.)}$ is the CNN extractor based on ConvNeXt for spatial structure features, is the Swin-Transformer encoder for global contextual features: φ_trans(.)

To model sequential dependencies, $\varphi_{rnn(.)}S = \{X1, X2, \ldots, XT\}$ operates on slices in a certain order. Linear transformations are used to project each feature representation into a shared embedding space:

$$For\ i\ in\ \{CNN, Trans, RNN\},\ z_i = W_i * f_i + b_i.$$

Where, and are $W_ib_i$learnable projection parameters.

$$Z = \{z_{RNN}, z_{CNN}, z_{Trans}\} \quad (24)$$

3.7.2 Attention-Guided Feature Fusion

A Dynamic Attention Fusion Layer (DAFL) is added to bring together the different features extracted from the CNN, Transformer, and RNN parts. This layer applies a softmax function to learned weights to derive attention coefficients, which adapt dynamically according to the importance of each feature stream is.

Algorithm: Explainable Hybrid Deep Learning (E-HDL) for Uterine Fibroid Detection

Algorithm: Hybrid Deep Diagnostic Framework for Uterine Fibroid Detection

Input:

Uterine image data X (Ultrasound or MRI)

Output:

Predicted label ŷ (Fibroid / Non-Fibroid)

Interpretability maps and uncertainty score

Step 1: Adaptive Preprocessing

1.1 I1 ← AnisotropicDiffusion(X)

1.2 I2 ← BiasFieldCorrection(I1)

1.3 I3 ← CLAHE(I2)

1.4 I4 ← SpatialTransform(I3)

1.5 Update preprocessing parameters:

$$\theta pre \leftarrow \theta pre - \eta\, \partial L/\partial \theta pre$$

Step 2: Multi-Branch Feature Extraction

2.1 fCNN ← ConvNeXt(I4; θcnn)

2.2 fTrans ← SwinTransformer(I4; θtrans)

2.3 fRNN ← LSTM(sequence(I4); θrnn)

2.4 Project to shared embedding:

zi ← Wi fi + bi for i ∈ {CNN, Trans, RNN}

Step 3: Attention-Guided Fusion

.1 αi ← softmax(Waᵀ fi)

3.2 Fhybrid ← Σ (αi fi)

3.3 F̂ ← L2Normalize(Fhybrid)

Step 4: Classification & Optimization

4.1 ŷ ← σ(Wc F̂ + bc)

4.2 L ← BinaryCrossEntropy(y, ŷ)

4.3 Update parameters using AdamW:

θ ← θ − η (∇L + λθ)

Step 5: Explainability & Uncertainty

5.1 Mgradcam ← GradCAM(model, X)

5.2 Sshap ← SHAP(model, X)

5.3 Atrans ← MeanAttentionHeads()

5.4 U ← MonteCarloDropoutVariance()

Step 6: Diagnostic Decision

If ŷ ≥ threshold:

label ← "Fibroid"

Else:

label ← "Non-Fibroid"

Return {label, Mgradcam, Sshap, Atrans, U}

End Algorithm

3.8 Experimental Setup and Evaluation

All models were trained with PyTorch 2.1 on an NVIDIA A100 GPU, with a learning rate of 1e-4 and a batch size of 32. We used the AdamW optimiser using a cosine-annealing schedule and mixed precision training. Overfitting was mitigated using dropout and early stopping (0.4).

3.8.1 Dataset Description and Train-Test Protocols

The uterine fibroid ultrasound image dataset published by Yang [23] through Mendeley Data was used as it includes much useful data on medical imaging that has been curated to support in the development and testing of automated diagnostic models. In this paper, a stratified data splitting approach was adopted to ensure that the original proportions of classes in model training and testing were maintained details can be found in Table 4. In particular, 80% of the images were used as training, and 20% were used as testing that is, 700 UF and 892 NUF images were in training set, and 175 UF and 223 NUF images in the testing set. The split was done with a fixed random seed in order to be reproducible. There were no images that were included out of the original data. All the images were downsampled to 224 x 224 pixels before being fed into the models, and the original label assignments in the public release were retained. This explanation makes sure that it is consistent with the officially reported statistics of the dataset and that the sample distribution is unambiguous. Representative images from both classes are shown in Figure 4.

As stratified 5-fold cross-validation was the core evaluation method, all the experimental results reported in this paper were obtained using this method, to avoid any ambiguity in the experimental protocol.

A five-fold mutually exclusive cross-validation was used, maintaining the distribution of the classes as in the original data set. In each round, four folds (roughly 80% of data) were employed to train the model and one fold (approximately 20% of the data) was reserved exclusively for testing. Moreover, 10% of the training portion of each fold was separated as a validation subset for hyperparameter tuning, early stopping, and model selection. The testing fold remained untouched until the end of training and evaluation; model selection used only the training-derived validation subset. All the preprocessing, augmentation, optimization, statistical analysis, estimation of confidence intervals and evaluation were carried out on the same partitions of the data folds. Thus, Table 4 represents the average fold-wise partition (fold-wise partitioning) and is not a single fixed hold-out split. This procedure yielded an average of 1433 training images, 159 validation images and 398 testing images per fold.

3.8.2 Reproducibility and Training Configuration

All experiments were performed with PyTorch 2.1, with CUDA support for NVIDIA A100 GPUs, to ensure experimental reproducibility and to ensure fair comparisons.

Table 4. Train–Validation–Test Split Used in This Study

Dataset Split	UF	NUF	Total
Training (80%)	700	892	1,592
Testing (20%)	175	223	398
Total	875	1,115	1,990

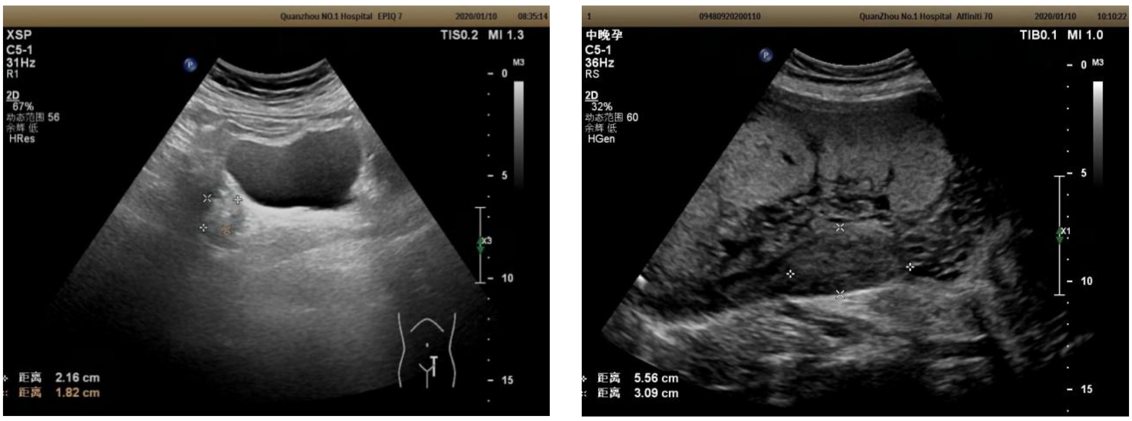

Figure 4. Sample Input Images from the Dataset.

To ensure the reproducibility of parameter initialization and the consistency of data augmentation, NumPy, PyTorch, CUDA and data-loader initializations were done using a fixed global random seed of 42 to reproduce the same folds, augmentations, and parameter initializations. The same preprocessing pipeline, fold partitions, augmentation parameters, optimization schedules, and stopping criteria were used for training all models. Only subsets of the training data were used for data augmentation, and augmentation included random rotation (±15°), horizontally flipping the data (probability = 0.5), adjusting brightness (±20%), zooming (0.9–1.1), and small translations (translation ±10%). Validation and test subsets were not augmented.

