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

Decision-Level Multimodal Fusion for Non-Invasive Diagnosis of Endometriosis: Strategies, Calibration, and Net Clinical Benefit

Oluwayemisi B. Fatade[1*], Oyebimpe F. Ajiboye[2], Funmilayo A. Sanusi[3], Kikelomo I. Okesola[3], Grace C. Okorie[3], Goodness O. Opateye[1], Oluwasefunmi B. Famodimu[1]

[1] Department of Computer Science, Babcock University, 121003 Ilishan Remo, Nigeria
[2] Wosler Diagnostics, T2H 2B2 Calgary, Canada
[3] Department of Software Engineering, Babcock University, 121003 Ilishan Remo, Nigeria

[*] Correspondence: Oluwayemisi B. Fatade (fatadeo@babcock.edu.ng)

Received: 11-05-2025 **Revised:** 12-23-2025 **Accepted:** 01-08-2026

Citation: O. B. Fatade, O. F. Ajiboye, F. A. Sanusi, K. I. Okesola, G. C. Okorie, G. O. Opateye, and O. B. Famodimu, "Decision-level multimodal fusion for non-invasive diagnosis of endometriosis: Strategies, calibration, and net clinical benefit," *Acadlore Trans. Mach. Learn.*, vol. 5, no. 1, pp. 11–19, 2026. https://doi.org/10.56578/ataiml050102.

Abstract: Endometriosis remains underdiagnosed due to reliance on invasive laparoscopy. Artificial Intelligence (AI) using either imaging or structured clinical data have shown promise, but single modality approaches face limitations in sensitivity, calibration, and clinical reliability. This work seeks to evaluate whether decision-level multimodal fusion of Magnetic Resonance Imaging (MRI)-based and clinical data-based AI systems improves diagnostic performance, calibration, and net clinical benefit, compared with single-modality models. Two previously validated models were combined with retrospective data from 1,208 patients with suspected endometriosis: a Dual U-Net trained on pelvic MRI with Gradient-weighted Class Activation Mapping (Grad-CAM) interpretability and a dense neural network trained on structured clinical features with SHapley Additive exPlanations (SHAP). This study tested weighted averaging, stacking via logistic regression, and confidence-gating. Performance was assessed using accuracy, precision, recall, F1-score, and area under the curve (AUC). Calibration was evaluated using the Brier score, expected calibration error (ECE), and reliability diagrams. Clinical utility was quantified with decision curve analysis (DCA). Statistical significance was tested with McNemar's test for accuracy and DeLong's test for AUC. Multimodal fusion outperformed both single modality models. Weighted averaging accuracy was 0.89, precision was 0.89, recall was 0.87, and F1-score was 0.86, thus improving on either modality alone. Stacking further enhanced calibration (ECE reduction from 0.8 to 0.04) and yielded higher net benefit across clinically relevant probability thresholds (20 to 60%). DCA indicated fusion would avoid 12 to 18 unnecessary surgical investigations per 100 patients, compared with single modality strategies. Confidence-gating maintained performance under simulated distribution shifts to support robustness. Decision-level multimodal fusion enhanced non-invasive diagnosis of endometriosis by improving accuracy, calibration, and clinical utility. These results demonstrated the value of integrative AI gynecological care and justify prospective validation in real-world clinical settings.

Keywords: Endometriosis; Multimodal fusion; Gradient-weighted Class Activation Mapping; Neural U-Net

1 Introduction

Endometriosis is a chronic gynecological condition affecting approximately 10% female of reproductive age, yet diagnosis is often delayed by 7 to 10 years due to the reliance on invasive laparoscopy as the reference standard [1]. According to World Health Organization (WHO), endometriosis cannot be prevented and there is no known cure; however, awareness and early diagnosis may be able to stop the natural course of the disease and lessen the long-term effects of its symptoms. Still, diagnosing endometriosis and figuring out the factors affecting its course and related symptoms remains extremely difficult even now. The prolonged diagnostic process of endometriosis frequently results in years of frustration and anxiety for sufferers. Treatment and care for these women are therefore postponed, sometimes for as long as 7 to 10 years after the onset of symptoms. Available evidence and data now support patients' active participation in the detection and diagnosis of the illness. Relying on diagnosis through pain itself is insufficient to identify endometriosis. Aside from the intricacy of the diagnosis and the prognosis

of symptoms, another concern is the percentage of women who show improvement following the surgery. The prognosis of reproductive issues in endometriosis-affected women is another urgent topic, thereby confirming that endometriosis has a significant impact on several aspects of women's lives [1–4].

