Abdominal Radiology
https://doi.org/10.1007/s00261-025-04942-8

RESEARCH

Advancing endometriosis detection in daily practice: a deep learning-enhanced multi-sequence MRI analytical model

Mana Moassefi[1] · Shahriar Faghani[1] · Ceylan Colak[1] · Shannon P. Sheedy[1] · Pamela L. Causa Andrieu[1] · Sherry S. Wang[1] · Rachel L. McPhedran[1] · Kristina T. Flicek[1] · Garima Suman[1] · Hiroaki Takahashi[1] · Candice A. Bookwalter[1] · Tatnai L. Burnett[1] · Bradley J. Erickson[1] · Wendaline M. VanBuren[1]

Received: 8 January 2025 / Revised: 31 March 2025 / Accepted: 6 April 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

Background and purpose Endometriosis affects 5–10% of women of reproductive age. Despite its prevalence, diagnosing endometriosis through imaging remains challenging. Advances in deep learning (DL) are revolutionizing the diagnosis and management of complex medical conditions. This study aims to evaluate DL tools in enhancing the accuracy of multi-sequence MRI-based detection of endometriosis.

Method We gathered a patient cohort from our institutional database, composed of patients with pathologically confirmed endometriosis from 2015 to 2024. We created an age-matched control group that underwent a similar MR protocol without an endometriosis diagnosis. We used sagittal fat-saturated T1-weighted (T1W FS) pre- and post-contrast and T2-weighted (T2W) MRIs. Our dataset was split at the patient level, allocating 12.5% for testing and conducting seven-fold cross-validation on the remainder. Seven abdominal radiologists with experience in endometriosis MRI and complex surgical planning and one women's imaging fellow with specific training in endometriosis MRI reviewed a random selection of images and documented their endometriosis detection.

Results 395 and 356 patients were included in the case and control groups respectively. The final 3D-DenseNet-121 classifier model demonstrated robust performance. Our findings indicated the most accurate predictions were obtained using T2W, T1W FS pre-, and post-contrast images. Using an ensemble technique on the test set resulted in an F1 Score of 0.881, AUROCC of 0.911, sensitivity of 0.976, and specificity of 0.720. Radiologists achieved 84.48% and 87.93% sensitivity without and with AI assistance in detecting endometriosis. The agreement among radiologists in predicting labels for endometriosis was measured as a Fleiss' kappa of 0.5718 without AI assistance and 0.6839 with AI assistance.

✉ Wendaline M. VanBuren
Vanburen.wendaline@mayo.edu

Mana Moassefi
Moassefi.mana@mayo.edu

Shahriar Faghani
Faghani.shahriar@mayo.edu

Ceylan Colak
Colak.ceylan@mayo.edu

Shannon P. Sheedy
Sheedy.Shannon@mayo.edu

Pamela L. Causa Andrieu
causaandrieu.pamela@mayo.edu

Sherry S. Wang
Wang.sherry@mayo.edu

Rachel L. McPhedran
mcphedran.rachel@mayo.edu

Kristina T. Flicek
flicek.kristina@mayo.edu

Garima Suman
Suman.Garima@mayo.edu

Hiroaki Takahashi
Takahashi.Hiroaki@mayo.edu

Candice A. Bookwalter
Bookwalter.candice@mayo.edu

Tatnai L. Burnett
burnett.tatnai@mayo.edu

Bradley J. Erickson
bje@mayo.edu

[1] Mayo Clinic, Rochester, USA

Published online: 15 April 2025

Conclusion This study introduced the first DL model to use multi-sequence MRI on a large cohort, showing results equivalent to human detection by trained readers in identifying endometriosis.