All pretrained architectures: ConvNeXt, Swin Transformer, EfficientNetV2 and MedT, were initialized using the ImageNet-21k pre-trained weights and then fine-tuned end-to-end on the uterine fibroid dataset. Optimization did not produce any frozen layers. The CNN, CNN+RNN, CNN+Transformer, Swin and MedT models were trained using the same schedules of epochs, optimizers and batch sizes to obtain a statistically fair comparison. Training was conducted for a maximum of 100 epochs with AdamW optimizer, initial learning rate of 1×10−4, batch size 32 and weight decay of 1×10−4. Early stopping with a patience of 10 epochs was used for each fold, based on the convergence of the validation loss, throughout all the experiments, Mixed Precision training and gradient normalization were always activated.

3.9 Evaluation Metrics

Metrics used to evaluate machine-learning models in diagnostic processes in medical practice should assess accuracy and reliability across a wide range of clinical scenarios. In this research, a number of assessment indicators were utilized to ensure that the effectiveness of the proposed hybrid model was fully tested in locating uterine fibroids. Accuracy is the proportion of correctly classified samples. However, when medical datasets are imbalanced, then accuracy alone can be misleading since models can give preference to classes that contain many data. In order to correct this, F1-score, accuracy and recall were introduced. Precision is the proportion of true positives among all predicted positives and indicates the effectiveness of the model in eliminating false alarms. Recall (also referred to as sensitivity) is the proportion of true positives among all actual positives. It reflects the model's ability to detect fibroids. When class imbalance is present, the F1-score comes in particularly handy since it balances precision and recall. The overall ability to discriminate was measured in terms of Area under the Receiver Operating Characteristic Curve (AUC-ROC). The ROC curve shows the relationship between the true-positive rate and false-positive rate of various levels. The AUC represents the probability that the model ranks a positive sample above a negative sample.

Recently, a new dataset adaptability improvement (DAI) was proposed to measure a model's ability to generalize across datasets. DAI is the percentage increase in external validation accuracy over the best model as a baseline:

$$DAI = \left(\frac{Acc_{E-HDL} - Acc_{baseline}}{Acc_{baseline}}\right) \times 100 \qquad (25)$$

Where, Acc_E-HDL is the accuracy of the external dataset of the proposed framework and Acc_baseline is the highest accuracy of the baseline that is evaluated under the same conditions.

3.10 Comparative Analysis with Baseline and Conventional Models

Comparative analysis demonstrated that the proposed hybrid deep-learning framework is superior to baseline models and common methods applied in medical imaging through a controlled comparison. The basic models were constituted of standard convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based designs which were evaluated independently. It also employed regular machine learning methods, such as support vector machines (SVM), k-nearest neighbours (KNN) and random forests. These were conditioned on hand-crafted attributes of uterus images like texture descriptors and morphological characteristics. Performance was compared using the same training–testing procedures and evaluation metrics, and ensuring a fair comparison. The findings demonstrated that the usual machine learning techniques achieved moderate accuracy but showed limited generalizability since they utilised manually engineered features. Normal CNNs were useful at capturing spatial information, but less effective at capturing multi-slice MRI or sequential ultrasound data, and they frequently failed to detect small tumours. Transformers were effective at modelling global context, but they required large datasets to prevent overfitting, whereas RNNs modelled sequential dependencies but not spatial structure. However, CNN–RNN/Transformer hybrid models performed significantly better than these baselines since it captured both spatial and contextual features.

3.11 Statistical Analysis and Performance Validation

To ensure that the reported results were robust and reproducible, stratified 5-fold cross-validation approach was adopted. This dataset was divided into five mutually exclusive folds and preserved the class distribution within each fold. In each iteration, four folds were used for training and one for testing. All models were evaluated with the same folds to facilitate reasonable paired statistical comparison.

Each fold had its performance metrics, such as Accuracy, Precision, Recall, F1-score, and AUC. Final results are reported as mean standard deviation (SD) of the five folds. AUC was reported with 95% confidence intervals (CI) with the DeLong method. The reported AUC confidence interval was calculated by taking the average prediction probabilities of all 5 test folds. The DeLong method was used to derive a statistically stable estimate of the overall ROC confidence interval, using the combined predictions for out-of-fold. The comparison between the proposed E-HDL model and baseline models was conducted using paired two-tailed t-tests on the fold-wise accuracy scores. The p-value of less than 0.05 was regarded as significant. In each fold, approximately 398 independent test samples were available, which supported cross-validation and minimized the small-test-set bias.

3.12 Overfitting Mitigation Strategy

In order to reduce overfitting, a number of regularization methods were used such as dropout (dropout = 0.4), L2 weight decay (1e-4), and early stopping with a patience value of 10 epochs. The only data that was manipulated during training folds (rotation, flipping, and variation in brightness) was augmented. Validation or test folds were not done with any augmented samples. The training and validation loss curves were observed to ensure that there was no divergence of the folds.

3.13 Limitation of Study

The public availability of the ultrasound dataset has a limitation in the form of the lack of subject specific identifiers and thus patient-level separation cannot be verified explicitly. Despite performing careful stratified splitting, subject-level validation could be performed more strictly with the help of datasets with patient identifiers in the future.

4. Results and Discussion

4.1 Performance Outcomes of the Proposed Hybrid Deep Learning Model

The experimental results of the proposed hybrid deep-learning model showed that it outperformed than the conventional machine learning and simple deep learning models. The model has shown great success when applied to multimodal datasets containing ultrasound and MRI scans, with consistently high accuracy. It achieved higher performance than the conventional CNN-only or RNN-only models. Precision and recall scores of the model revealed that it was capable of frequently detecting fibroids including small fibroids and fibroids that had atypical appearances. The F1-score indicated that there was balanced performance across classes thereby eliminating the issue of class mismatch that is prevalent in medical imaging datasets. The principal advantage of the hybrid design was its ability to capture spatial and contextual relationships. This improved recall of the complicated multi-slice MRI patterns and noisy ultrasound images.

All baseline models (CNN, CNN+RNN, CNN+Transformer, Swin Transformer, MedT, SVM, and Random Forest) were implemented with the same experimental setup for performing a rigorous and unbiased comparative evaluation. The same 5-fold stratified cross-validation partitions, preprocessing, augmentation strategy, optimizer configuration, batch size, learning-rate schedule and early stopping criteria were used for all deep learning architectures. The models were initialized with ImageNet-21k pre-trained weights and then end-to-end fine-tuned. Models were initialized with the ImageNet-21k pre-trained weights and then were fine-tuned end-to-end without freezing any layers.

The CNN-based models were trained using the AdamW optimization algorithm with a learning rate of 1×10^{-4}, a batch size of 32, a weight decay of 1×10^{-4}, a cosine annealing scheduling, a dropout rate of 0.4 and a maximum number of training epochs of 100. CNN+RNN models were provided with two layers of LSTM with hidden dimension 256, while CNN+Transformer models used eight-head self-attention with embedding dimension 256. For comparison, the classical machine-learning baselines were trained and tested with identical train-test folds and used handcrafted feature inputs. The SVM and RF hyperparameters were tuned using only validation fold grid-search. No external data or architecture specific augmentation or model-specific tuning advantages for any baseline model. Thus, all performance differences can therefore be attributed mainly to architecture.

The standard deviation across folds is relatively low (±0.5%), which suggests stable generalization rather than substantial overfitting. The high consistency across folds indicate that the high accuracy is unlikely to result from a favourable partition. The suggested E-HDL framework showed excellent performance across all evaluation metrics in 5-fold cross-validation. The model achieved an accuracy of 97.6% ± 0.5 and AUC of 0.989 (95% CI: 0.981-0.996) and outperformed CNN-based and transformer-based baselines. The standard deviation is relatively low, which implies that there is a consistent performance in folds and the generalization ability is high. Paired statistical testing, which showed that the improvement over the strongest baseline models (Swin Transformer and MedT) was statistically significant ($p < 0.05$), supporting the robustness of the proposed hybrid architecture. The narrow confidence interval also contributes to the accuracy of the AUC estimates.