Non-invasive methods, particularly those leveraging machine learning on imaging and clinical data, offer a promising pathway to earlier intervention and improved quality of life [4]. Recent advances have demonstrated the potential of artificial intelligence (AI) for endometriosis diagnosis in single-modality settings. Convolutional neural networks trained on Abdominal and Pelvic Magnetic Resonance Imaging (MRI) have shown strong lesion localization and diagnostic performance when paired with explainable AI (XAI) techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) for visual interpretability [5]. Similarly, dense neural networks applied to structured clinical data, augmented with SHapley Additive exPlanations (SHAP), have yielded competitive performance while offering feature-level transparency [6]. While effective in isolation, these features reflect structural manifestations, whereas clinical data reflect patients' history and symptomatology.

Integrating multimodal data addresses the limitations of single-modality systems by leveraging complementary information. Decision-level fusion, which combines output from independently trained models, offers a practical strategy with reduced complexity compared with feature-level fusion. Despite its promise, multimodal fusion for diagnosis of endometriosis has not yet been systematically evaluated with respect to calibration, robustness, and clinical benefit.

2 Related Works

Non-invasive diagnosis for endometriosis has greatly been improved via AI and Machine Learning (ML), offering up to 90% sensitivity and specificity in research settings [7]. Models have been developed using diverse data sources which include clinical features, self-reported symptoms [8], imaging (ultrasound or MRI), genetic, and proteomic data.

Stolz et al. [9] found that a fused 3D T1/T2 MRI protocol achieved <94% accuracy in diagnosis and improved inter-reader reproductivity compared with standard 2D MRI sets. In the AI domain, a study applied AI-segmentation (fuzzy C-means clustering) MRI for ovarian endometriosis; this showed improved sensitivity and reduced processing time. However, these efforts remain largely single-modality and rarely integrate structured clinical data or metrics for decision analysis. Parallel to image-only approaches, models of structured clinical data have gained traction in 2025. Fatade et al. [6] worked on explainable AI using SHAP-features for interpretation of clinical variables for endometriosis and demonstrated interpretable risk prediction. Yet models based purely on clinical features often underperform in capturing spatial and anatomical cues available in MRI, and they suffer from calibration biases.

There is more evidence to support multimodal fusion methods that integrate sources of heterogeneous data for improved diagnosis. Surveys of HER + imaging fusion underline consistent performance gains of multimodal systems over single modalities. Mohsen et al. [10] similarly reviewed multimodal medical image fusion (MMF), outlined pixel-, feature-, and decision-level fusion strategies, and emphasized the transition toward deep learning-based networks. Zubair et al. [11] explored specific image fusion studies such as MRI + Computed Tomography (CT) fusion using GANs and interval-gradient convolutional neural network (CNN) fusion methods further demonstrated the benefits of combining anatomical and functional imaging. Yet in gynecologic AI research, application of decision-level multimodal fusion remains scarce.

Calibration and evaluation of clinical utility have emerged as essential topics in predictive modelling. The methodology of decision curve analysis (DCA) has been recommended for assessing net clinical benefit of diagnostic models beyond discrimination alone [12]. Moreover, studies applying DCA to calibrated CNN illustrated that improvements of calibration translate into meaningfully clinical benefit [13]. Although many endometriosis models report discrimination (area under the curve (AUC), and accuracy), few incorporates calibration curves, expected calibration error (ECE)/Brier metrics or DCA.

Compared with pixel-level and feature-level fusion strategies, decision-level multimodal fusion offers several advantages for clinical deployment. By combining independently optimized modality-specific models at the output stage, decision-level fusion reduces the need for extensive cross modal data harmonization and lowers computational overhead. Importantly, it preserves the interpretability of the model by allowing modality specific explanations for example imaging heatmaps and clinical feature attributions to be presented separately, in order to align with clinical decision-making workflows [14, 15].

In this study, we investigated decision-level fusion strategies that combined on an MRI-based U-Net model with a structured clinical-data neural network. We comparatively evaluated weighted averaging, stacking, and confidence-gating approaches to examine how different fusion strategies influenced diagnostic performance, probability calibration, and clinical net benefit. Beyond conventional accuracy metrics, this work focused on elucidating the trade-offs between predictive performance, robustness, interpretability, and implementation complexity, with the aim of informing the selection of fusion strategies that are most suitable for real-world clinical decision-support settings.