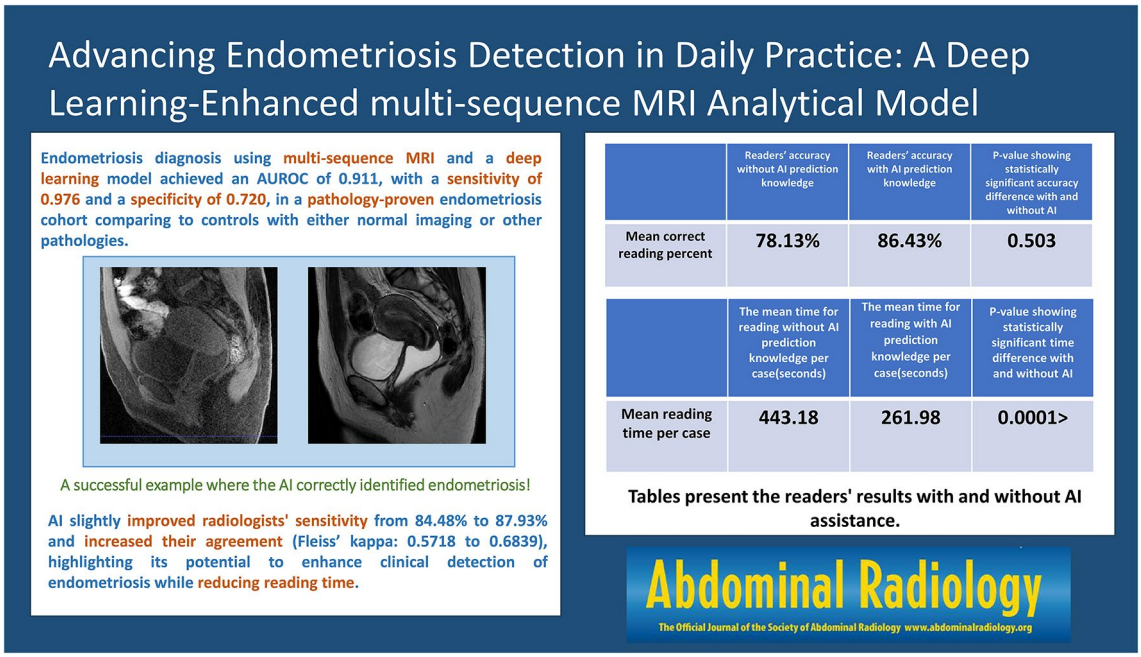
Key findings

- Endometriosis diagnosis using multi-sequence MRI and a 3D-DenseNet-121 deep-learning model showed promising results. The model achieved F1 Score: 0.881, AUROCC: 0.911, sensitivity: 0.976, and specificity: 0.720.
- AI slightly improved radiologists' sensitivity from 84.48 to 87.93% and increased their agreement (Fleiss' kappa: 0.5718 to 0.6839), highlighting its potential to enhance clinical detection of endometriosis while reducing reading time.

Highlights

Deep-learning models leveraging multi-sequence MRI significantly enhance endometriosis detection accuracy, surpassing traditional methods and improving radiologists' performance, marking an advance in imaging-based diagnostics.

Graphical abstract



Keywords Endometriosis · Deep learning · MRI · Diagnosis · Artificial intelligence

Abbreviations

CNNs	Convolutional Neural Networks
DL	Deep Learning
MRI	Magnetic Resonance Imaging
AI	Artificial Intelligence
ROS	Reactive Oxygen Species
T2W	T2-weighted imaging
T1W	T1-weighted images
CE-T1W	Contrast-enhanced T1-weighted images
AUROCC	Area Under the ROC Curve
US	Ultrasonography or Ultrasound
FS	Fat saturated

Introduction

Endometriosis is an estrogen-dependent disorder defined as the presence of endometrial tissue outside of the uterus. This condition, affecting 5–10% of women of reproductive age, triggers a chronic inflammatory response, leading to symptoms that can significantly impact quality of life, including pain and subfertility [1, 2]. Studies also indicate a prevalence of 20–47% in infertile women, and up to 1% may experience malignant transformation of endometriosis, predominantly affecting the ovary [3–6]. The delay in diagnosis is approximately 6–10 years [7]. 50% of women with endometriosis suffer from infertility, and 50–80% experience pelvic pain [8, 9]. In 2008, the annual healthcare costs

for endometriosis in the United States were estimated to be approximately $4,000 per affected woman. These costs are comparable to those associated with chronic diseases such as type 2 diabetes, rheumatoid arthritis, and Crohn's disease [10]. The most common sites affected include the ovaries, uterine ligaments, pelvic peritoneum, bowel, bladder, and vagina [11].

The disease's multifaceted nature demands a nuanced approach to diagnosis, where clinical evaluation and imaging studies play crucial roles. Laparoscopy, complemented by histological verification of ectopic endometrial-like tissue, remains the definitive diagnostic criterion for endometriosis, however, international recommendations now include pre-operative imaging [12–15]. [16–18]. Magnetic Resonance Imaging(MRI) has emerged as an indispensable tool in the diagnostic and pre-surgical strategy for endometriosis treatment planning [16–18, 6, 7]. MRI provides high reproducibility and the ability to visualize a large field of view in a multi-planar and multiparametric manner [8, 9]. With its superior soft tissue contrast resolution, MRI, through a combination of various sequences, offers intricate details on the localization and histological features of endometriosis, thereby enriching the understanding and management of this complex condition [4, 19–22]. The development and progression of endometriosis involve oxidative stress and an imbalance in reactive oxygen species (ROS), leading to cellular damage and altered gene expression. [23–29]. Excess iron from lysed erythrocytes in hemorrhagic lesions contributes to oxidative injury, inflammation, and tissue damage in the pelvic and abdominal cavities. [30–32].