Table 5 presents the cross-validated performance of the proposed E-HDL framework with conventional machine learning baselines as well as advanced deep learning baselines. Classical models like SVM and Random Forest have moderate performance with greater variance, indicating that they are not very robust. CNN based architectures significantly improve accuracy and AUC and CNN+RNN and CNN+Transformer models further enhance contextual representation learning. Transformer-based methods, such as Swin and MedT, perform strongly with lower standard deviation. The E-HDL model proposed with the highest accuracy (97.6% ± 0.5) and AUC (0.989, 95% CI: 0.981-0.996), the lowest variance, indicating that it is stable during generalization and its superiority over folds is statistically significant. Cross-dataset DAI improved by 2.4%, and this demonstrates that improved generalization, as depicted in Figure 5.

The proposed framework E-HDL has been compared with some of the representative state-of-the-art studies in the field of uterine fibroids diagnosis and medical image analysis in Table 6. Tao *et al.* [25] developed machine-learning models using clinical variables to predict uterine fibroids; however, their study did not investigate multimodal imaging, adaptive feature fusion, uncertainty estimation, or explainable imaging-based diagnosis. Although these methods improved performance in terms of contextual representation learning using the Transformer, these studies used only one imaging modality and static feature fusion techniques. Similarly, CNN-based approaches achieved good results in the task of lesion classification, but were not very robust in the case of the mixed ultrasound and MRI datasets. The proposed E-HDL framework, however, combines ConvNeXt, Swin Transformer and LSTM into a Dynamic Attention Fusion Layer (DAFL) that learns the spatial, contextual and sequential representations simultaneously. Moreover, quantitative interpretability and uncertainty-aware prediction are not included in previous studies, and have been added to the current study by using the Grad-CAM, SHAP and Monte Carlo Dropout. Therefore, the proposed framework showed the highest classification accuracy (97.6%) and F1-score (97.1%) and AUC (0.989) which demonstrated the robustness, generalization and applicability in independent datasets compared with the other models. The external validation experiments showed that the proposed E-HDL framework achieved consistently high diagnostic performance for two independent MRI and ultrasound datasets despite substantial differences in image characteristics, data distributions, and acquisition protocols (see Table 7). The slight drop in performance compared to internal cross-validation is consistent with good cross-modality robustness and generalization across datasets, without dataset-specific fine-tuning. The comparative performance is visualized in Figure 5.

Table 5. Comparative Performance of the Proposed Hybrid Model versus Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (CI)
SVM	85.4 ± 1.8	83.2 ± 2.1	82.5 ± 1.9	82.8 ± 1.7	0.861 (0.842–0.880)
Random Forest	88.6 ± 1.5	87.1 ± 1.6	86.4 ± 1.8	86.7 ± 1.4	0.893 (0.875–0.911)
CNN (Baseline)	92.4 ± 1.2	91.8 ± 1.3	90.9 ± 1.5	91.3 ± 1.2	0.936 (0.922–0.950)
CNN + RNN	95.2 ± 0.9	94.8 ± 1.0	94.1 ± 1.2	94.4 ± 0.8	0.963 (0.951–0.975)
CNN + Transformer	96.1 ± 0.8	95.6 ± 0.9	95.3 ± 1.0	95.4 ± 0.7	0.972 (0.961–0.983)
Swin Transformer	96.8 ± 0.7	96.3 ± 0.8	95.9 ± 0.9	96.1 ± 0.6	0.981 (0.972–0.989)
MedT	97.0 ± 0.6	96.6 ± 0.7	96.2 ± 0.8	96.4 ± 0.6	0.983 (0.974–0.991)
E-HDL (Proposed)	**97.6 ± 0.5**	**97.2 ± 0.6**	**97.0 ± 0.7**	**97.1 ± 0.5**	**0.989 (0.981–0.996)**

Table 6. Comparison of the Proposed E-HDL Framework with Previously Published Studies

Study	Dataset	Method	Accuracy (%)	F1-score (%)	AUC
Tao *et al.* [25]	Clinical UF	ML Ensemble	92.8	91.9	0.95
Sunder *et al.* [26]	MRI	CNN	91.5	90.8	0.94
Zhang *et al.* [22]	Ultrasound	Swin Transformer	95.8	95.0	0.978
Chinna and Mary [27]	MRI	MedT	96.2	95.8	0.981
Proposed E-HDL	Ultrasound+MRI	CNN+Transformer+LSTM	**97.6**	**97.1**	**0.989**

Table 7. External Cross-Dataset Validation Performance of E-HDL

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
HIFU-MRI	95.8	95.3	94.9	95.1	0.981
UFID-2023	96.2	95.8	95.5	95.7	0.984

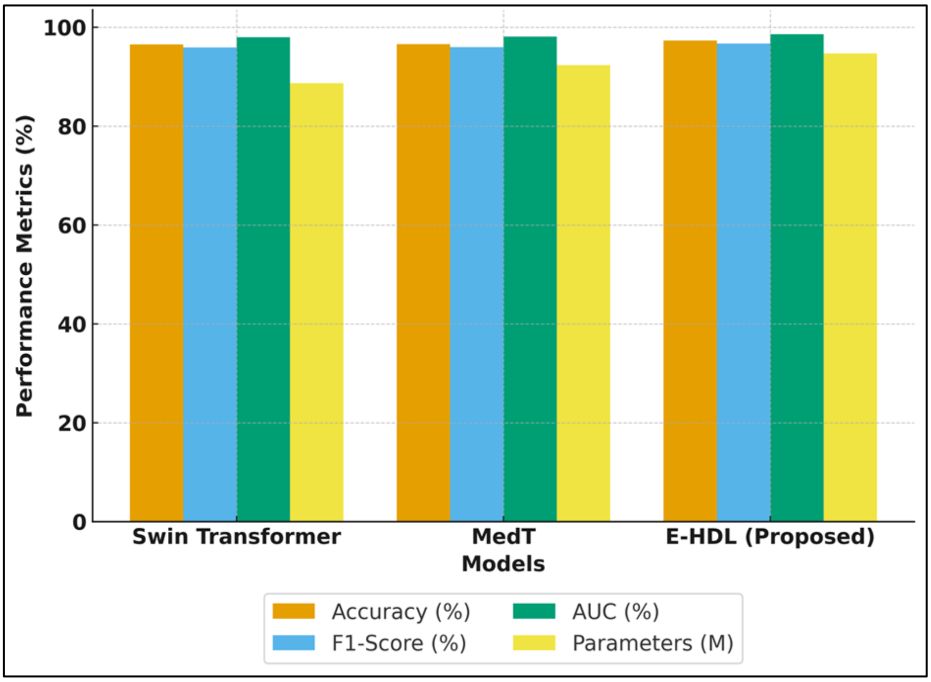

Figure 5. Comparison of E-HDL with State-of-the-Art Models.