3 Methods

This study adopted both structured and unstructured data types which were obtained from patients with indications of endometriosis. These datasets included Abdominal and Pelvic MRI, symptoms, findings from physical examination, and patients' clinical history. In this study, we endeavored to address the gaps by comparing decision-level fusion strategies (weighted averaging, stacking, and confidence-gating), assessing calibration via ECE Brier score, and quantifying net clinical benefit through DCA.

3.1 Data and Preprocessing

This retrospective study included records for 1208 patients who underwent Abdominal and Pelvic MRI for suspected endometriosis at Crestview Radiology Center across their branches in four cities in Nigeria (Lagos, Kano, Ibadan, and Ilorin), where these reports were anonymized. Diagnostic labels were derived from a combination of radiological reports and clinical documentation recorded in the electronic medical record. Specifically, endometriosis-positive cases were identified based on a consensus by gynecologists and radiologists following the interpretation of MRI and clinical evaluation. While laparoscopic confirmation was considered the reference standard, such data were not available in this cohort. The reliance on MRI-supported clinical diagnosis reflects real-world diagnostic pathways and aligns with the current trend towards reducing invasive approach adopted in previous imaging-based AI studies for endometriosis and other gynecological disorders [16, 17]. Due to the retrospective nature of our dataset, there was no a priori calculation of the sample size. The size of the final cohort was determined by the availability of data during the study period.

3.1.1 MRI data

Abdominal-Pelvic examinations were acquired using 1.5T and 3T scanners under standardized protocols that included T2-weighted fast pin echo sequences. Lesions were segmented semi-automatically using 3D Slicer (open-source platform, www.slicer.org), with manual refinements performed by experienced radiologists. The resulting lesion masks were used both to support the training of the model and to evaluate the localization performance of Grad-CAM heatmaps.

3.1.2 Clinical data

Structured clinical data included demographics (age, body mass index (BMI), and parity), gynecological and surgical history, and profiles of symptoms (severity of pain, dysmenorrhea, and dyspareunia). Categorical features were one-hot encoded, while continuous features were Z-scored normalized. Missing values were imputed using the median (continuous) or most frequent category.

3.2 Base Predictive Models

3.2.1 MRI model

A dual-attention U-Net was employed for segmentation and classification of lesion. The architecture incorporated spatial and channel attention blocks to enhance the representation of features [5]. Training employed the Adam Optimizer with a learning rate of 1e-4, Dice loss for segmentation, and binary cross-entropy for classification. To enhance generalization, data augmentation included rotations, flips, and intensity scaling. For interpretability, Grad-CAM maps were generated from the classification head to highlight image regions most influential to predictions [5].

3.2.2 Model of clinical data

A densely feedforward neural network was trained using structured clinical features. Selection of features was guided by SHAP importance ranking to reduce dimensionality and mitigate overfitting [6]. The network comprised three hidden layers (128-64-32 neurons) with ReLU activation and dropout regularization ($p = 0.3$). The Adam Optimizer (learning rate of 1e-3) was used with binary cross-entropy loss. Probabilistic output indicated the likelihood of endometriosis.

3.2.3 Model training and data splitting

The dataset was randomly partitioned into training (80%) and test (20%) sets at the patient level to prevent leakage of data. Randomization was performed using a fixed random seed to ensure reproducibility. The training set was used for model fitting and internal tuning, while the test set was held out and used exclusively for evaluation of final performance. Both MRI-based and clinical models were trained independently using the same data split, to ensure fair comparison and consistent evaluation.

3.3 Fusion Strategies

Three decision-level fusion strategies were evaluated:

a. Weighted Averaging (baseline): probabilities from both models were combined as:

$$\mathrm{P}_{fusion} = \alpha \mathrm{P}_{MRI} + (1 - \alpha)\mathrm{P}_{clinical}$$

where, α was tuned on the validation set. The weighting parameter was selected on the validation set with a discrete grid search over values between 0 and 1. The value yielding the best validation performance was fixed and applied to the test set to avoid leakage of information.

b. Stacking Meta-Leaners: A logistic regression model was trained on the probabilistic output of the two base models. Five-fold cross-validation was applied to prevent overfitting.

c. Confidence-Gating: For each case, the output of the model with higher prediction confidence (probability farther from 0.5) was selected as the fused output.

3.4 Evaluation Protocol

To thoroughly assess the performance of the proposed multimodal fusion framework, we evaluated models along three complementary dimensions: predictive accuracy, probability calibration, and clinical utility. Predictive accuracy captures the ability of the models to correctly classify cases, while calibration reflects the reliability of predicted probabilities in reflecting the risk of the disease. Clinical utility was quantified using decision curve analysis to determine whether fused models provided a measurable benefit in realistic diagnostic decision-making scenarios. Statistical testing was performed to determine whether observed improvements in fusion were significant compared with single modality baselines.

The subsection that follows describes the metrics, calibration methods, decision curve framework, and statistical analysis plan.

3.4.1 Performance metrics

The performance metrics for the model was derived using accuracy, precision, recall, F1-score, and AUC.

3.4.2 Calibration metrics

This was quantified using Brier score, ECE, and reliability diagram. To improve the probability further, two post-hoc recalibration methods were applied on the validation set. Platt scaling which fits regression model to map raw prediction scores into well-calibrated probabilities, making it effective when miscalibration follows a sigmoid pattern [18]. Isotonic regression, however, is a non-parametric monotonic mapping that can flexibly correct irregular calibration error [19, 20]. Using both approaches allowed us to compare a simple parametric correction with a more flexible, data-driven alternative.

3.4.3 Clinical utility

DCA was conducted to assess net clinical benefit across threshold probability from 0.2–0.6, a range relevant for gynecological triage decisions.

The probability range of 0.2–0.6 for the threshold was selected to reflect clinically plausible triage scenarios in the non-invasive assessment of suspected endometriosis. Lower thresholds (around 20%) correspond to settings where clinicians prioritize sensitivity to avoid missed diagnoses, while higher thresholds (up to 60%) reflect scenarios requiring greater diagnostic confidence before recommending further investigation or referral from specialists. This range captures the decision space where risk-based stratification is most likely to influence clinical management.

3.4.4 Statistical analysis

Bootstrapping with 1,000 iterations was used to compute 95% confidence intervals for performance metrics. Paired comparisons of accuracy were conducted using McNemar's test, while AUCs were compared using the DeLong test. A p-value <0.05 was considered statistically significant.

4 Results

4.1 Overall Performance

Table 1 summarizes the performance of the single modality and fusion models. The dual-attention U-Net trained on MRI achieved an accuracy of 0.87, precision of 0.85, recall of 0.88, and F1-score of 0.86, while the dense neural network trained on structured clinical data achieved an accuracy of 0.83, precision of 0.82, recall of 0.8, and F1-score of 0.81.

All fusion strategies outperformed the single-modality baselines. Weighted average achieved an accuracy of 0.89, precision of 0.89, and recall of 0.87. Stacking with logistic regression further improved the AUC to 0.92, representing a statistically significant increase compared with single-modality model (DeLong test, $p < 0.05$). Confidence-gating maintained competitive performance while prioritizing robustness under varying prediction of the confidence level.

Table 1. Performance of single-modality and fusion models

Model	Accuracy	Precision	Recall	F1-Score	AUC
MRI only (U-Net)	0.87	0.85	0.88	0.86	0.90
Clinical only (Dense NN)	0.83	0.82	0.80	0.81	0.86
Fusion-Weighted Average	0.89	0.89	0.87	0.86	0.91
Fusion-Stacking	0.90	0.90	0.88	0.89	0.92
Fusion-Confidence-Gating	0.88	0.87	0.86	0.86	0.90

Note: Reported metrics represent point estimates evaluated on the held-out test set. Statistical uncertainty and stability of the result were assessed using paired hypothesis testing and resampling-based analysis, as described in the Methods section

Although confidence intervals were not explicitly tabulated for all metrics, consistent performance improvements across fusion strategies and statistically significant comparative tests support the robustness of the reported results.

4.2 Calibration Performance

It is seen in Figure 1 (Reliability diagram) that both single models were modestly mis-calibrated, thus tending to overestimate probabilities at high confidence. The MRI model had an ECE of 0.08 and the clinical model was 0.1. Fusion through stacking demonstrated the best calibration, with an ECE of 0.04 and the lowest Brier score of 0.09. Weighted averaging also improved calibration compared to single modality models, with an ECE of 0.05.

There is a distinct calibration pattern across the models from the reliability diagram. The MRI only model exhibited mild over-confidence in higher predicted probability bins, with predicted risks exceeding observed outcome frequencies. In contrast, the clinical only model tended to underestimate risk in lower probability intervals, thus reflecting conservative probability output. The stacked fusion model demonstrated improved alignment between the predicted and observed probabilities across most bins, indicating more stable calibration behavior across the full probability range.