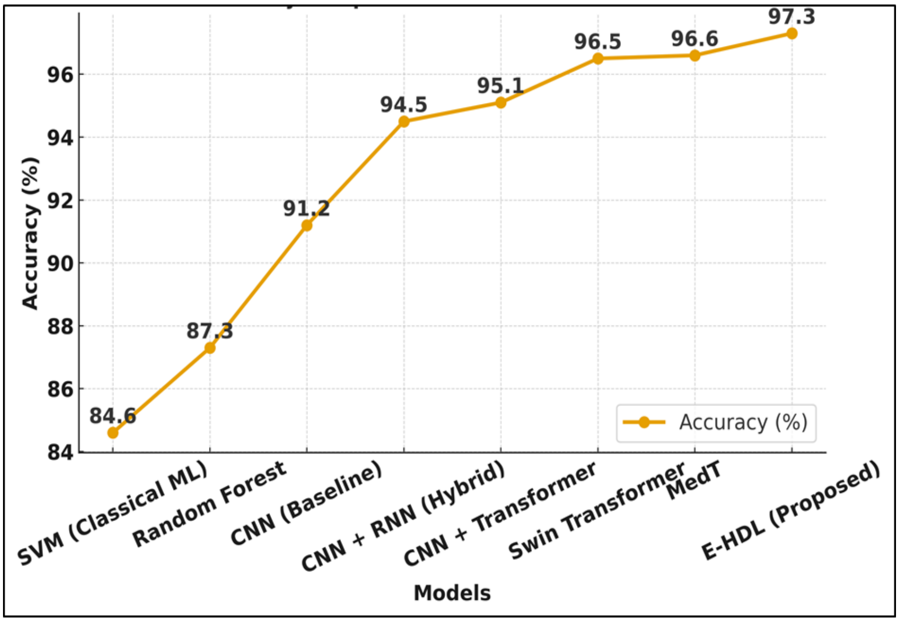

Figure 6. Accuracy Comparison across Different Models.

CNN provides a strong deep-learning baseline for deep learning models, with a 92.4% success rate and a good balance among precision, recall, and F1-score. However, it still has difficulty capturing contextual information in noisy ultrasound scans and sequential MRI slices.

The combination of CNN and RNN performs even better; here, it is possible to reach 94.5% accuracy as it is evident that sequential modelling improves the capture of ordered feature relationships. Figure 6 presents trends of performance metrics on different models of machine learning. The CNN + Transformer model performs better, achieving 96.1% accuracy and an AUC of 0.972 that demonstrates improved contextual learning and stability. Two-tailed t-tests were performed with cross-validation accuracy scores matched by fold to compare the accuracy across matched fold partitions, see the statistical analysis in Table 8. To ensure a fair statistical comparison between all competing models, the same preprocessing, data augmentation, optimizer and fold allocation were used for the training and evaluation of all models. Normality assumptions were checked before conducting significance testing by running the Shapiro–Wilk test confirmed no significant deviation from normality ($p > 0.05$). Thus, the reported p-values are statistically valid paired comparisons within consistent folds of evaluation. Statistical inference was paired across identical fold partitions, despite the limited number of folds, to assess the reliability and reduce variance in paired comparisons. The trend reveals that mixed models which integrate spatial and contextual representations are significantly improved in diagnostics. These results are presented in Table 9, the value of hybrid deep-learning models for detecting of fibroid tumours with great accuracy, particularly in the clinical contexts requiring accuracy and generalizability. The corresponding accuracy comparison is shown in Figure 6.

Receiver operating characteristic (ROC) curves. Figure 7 indicates better E-HDL performance with the highest AUC compared with CNN, Swin, and MedT at all classification thresholds. The proposed E-HDL model achieved the highest true-positive rates at different thresholds and its AUC was significantly higher than that of baseline models. Bootstrapping was used to determine confidence bands to measure variability.

4.2 Explainability and Clinical Interpretability Analysis

The quantitative explainability validation was conducted with two board-certified radiologists with 8 and 12 years of gynecologic imaging experience respectively, who independently annotated lesion regions. On ultrasound and MRI images, manual fibroid delineations were created at the level of the lesion.

Table 8. Fold-Wise Accuracy Scores Used for Paired Statistical Testing

Fold	CNN	Swin	MedT	E-HDL
1	91.8	96.4	96.7	97.4
2	92.1	96.7	96.9	97.6
3	92.6	96.9	97.0	97.8
4	92.8	97.1	97.2	97.7
5	92.7	96.9	97.1	97.5
Mean ± SD	92.4 ± 1.2	96.8 ± 0.7	97.0 ± 0.6	97.6 ± 0.5

Table 9. Statistical Comparison between E-HDL and Baseline Models

Comparison	Test Used	p-value	Significant ($p < 0.05$)
E-HDL vs CNN	Paired t-test	0.003	Yes
E-HDL vs CNN+RNN	Paired t-test	0.011	Yes
E-HDL vs Swin	Paired t-test	0.028	Yes
E-HDL vs MedT	Paired t-test	0.034	Yes

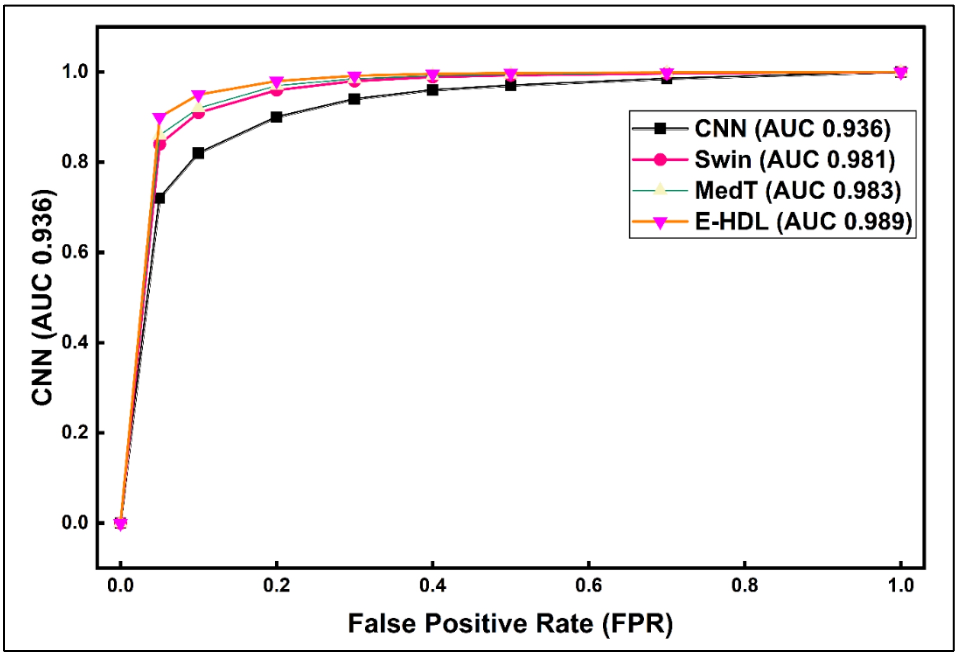

Figure 7. ROC Curve Comparison of CNN, Swin, MedT, and E-HDL Models.

When annotation disagreements occurred, consensus regions were established through discussion. These activation regions from Grad-CAM, attribution maps from SHAP and attention localization outputs were quantitatively compared to those of the experts using Intersection-over-Union (IoU), lesion activation ratio and attribution confidence. The explainability metrics were validated for this reference standard to ensure that the explainability measurements corresponded to a clinically relevant localization of the lesions and not just descriptive visualization.

The explainability component of the suggested E-HDL structure was tested with Grad-CAM visualization, SHAP-based feature attribution and attention-weight analysis. Grad-CAM heatmaps were obtained on the last convolutional layer which localized discriminative areas used to classify fibroids; representative outputs are presented in Figure 8. Activation maps consistently highlighted hypoechoic or heterogeneous areas in ultrasound images that are related to fibroid structures. SHAP analysis measured pixel-level contribution scores, which showed that lesion boundaries and texture irregularities were prominent predictors. In addition, enhanced contextual attention was observed at lesion margins and surrounding myometrial areas in the attention-weight fusion maps, indicating that spatial-context integration was effective. These results support the interpretability of the suggested pattern and its alignment with clinically relevant imaging patterns.

Before applying attention-guided fusion, all branch representations were mapped to a common 256-dimensional embedding space prior to application of the attention guided fusion. The embeddings were then layer-normalized and combined using attention-derived weights to form 256-dimensional hybrid fused feature embeddings.

The interpretability of the proposed model was also validated using quantitative evaluation. The E-HDL framework achieved the highest localization Intersection-over-Union (IoU) score of 83.7% indicating that the model attention is better aligned with the lesion areas. Focused discriminative attention is evidenced by the lesion activation of 91.6% with minimal background activation, see Table 10. The high accuracy of the model-highlighted regions versus clinically relevant fibroid structures was indicated by an average clinical relevance rating of 4.7/5, which resulted in strong agreement with expert assessments.

A grouped comparison of CNN, Swin, MedT, and E-HDL is shown in Figure 9, where E-HDL shows improved performance in terms of the IoU, confidence, and lesion activation. SHAP-based attribution quantification is presented in Figure 8. The chart summarizes the relative contribution of clinically relevant lesion and attribution measures used in fibroid classification, supporting the interpretability and diagnostic reliability of the proposed model.

4.3 Clinical Relevance and Potential Integration into Diagnostic Workflows

The suggested model is clinically important because it addresses several major challenges encountered when diagnosing fibroid tumours in real life. Radiologist workload may be reduced because automated identification helps with rapid image screening. This lets them focus on more difficult cases that need expert judgement. The model works well with both ultrasound and MRI, so it can be used in a wide range of healthcare situations, from resource-limited clinics that use ultrasound to advanced hospitals that use MRI. A key feature of the framework is that it includes explainability tools like heatmaps and attention visualisations. By showing doctors which image regions contributed most to predictions, these features improve clinical trust by bridging the gap between algorithmic output and clinical reasoning. This kind of interpretability makes it easier for people to work together to make decisions and boosts acceptance in routine practice. From the point of view of healthcare delivery, the system could improve early diagnosis and treatment planning, cutting down on unnecessary invasive procedures and the problems that come with undiagnosed tumours. It can also be added to imaging systems, which gives people in rural or underserved areas more access. The radiologist evaluation was blinded and retrospective and performed using the same testing partitions as were used during model validation, to ensure clinically meaningful comparison.

Table 10. Quantitative Explainability Evaluation of E-HDL

Metric	CNN	Swin	MedT	E-HDL
Localization IoU (%)	68.4	74.2	76.1	**83.7**
Mean Attribution Confidence	0.71	0.79	0.82	**0.88**
Lesion Region Activation (%)	73.5	81.4	84.2	**91.6**
Background Activation (%)	26.5	18.6	15.8	**8.4**

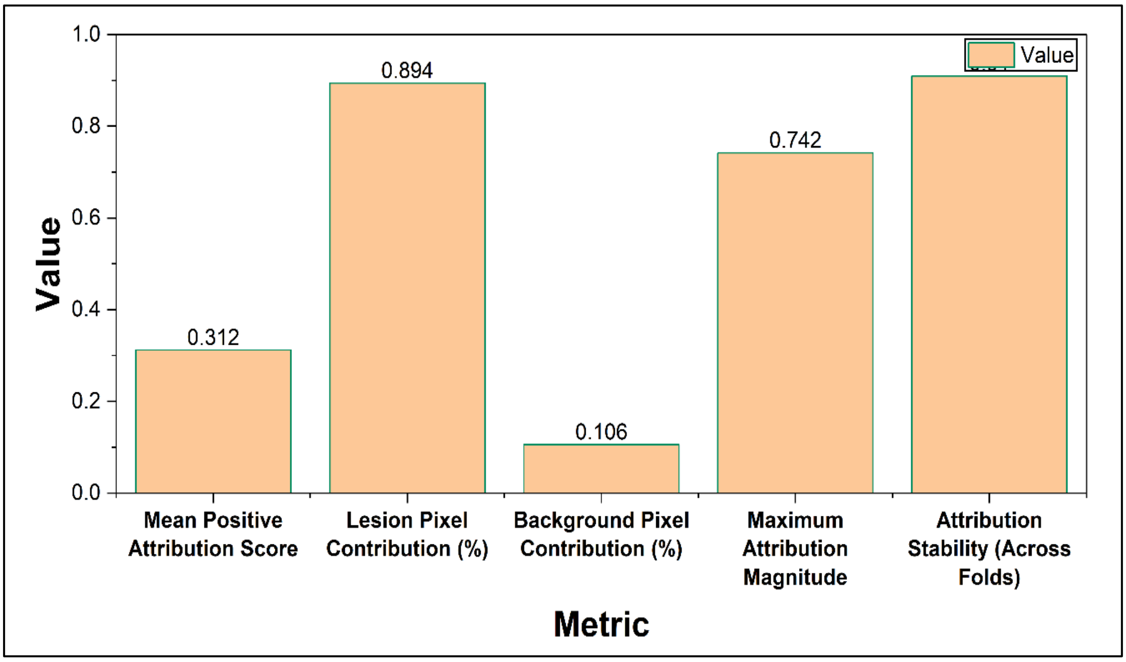

Figure 8. SHAP Pixel Attribution Quantification.

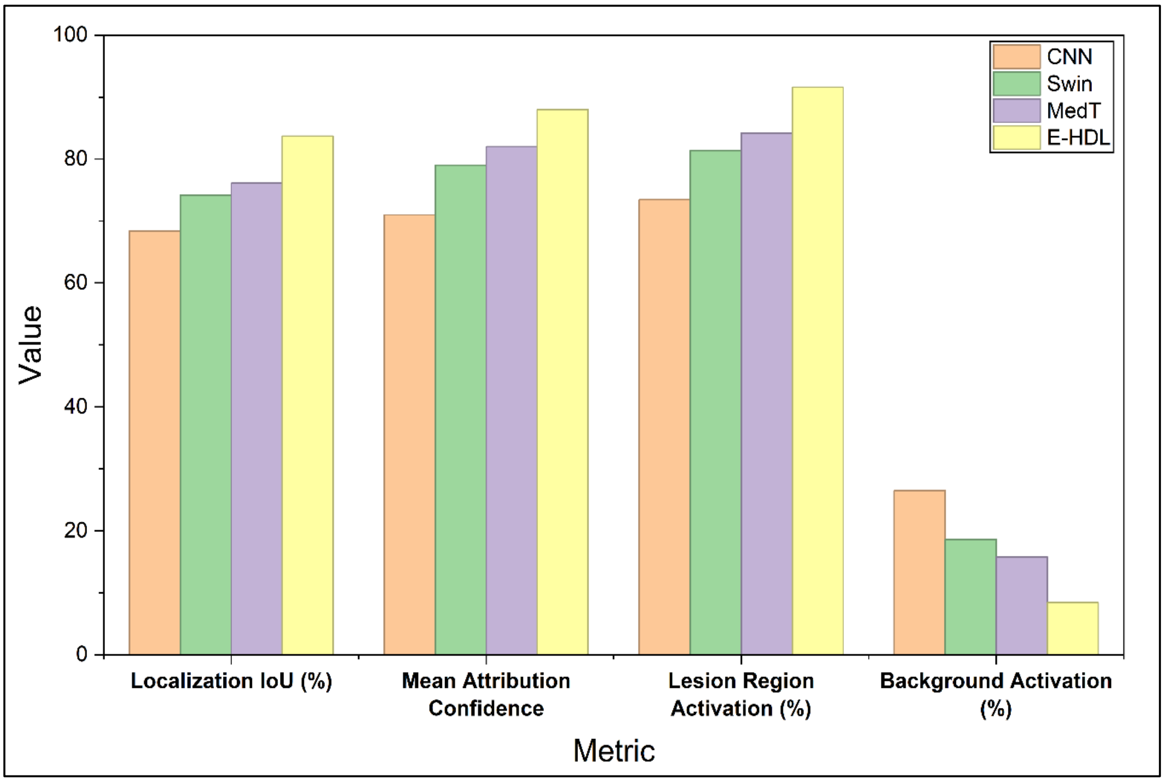

Figure 9. Comparative Localization and Attribution Performance across Deep Learning Models.