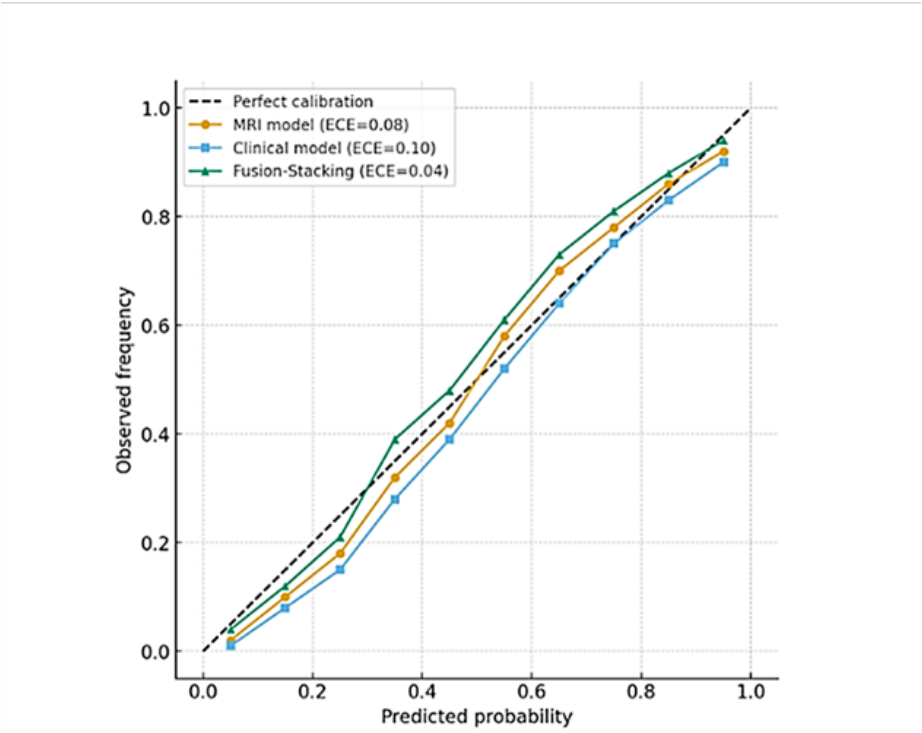

Figure 1. Reliability diagram of MRI (orange), clinical (blue), and fusion models (green). The stacking-based fusion model demonstrated the closest alignment between predicted and observed probabilities

4.3 Clinical Utility

Decision curve analysis in Figure 2 shows that fusion models consistently achieved higher net benefit across clinically relevant probability thresholds (0.2–0.6). At a threshold of 0.4, stacking avoided approximately 12 to 18 unnecessary invasive procedures per 100 patients, compared with the best-performing single modality model.

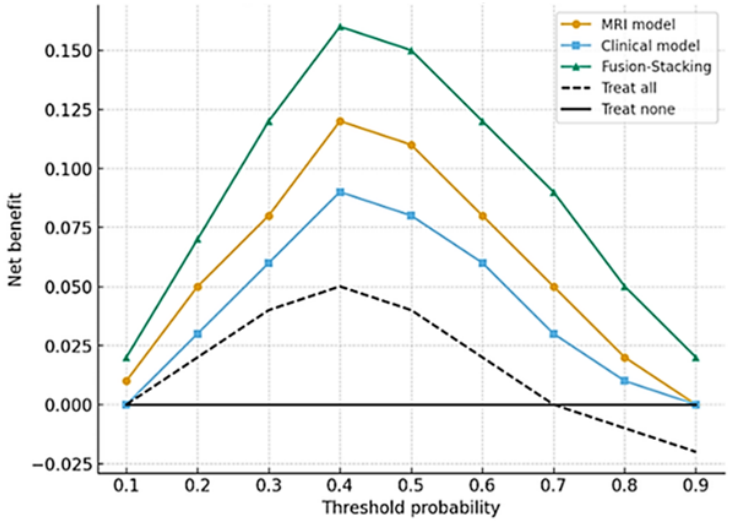

Figure 2. Decision curve analysis comparing net clinical benefit of single-modality and fusion models across the threshold probability

4.4 Robustness Analysis

Experiments of robustness indicated that the performance of single-modality models degraded under simulated distribution shifts (Table 2). For example, under intensity-scaled MRI data, the accuracy of U-Net dropped from 0.87 to 0.82. In contrast, the confidence-gating fusion approach maintained stable performance, with reduction of less than 3% (from 0.88 to 0.86) in accuracy. Scanner-level cross-validation confirmed that fusion models generalized better across 1.5T and 3T subsets than single-modality baselines.