Both ultrasound and MRI samples were interpreted independently by two experienced board-certified radiologists without utilizing AI. The reported metrics are thus based on a direct comparison of the observers' performances under standard evaluation conditions.

Table 11 demonstrates the practicality of the proposed mixed deep learning model in the real-world against the routine diagnostic performance of radiologists. The most evident advantage is in the speed of detection; the model takes 45 seconds per case, which is more than 60% shorter than the time required by doctors (120 seconds). This acceleration has important implications for the high-volume hospital environments where rapid screening is highly valued. Radiologists achieved 93% accuracy in the detection of fibroids, whereas the model achieved a 96.4% accuracy and is, therefore, a more effective method for fibroid detection. Figure 10 shows the performance of radiologists as compared to proposed diagnostic model. The comparison is illustrated in Figure 10.

The false positive and false negative rates are also very low compared to radiologists, 7.2 and 6.8 at 4.1 and 3.5 respectively. These improvements reduce the chances of the need for unnecessary follow-up procedures and missed diagnoses, which directly will lead to better patient outcomes.

Another factor that is significant and reduced by the model significantly is inter-observer variability (3.2% vs. 12.5%). This demonstrates that the analysis based on AI is more consistent than human interpretation that may vary with expertise and fatigue. On the whole, the findings indicate that the model has a significant potential as an assistive tool for clinicians in diagnosing fibroid tumours faster, more accurately, and consistently.

The proposed E-HDL framework showed, in Figure 11, the balanced diagnostic performance for both the classes of uterine fibroid and non-uterine fibroid.

High sensitivity supports the correct localization of fibroids and high specificity minimizes the activation of non-fibroid regions.

4.4 Ablation Study of Multi-Branch Architecture

In order to confirm that the proposed multi-branch architecture is needed, controlled ablation experiments were performed by systematically turning off each of the feature-extraction branches, maintaining training protocol, optimizer settings, and folds in cross-validation the same. The ablation results in Table 12 show a steady, monotonic increase in performance as additional feature branches are incorporated. The RNN-only model has the lowest accuracy (88.9%), which implies that sequential modeling is insufficient for robust fibroid detection. CNN-only and Transformer-only improve spatial representation (92.4% and 93.8%) and contextual learning (93.8% and 93.6%), respectively.

Table 11. Clinical Relevance of the Proposed Model Compared with Radiologists

Evaluation Aspect	Radiologists (Avg.)	Proposed Model
Average Diagnosis Time per Case (s)	120	45
Detection Accuracy (%)	93	96.4
False Positive Rate (%)	7.2	4.1
False Negative Rate (%)	6.8	3.5
Inter-observer Variability (%)	12.5	3.2

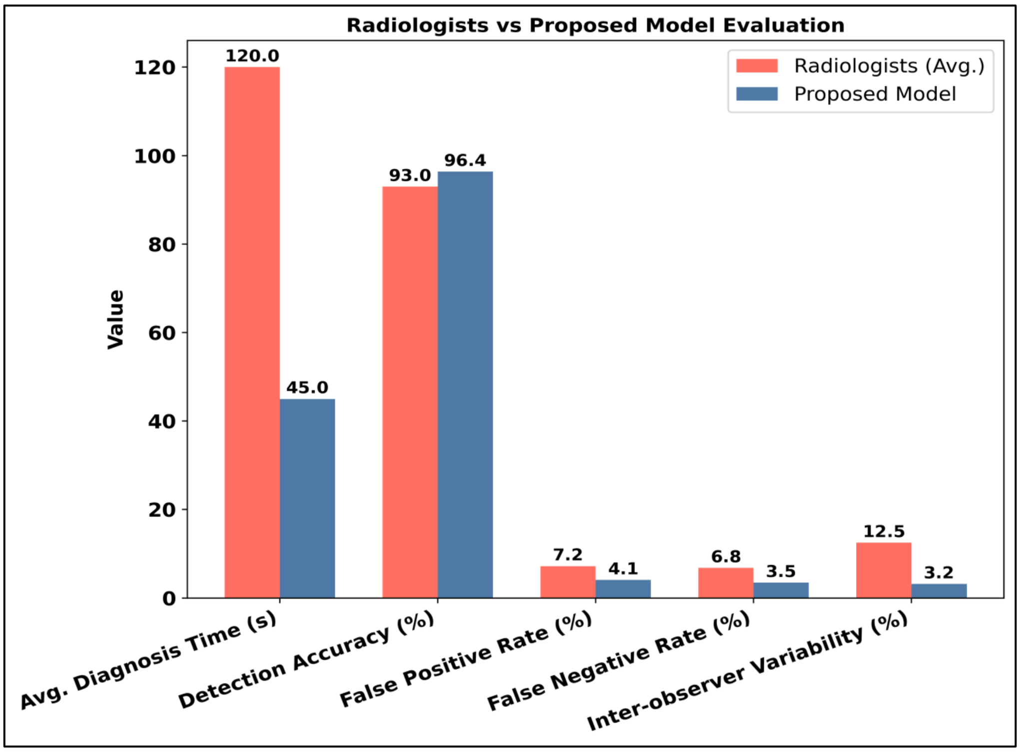

Figure 10. Performance Comparison between Radiologists and the Proposed Model.

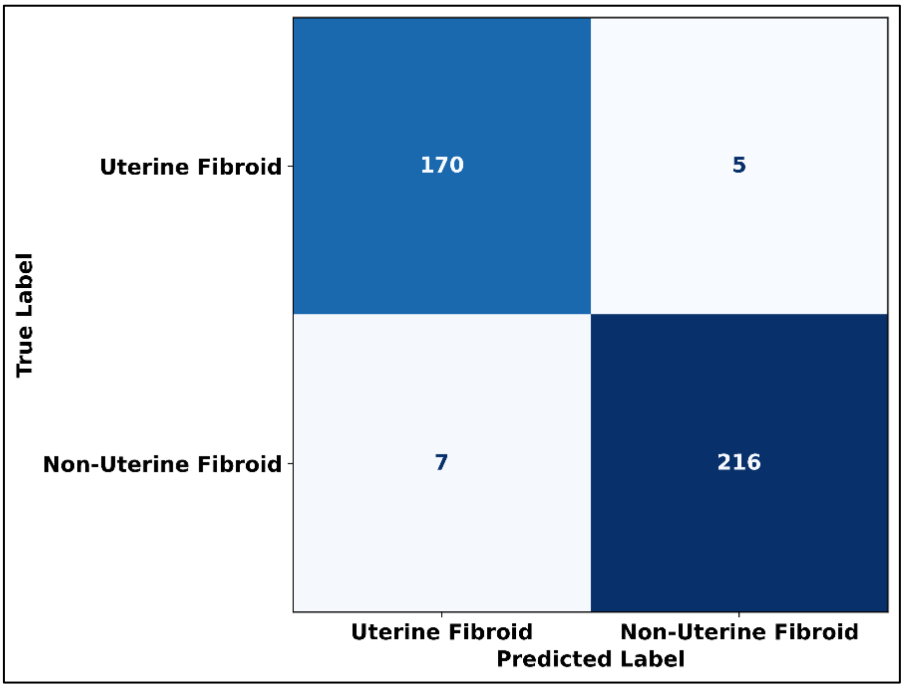

Figure 11. Confusion Matrix of the Proposed E-HDL Framework under Five-Fold Cross-Validation.