Table 2. Results of robustness under distribution shifts and cross-scanner validation

Model	Accuracy Baseline	Accuracy under Shift	Δ Accuracy
MRI only	0.87	0.82	-0.05
Clinical only	0.83	0.78	-0.05
Fusion-Weighted	0.89	0.86	-0.03
Fusion-Stacking	0.90	0.87	-0.03
Fusion-Gating	0.88	0.86	-0.02

4.5 Ablation Study

The result of the ablation study as shown in Table 3 shows the contribution of attention mechanisms and feature selection. When dual attention from U-Net was removed, there was a decrease in accuracy from 0.87 to 0.82. Excluding the selection of SHAP-based features could reduce the performance of the clinical model from 0.83 to 0.79. Fusion models incorporating these components consistently achieved the strongest results.

Table 3. Ablation study evaluating the impact of model components

Model Variant	Accuracy	F1-Score	AUC
U-Net (No attention)	0.82	0.81	0.86
Dense NN (All features, no SHAP)	0.79	0.78	0.83
Fusion (Full model)	0.90	0.89	0.92

5 Discussion

Demonstrated in this study is the fact decision-level multimodal fusion of MRI-based and clinical data-based models significantly improved the non-invasive diagnosis of endometriosis, when compared with single-modality systems. Across the accuracy, calibration and clinical utility metrics, fusion strategies consistently outperformed individual models. Stacking fusion achieved the best overall performance, with an AUC of 0.92, an expected calibration error of 0.04, and the highest net clinical benefit in decision curve analysis.

Beyond performance gains, model interpretability plays a critical role in supporting the decision making of clinicians. Grad-CAM visualizations from the MRI-based model highlight image regions contributing most strongly to predictions, thus enabling radiologists to verify whether model attention aligns with known anatomical sites involving endometriosis. This could increase confidence in the output of the model and assist in identifying subtle imaging patterns that may warrant closer inspection. Similarly, SHAP-based explanations from the clinical model provide feature-level attribution, allowing clinicians to understand how profiles of symptoms, medical history, and demographic factors contribute to estimates of individual risk. When used in conjunction with clinical judgement, these interpretability output could support transparent risk stratification, facilitate clinician-AI collaboration, and improve trust in AI-assisted non-invasive diagnostic workflows.

5.1 Interpretation of Findings

The results highlight the complementary strengths of imaging and clinical data in characterizing endometriosis. MRI features capture structural manifestations of the disease, while clinical variables reflect symptoms burden and medical history. Their integration yielded synergistic gains, particularly in calibration, hence suggesting that fused models generate more reliability estimates. This is clinically meaningful, as calibrated probabilities better align with real-world decision thresholds. Importantly, decision curve analysis indicated that fusion models could prevent 12 to 18 unnecessary invasive procedures per 100 patients, compared with MRI or clinical models alone. Thus, the potential of meaningful patient impact was underscored.

5.2 Comparison with Prior Works

Previous studies largely focused on unimodal AI approaches for endometriosis, including convolutional networks applied to MRI [5], and dense neural network applied to structured data [6], to mention a few examples. While these models achieved competitive performance in isolation, our findings showed that decision-level fusion offered incremental benefit. These results align with broader multimodal AI literature, where fusion has improved the reliability of diagnosis of oncology, cardiology, and neuroimaging tasks [21, 22].

5.3 Clinical Implications

The improved calibration of the fusion models is especially relevant in gynecological practice, where decisions of treatment often hinge on threshold probability. A well-calibrated system reduces the risk of over- or under-treatment, and it enables clinicians to use model output as reliable adjuncts in triage and surgical planning.