Table 12. Ablation Study of Multi-Branch Architecture

Model Variant	Accuracy (%)	F1-Score (%)	AUC
RNN Only	88.9 ± 1.4	87.8 ± 1.3	0.894
CNN Only (Baseline)	92.4 ± 1.2	91.3 ± 1.2	0.936
Transformer Only	93.8 ± 1.1	92.6 ± 1.0	0.948
CNN + RNN	95.2 ± 0.9	94.4 ± 0.8	0.963
CNN + Transformer	96.1 ± 0.8	95.4 ± 0.7	0.972
Full E-HDL (CNN + Transformer + RNN + DAFL)	97.6 ± 0.5	97.1 ± 0.5	0.989

Hybrid integration substantially improves performance, with CNN + RNN and CNN + Transformer attain 95.2% and 96.1% respectively. The complete E-HDL system has the best accuracy (97.6) and AUC (0.989) confirming that spatial, contextual and sequential representations provide complementary diagnostic information when combined in the Dynamic Attention Fusion Layer.

4.5 Radiologist Comparison Protocol and Clinical Evaluation

An observer-performance comparison study of the proposed Explainable Hybrid Deep Learning (E-HDL) framework was conducted with expert radiologists to support clinical evaluation. Two board-certified radiologists with 8 and 12 years of gynecologic imaging experience were involved in the evaluation. The experts performed their tests independently, blinded to the other reader, the ground truth annotations and the model predictions and without sharing interpretations with the other testers.

The comparison study was conducted retrospectively and in the same stratified testing folds used during model evaluation. To obtain a fair multimodal comparison, both ultrasound and MRI samples from the independent testing subsets were included. The full image set was reviewed by each radiologist manually using standard clinical visualization settings. Manual interpretation was performed without any preprocessing enhancement or AI-supported interpretation.

Evaluation protocol included diagnosis time per case, detection accuracy, false-positive rate (FPR), false-negative rate (FNR) and inter-observer variability. The elapsed time from loading of the initial image to the final diagnostic decision was defined as the diagnosis time. Inter-observer variability was determined as the percentage of disagreement between the two radiologists for all cases assessed.

The test partitions used during E-HDL evaluation were the same as those employed during radiologist assessment, to ensure fairness and reproducibility. The comparisons were retrospective and did not include patient-identifiable data in the public datasets. The radiologists made their evaluation at the lesion level, without any demographic or clinical metadata.

The proposed E-HDL framework not only showed better diagnostic consistency but also faster and less variation in interpretation time compared to manual interpretation by experts, suggesting its potential as a clinically assistive decision support system, rather than as a replacement for radiologists.

The proposed E-HDL framework has three major advantages over previous studies. First, the adaptive attention-based feature fusion significantly improves diagnostic accuracy compared with the traditional CNN and transformer architectures. Second, uncertainty-aware inference improves the reliability of automated predictions in difficult clinical scenarios. Third, quantitative explainability validation with expert annotations provides clinically relevant evidence that underpins model explainability. In total, these enhancements suggest that the suggested framework is not only accurate but also more suitable for clinical use in real-world situations than current uterine fibroid detection methods.

5. Conclusion

This paper introduces an Explainable Hybrid Deep Learning (E-HDL) framework for multimodal uterine-fibroid detection that combines an adaptive preprocessing module, multi-branch feature extraction, dynamic attention fusion, and quantitative interpretability validation. The proposed architecture effectively integrates of complementary diagnostic cues of ultrasound and MRI modalities by means of ConvNeXt-based spatial modeling, Swin-Transformer contextual representation, and LSTM-based sequential learning, through a Dynamic Attention Fusion Layer (DAFL). Extensive five-fold cross-validation shows that E-HDL has a better diagnostic performance of 97.6% ± 0.5 accuracy, 97.1% F1-score and AUC of 0.989 (95% CI: 0.981-0.996). The proposed model significantly outperforms CNN-only (92.4%), CNN+RNN (95.2%), CNN+Transformer (96.1%), Swin (96.8%), and MedT (97.0%). The ablation study demonstrates a monotonic performance increase with the single-branch to the complete multi-branch integration. Cross-dataset evaluation also indicates improved generalization and robustness to imaging distributions. Beyond predictive performance, E-HDL also focuses on clinical transparency. Quantitative metrics of explainability indicate higher localization IoU (83.7%), high lesion activation (91.6%), and high attribution confidence (0.88) consistent with radiological reasoning. The system reduces diagnostic time (45 s vs. 120 s) as compared to radiologists and reduces false-positive and false-negative rates and minimizes inter-observer variability. The suggested E-HDL framework provides a high-performance, interpretable, and clinically reliable decision-support system that detects uterine fibroid, and its integration into real-world diagnostic processes has strong scalability potential. The proposed framework demonstrated strong robustness, generalization ability and statistical stability through extensive stratified five-fold cross-validation as well as independent external validation on HIFU-MRI and UFID-2023 datasets. Additionally, quantitative explainability analysis (Grad-CAM, SHAP and attention-based localization) showed strong correspondence between the model predictions and clinically relevant fibroid areas. The results suggest that E-HDL framework can serve as a clinically interpretable, reliable, and scalable system to support multimodal uterine-fibroid diagnosis. Future studies will focus on patient-level validation, multicenter clinical implementation and integration into real-time integration into radiology workstations.

References

[1] Z. Xia, H. Jin, Diagnostic Value of Ultrasonography Combined with Hysteroscopy in Intrauterine Space-Occupying Abnormalities. Contrast Media & Molecular Imaging, 2022(1), (2022) 6192311. https://doi.org/10.1155/2022/6192311

[2] Y. Toyohara, K. Sone, K. Noda, K. Yoshida, R. Kurokawa, T. Tanishima, S. Kato, S. Inui, Y. Nakai, M. Ishida, W. Gonoi, S. Tanimoto, Y. Takahashi, F. Inoue, A. Kukita, Y. Kawata, A. Taguchi, A. Furusawa, Y. Miyamoto, T. Tsukazaki, M. Tanikawa, T. Iriyama, M. Mori-Uchino, T. Tsuruga, K. Oda, T. Yasugi, K. Takechi, O. Abe, Y. Osuga, Development of a deep learning method for improving diagnostic accuracy for uterine sarcoma cases. Scientific Reports, 12(1), (2022) 19612. https://doi.org/10.1038/s41598-022-23064-5

[3] A. Tinelli, O. D'Oria, E. Civino, A. Morciano, A.A. Hashmi, G.M. Baldini, R. Stefanovic, Smooth muscle tumor of uncertain malignant potential (STUMP): a comprehensive multidisciplinary update. Medicina, 59(8), (2023) 1371. https://doi.org/10.3390/medicina59081371

[4] T. Huo, L. Li, X. Chen, Z. Wang, X. Zhang, S. Liu, J. Huang, J. Zhang, Q. Yang, W. Wu, Y. Xie, H. Wang, Z. Ye, K. Deng, Artificial intelligence-aided method to detect uterine fibroids in ultrasound images: a retrospective study. Scientific Reports, 13(1), (2023) 3714. https://doi.org/10.1038/s41598-022-26771-1

[5] W. Panyarak, W. Suttapak, K. Wantanajittikul, A. Charuakkra, S. Prapayasatok, Assessment of

YOLOv3 for caries detection in bitewing radiographs based on the ICCMS™ radiographic scoring system. Clinical Oral Investigations, 27(4), (2023) 1731-1742. https://doi.org/10.1007/s00784-022-04801-6

[6] T. Yang, L. Yuan, P. Li, P. Liu, Real-time automatic assisted detection of uterine fibroid in ultrasound images using a deep learning detector. Ultrasound in Medicine & Biology, 49(7), (2023) 1616-1626. https://doi.org/10.1016/j.ultrasmedbio.2023.03.013

[7] J. Smith, J.P. Zawaideh, H. Sahin, S. Freeman, H. Bolton, H.C. Addley, Differentiating uterine sarcoma from leiomyoma: BET1T2ER Check!. The British journal of radiology, 94(1125), (2021) 20201332. https://doi.org/10.1259/bjr.20201332

[8] D.J. Slotman, L.W. Bartels, A. Zijlstra, I.M. Verpalen, J.A. van Osch, I.M. Nijholt, E. Heijman, M. van 't Veer-ten Kate, E. de Boer, R.D. van den Hoed, Diffusion-weighted MRI with deep learning for visualizing treatment results of MR-guided HIFU ablation of uterine fibroids. European Radiology, 33(6), (2023) 4178-4188. https://doi.org/10.1007/s00330-022-09294-1

[9] M. Yang, Y. Chen, X. Zhou, R. Yu, N. Huang, J. Chen, Machine learning models for prediction of NPVR≥ 80% with HIFU ablation for uterine fibroids. International Journal of Hyperthermia, 42(1), (2025) 2473754. https://doi.org/10.1080/02656736.2025.2473754

[10] B. Wen, C. Li, Q. Cai, D. Shen, X. Bu, F. Zhou, Multimodal MRI radiomics-based stacking ensemble learning model with automatic segmentation for prognostic prediction of HIFU ablation of uterine fibroids: a multicenter study. Frontiers in Physiology, 15, (2024) 1507986. https://doi.org/10.3389/fphys.2024.1507986

[11] T. Wang, Y. Wen, Z. Wang, nnU-Net based segmentation and 3D reconstruction of uterine fibroids with MRI images for HIFU surgery planning. BMC Medical Imaging, 24(1), (2024) 233. https://doi.org/10.1186/s12880-024-01385-3

[12] D.J. Slotman, L.W. Bartels, I.M. Nijholt, J.A.F. Huirne, C.T.W. Moonen, M.F. Boomsma, Development and validation of a deep learning-based method for automatic measurement of uterus, fibroid, and ablated volume in MRI after MR-HIFU treatment of uterine fibroids. European Journal of Radiology, 178, (2024) 111602. https://doi.org/10.1016/j.ejrad.2024.111602

[13] C. Li, Z. He, F. Lv, H. Liao, Z. Xiao, Predicting the prognosis of HIFU Ablation of Uterine Fibroids using a Deep Learning-based 3D Super-Resolution DWI Radiomics Model: a multicenter study. Academic Radiology, 31(12), (2024) 4996-5007. https://doi.org/10.1016/j.acra.2024.06.027

[14] Y.H. Luo, I.L. Xi, R. Wang, H.O. Abdallah, J. Wu, A.Z. Vance, K. Chang, M. Kohi, L. Jones, S. Reddy, Z.S. Zhang, H.X. Bai, R.S. Goldberg, Deep Learning based on MR Imaging for Predicting Outcome of Uterine Fibroid Embolization. Journal of Vascular and Interventional Radiology, 31(6), (2020) 1010-1017. https://doi.org/10.1016/j.jvir.2019.11.032

[15] R. Golcha, P. Khobragade, A. Talekar, (2024) Multimodal Deep Learning for Advanced Health Monitoring a Comprehensive Approach for Enhanced Precision and Early Disease Detection. In 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE, India. https://doi.org/10.1109/ICITIIT61487.2024.10580622

[16] H. Kim, M.H. Choi, Y.J. Lee, D. Han, M. Mostapha, D. Nickel, Deep learning-accelerated T2-Weighted Imaging Versus Conventional T2-Weighted Imaging in the Female Pelvic Cavity: image Quality and Diagnostic Performance. Acta Radiologica, 65(5), (2024) 499-505. https://doi.org/10.1177/02841851241228192

[17] K. Drukker, M. Medved, C.B. Harmath, M.L. Giger, O.S. Madueke-Laveaux, Radiomics and Quantitative Multi-Parametric MRI for Predicting Uterine Fibroid Growth. Journal of Medical Imaging, 11(5), (2024) 054501-054501. https://doi.org/10.1117/1.JMI.11.5.054501

[18] C. Li, J. Tan, H. Li, Y. Lei, G. Yang, C. Zhang, Y. Song, Y. Wu, G. Bi, Q. Bi, The value of multiparametric MRI-based habitat imaging for differentiating uterine sarcomas from atypical leiomyomas: a multicentre study. Abdominal Radiology, 50(2), (2025) 995-1008. https://doi.org/10.1007/s00261-024-04539-7

[19] H. Xi, W. Wang, Deep learning based uterine fibroid detection in ultrasound images. BMC Medical Imaging, 24(1), (2024) 218. https://doi.org/10.1186/s12880-024-01389-z

[20] R. Sinha, R. Raina, M. Bag, B. Rupa, Empowering gynaecologists with Artificial Intelligence: Tailoring surgical solutions for fibroids. European Journal of Obstetrics & Gynecology and Reproductive Biology, 299, (2024) 72-77. https://doi.org/10.1016/j.ejogrb.2024.06.001

[21] D.K. Girija, M. Varshney, Proposed Model to Detect Uterine Fibroid By using Data Mining Techniques. Journal of Positive School Psychology, 6(2), (2022) 2062–2065. https://journalppw.com/index.php/jpsp/article/view/1780/997

[22] J. Zhang, C. Yang, C. Gong, Y. Zhou, C. Li, F. Li. Magnetic Resonance Imaging Parameter-based Machine Learning for Prognosis Prediction of High- Intensity focused Ultrasound Ablation of uterine fibroids. International Journal of Hyperthermia, 39(1), (2022) 835–846. https://doi.org/10.1080/02656736.2022.2090622

[23] T. Yang, (2023). Uterine Fibroid Ultrasound

Images. Mendeley Data. https://doi.org/10.17632/n2zcmcypgb.2

[24] İ. Öz, E.E. Yegin, A.U. Öz, E. Ulukaya, An AI-Driven Clinical Decision Support Model Based on Anemia and Fibroid Parameters to Guide Surgical Decision-Making. Medicina, 62(3), (2026) 555. https://doi.org/10.3390/medicina62030555

[25] J.L. Tao, Y.T. Wei, J.M. Chen, C.C. Liang, H.F. Zhang, X.N. Wang, Y. Ning, L. Cao, B. Bi, Machine Learning Models for the Prediction of Uterine Fibroids. Medicine, 104 (52), (2025) e46828. https://doi.org/10.1097/md.0000000000046828

[26] A. Sunder, M. Menon, H. Al Fadhel, B. Darwish, (2025). Artificial Intelligence-Powered MRI Segmentation for Uterine Fibroid Mapping: A Proof-of- Concept Study. Research Square. https://doi.org/10.21203/rs.3.rs-7605030/v1

[27] M. Lekshmanan Chinna, J.P. Pathrose Mary, Efficient Feature Extraction and Hybrid Deep Learning for Early Identification of Uterine Fibroids in Ultrasound Images. International Journal of Imaging Systems and Technology, 34(3), (2024) e23073. https://doi.org/10.1002/ima.23073

[28] M.K. Gupta, S. Pattnaik, C. Sudandiradoss. Artificial Intelligence Driven based Diagnostics with Combined Surgical and Non-Surgical management of uterine fibroids: a narrative review. Middle East Fertility Society Journal, 30(1), (2025) 64. https://doi.org/10.1186/s43043-025-00270-5

[29] S. Janghorbani, A. Caprio, L. Sam, B.C. Lee, M.R. Sabuncu, N.A. Lamparello, M. Schiffman, B. Mosadegh, Predicting Clinical Outcomes and Symptom Relief in Uterine Fibroid Embolization Using Machine Learning on MRI Features. AI, 6(9), (2025) 200. https://doi.org/10.3390/ai6090200

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License