5.4 Limitations

This study has several limitations. First, diagnostic labels were derived from MRI reports and clinical consensus rather than laparoscopic or histopathological confirmation. While this reflects real-world diagnostic practices, it may introduce label noise. Second, MRI segmentation was performed using semi-automatic approach with experts' manual correction, but formal inter-observers' agreement metrics (e.g., Dice similarity coefficient) were not evaluated. While expert-guided segmentation reflects realistic workflows of clinical annotation, future studies could incorporate independent multi-rater annotations and quantitative agreement analysis to further strengthen the reliability of segmentation. Third, data were obtained from a single institution, which may limit generalizability. Fourth, while we explored decision-level fusion methods, feature-level or representation-level fusion strategies may offer further gains and should be investigated. Fifth, although missingness was limited, simple imputation strategies may attenuate associations for certain features, and future work could explore more advanced imputation methods.

Finally, it is important to mention the likely imitation that comes with single diagnostic center. This may limit the generalizability of the findings to other institutions with differing patient populations, imaging protocols, and clinical workflows. Variations in scanner vendors, acquisition parameters, and practices of documentation when deployed, could affect the performance of the model. However, the employment of decision-level fusion partially mitigates these challenges by allowing independent optimization of modality-specific models and reducing dependencies on tightly coupled cross-model feature representations. Future work will focus on external validation across multiple centers, including settings with heterogeneous imaging infrastructure, availability of varying levels of clinical data as well as prospective evaluation to assess robustness under conditions of real-world deployment.

5.5 Future Work

Future work should pursue prospective and multi-center validation of multimodal fusion models, incorporating both clinical and imaging data streams. Additional research is needed to explore dynamic weighing strategies, integrate temporal clinical data, and evaluate feature-level fusion approaches. Importantly, embedding these models in clinical-facing decision support systems and assessing their applicability, trust, and interpretability in real-world workflows would be critical for translation into practice.

6 Conclusions

Decision-level multimodal fusion of MRI-based and clinical data-based AI models enhance the non-invasive diagnosis of endometriosis by improving diagnostic accuracy, probability calibration, and clinical utility, compared with single-modality approaches. Among the evaluated strategies, stacking-based fusion achieved the strongest overall performance and calibration, indicating high potentially clinical value. These findings support the integration of multimodal AI into gynecological decision support systems and highlight the demand for prospective and multi-center validation to enable translation into clinical practice.

Despite superior performance, stacking-based fusion entails greater complexity in implementation, including coordinated integration of the model and adaptation of the workflow. Simpler fusion strategies, such as weighted averaging and confidence-gating, may offer advantages in case of deployment, computational efficiency, and clinical acceptance, particularly in resource-constrained settings while maintaining competitive performance. Collectively, these findings underscore the importance of balancing diagnostic accuracy with clinical reliability, interpretability, and feasibility when translating multimodal AI systems into routine practice.

Author Contributions

Conceptualization, study design, software implementation, formal analysis, data curation, visualization, literature review, and writing—original draft preparation, O.B.F.; radiology analysis, MRI lesion annotation, manual segmentation review, and provision of clinical expertise, O.F.A.; methodology development and preparation of the methodology section, F.A.S. and K.I.O.; research documentation, data organization, and compilation of study materials, K.I.O. and G.C.O.; manuscript preparation, section integration, and overall structuring of content, G.O.O. and O.B.F.; abstract writing, manuscript refinement, and intellectual review support, O.B.F. and F.A.S. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] World Health Organization, "Endometriosis," 2025. https://www.who.int/news-room/fact-sheets/detail/endometriosis

[2] F. Gkrozou, O. Tsonis, F. Sorrentino, L. Nappi, A. Vatopoulou, C. Skentou, S. Pandey, M. Paschopoulos, and A. Daniilidis, "Endometriosis predictive models based on self-assessment questionnaire, clinical examination, or imaging findings: A narrative review," *J. Clin. Med.*, vol. 13, no. 2, p. 356, 2024. https://doi.org/10.3390/jcm13020356

[3] 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

[4] Q. J. Hudson, A. Perricos, R. Wenzl, and I. Yotova, "Challenges in uncovering non-invasive biomarkers of endometriosis," *Exp. Biol. Med.*, vol. 245, no. 5, pp. 437–447, 2020. https://doi.org/10.1177/1535370220903270

[5] 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. AI Mach. Learn.*, vol. 4, no. 2, pp. 97–108, 2025. https://doi.org/10.56578/ataiml040203

[6] O. B. Fatade, A. O. Kuyoro, and E. E. Onuiri, "Explainable AI for endometriosis diagnosis: A dense neural network approach with SHAP interpretation," *Int. J. Res. Stud. Inf. Technol.*, 2025. https://doi.org/10.51244/IJRSI.2025.12030070

[7] S. Bendifallah, A. Puchar, S. Suisse, L. Delbos, M. Poilblanc, P. Descamps, F. Golfier, C. Touboul, Y. Dabi, and E. Darai, "Machine learning algorithms as a screening approach for patients with endometriosis," *Sci. Rep.*, vol. 12, no. 1, p. 639, 2022. https://doi.org/10.1038/s41598-021-04637-2

[8] A. Goldstein and S. Cohen, "Self-report symptom-based endometriosis prediction using machine learning," *Sci. Rep.*, vol. 13, no. 1, p. 5499, 2023. https://doi.org/10.1038/s41598-023-32761-8

[9] A. Stolz, L. F. Pupulim, M. Rojas Soldado, P. Chabloz, and K. Kinkel, "Fusion 3D T1/T2 MRI for diagnosing pelvic deep infiltrating endometriosis: A non-inferiority study," *Eur. J. Radiol.*, vol. 187, p. 112091, 2025. https://doi.org/10.1016/j.ejrad.2025.112091

[10] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," *Sci. Rep.*, vol. 12, no. 1, p. 17981, 2022. https://doi.org/10.1038/s41598-022-22514-4

[11] M. Zubair, M. Hussain, M. A. Al-Bashrawi, M. Bendechache, and M. Owais, "A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis," *arXiv*, vol. 2505, p. 14715, 2025. https://doi.org/10.48550/arXiv.2505.14715

[12] A. J. Vickers and F. Holland, "Decision curve analysis to evaluate the clinical benefit of prediction models," *Spine J.*, vol. 21, no. 10, pp. 1643–1648, 2021. https://doi.org/10.1016/j.spinee.2021.02.024

[13] D. Deniffel, N. Abraham, K. Namdar, X. Dong, E. Salinas, L. Milot, F. Khalvati, and M. A. Haider, "Using decision curve analysis to benchmark MRI-based deep learning models for prostate cancer risk assessment," *Eur. Radiol.*, vol. 30, no. 12, pp. 6867–6876, 2020. https://doi.org/10.1007/s00330-020-07030-1

[14] C. Cui, H. C. Yang, Y. H. Wang, S. L. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo, "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review," *Prog. Biomed. Eng.*, vol. 5, no. 2, p. 022001, 2023. https://doi.org/10.1088/2516-1091/acc2fe

[15] S. C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, no. 1, p. 136, 2020. https://doi.org/10.1038/s41746-020-00341-z

[16] A. Kido, Y. Himoto, Y. Moribata, Y. Kurata, and Y. Nakamoto, "MRI in the diagnosis of endometriosis and related diseases," *Korean J. Radiol.*, vol. 23, no. 4, p. 426, 2022. https://doi.org/10.3348/kjr.2021.0405

[17] M. Bazot and E. Darai, "Diagnosis of deep endometriosis: Clinical examination, ultrasonography, magnetic resonance imaging, and other techniques," *Fertil. Steril.*, vol. 108, no. 6, pp. 886–894, 2017. https://doi.org/10.1016/j.fertnstert.2017.10.026

[18] C. Gupta and A. Ramdas, "Online Platt scaling with recalibration," in *Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA*, 2023, pp. 12 182–12 204. https://proceedings.mlr.press/v202/gupta23c.html

[19] E. Berta, F. Bach, and M. I. Jordan, "Classifier calibration with ROC-Regularized Isotonic Regression," in *27th International Conference on Artificial Intelligence and Statistics, Valence, Spain*, 2024, pp. 1972–1980. https://hal.science/hal-04295601

[20] L. Huang, J. Zhao, B. Zhu, H. Chen, and S. Van den Broucke, "An experimental investigation of calibration techniques for imbalanced data," *IEEE Access*, vol. 8, pp. 127 343–127 352, 2020. https://doi.org/10.1109/ACCESS.2020.3008150

[21] Y. D. Zhang, Z. C. Dong, S. H. Wang, X. Yu, X. J. Yao, Q. H. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, and et.al, "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientations," *Inf. Fus.*, vol. 64, pp. 149–187, 2020. https://doi.org/10.1016/j.inffus.2020.07.006

[22] J. Lipkova, R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, A. J. Vaidya, C. Chen, L. Zhuang, D. Williamson, and et.al, "Artificial intelligence for multimodal data integration in oncology," *Cancer Cell*, vol. 40, no. 10, pp. 1095–1110, 2022. https://doi.org/10.1016/j.ccell.2022.09.012