Recent advances in Artificial Intelligence (AI) and Deep Learning (DL), particularly through Convolutional Neural Networks (CNNs), are transforming medical diagnostics. Their image classification performance enables sophisticated analysis of medical images, marking a pivotal improvement in diagnostic precision. The fibrosis, inflammatory damage and iron deposition associated with endometriosis may manifest as specific imaging features that are difficult for the human eye to discern. However, CNNs, which analyze all pixel values and their correlations, could potentially capture these subtle features or localized lesions.

Given the processes and pathophysiology of endometriosis, particularly the role of iron overload and inflammation in the abdominal area, we aim to explore whether these pathological changes can be detected through MRI pixel values. The fibrosis, inflammatory damage, and iron deposition associated with endometriosis may manifest as specific imaging features that are difficult for the human eye to discern. However, CNNs, which analyze all pixel values and their correlations, could potentially capture these subtle features or localized lesions.

This paper aims to illuminate the advances in utilizing DL for MRI analysis in diagnosing endometriosis, exploring how this innovative approach can enhance our understanding of a condition. We are investigating how the application of AI impacts radiologists' reading time and accuracy when analyzing these images for this specific condition. Through this lens, we explore the potential of DL to revolutionize the diagnostic process, offering a more precise, efficient, and personalized approach to combating endometriosis and its numerous repercussions on women's health.

Method

Our institutional review board approved the study and granted a waiver of informed consent.

Study population

We reviewed our institutional database to identify individuals who had histopathologically proven abdominal or pelvic endometriosis (regardless of severity of the disease) from January 1, 2015, to January 1, 2024. Patients who underwent MR imaging within three months before their diagnosis were selected. Individuals from our institutional database who underwent the same MRI protocol without evidence of endometriosis in their pathology and radiology reports were included as the control [17]. This group comprised patients who had undergone surgeries at our institution with no evidence of endometriosis in pathology, radiology, and surgical reports. The control group consisted of images with various pathologies, not necessarily normal images. They were included regardless of the specific diagnosis. All endometriosis surgeries were performed by fellowship-trained minimally invasive gynecological surgeons in a high-volume subspecialized endometriosis practice and pathology was evaluated by a high-volume, subspecialized international laboratory. The patients (subjects and controls) were between 12 and 55 years old. Tests confirmed that the population followed a normal distribution. Subsequently, a t-test was conducted to investigate any differences in age distributions.

Dataset

All participants in our cohort underwent a standardized MRI protocol that included sagittal T2-weighted imaging (T2W), sagittal fat-saturated T1-weighted images (T1W FS), and sagittal contrast-enhanced fat-saturated T1-weighted images (CE-T1W FS) [17]. Pelvic MRI examinations were conducted using 1.5 T or 3 T MRI systems. The usage of

glucagon and vaginal gel or a mixture of those during these examinations was documented.

Model architecture and training process

We partitioned our dataset at the patient level (to avoid information leakage) into training/validation and test sets, stratified by diagnosis (endometriosis versus non-endometriosis) using the Scikit-learn package [36, 37]. We set aside 12.5% of the data as the final test set for evaluating the ultimate models, while the remaining 87.5% was used for training and validation. For the training/validation portion, we performed 7-fold cross-validation to assess the model's stability in response to data variations. We explored combinations of the three sequences (T2W individually, T1W pre- and post-contrast together, and all three combined). Where the models utilized multiple sequences, we considered the different sequences as separate input channels, normalizing the voxel intensity for each channel to zero mean and unit standard deviation. The models accepted one, two, or three MRI sequences (T2W, T1W FS, CE-T1W FS) as inputs. In scenarios involving multiple sequences, the model integrated these modalities into a 4-dimensional array (c, x, y, z), with "c" indicating the channel number (T2W, T1W FS, CE-T1WFS) and (x, y, z) denoting the dimensions (width, height, slice) of each individual MRI sequence dataset.

We adjusted the width and height dimensions of each channel to 300×300 pixels and applied symmetric zero padding along the z dimension, standardizing all images to match the maximum z dimension observed in our dataset (200).

To enhance the model's generalizability and mitigate the risk of overfitting, we employed several augmentation techniques, including random shifts in intensity (RandShiftIntensityd with a probability of 0.5 and offsets of 100), random histogram shifts (RandHistogramShiftd with a probability of 1 and three control points), random bias field application (RandBiasFieldd with a degree of 3, a coefficient range of 0.0 to 0.1, and a probability of 0.5), and the addition of random Gaussian noise (RandGaussianNoised with a probability of 0.1, a mean of 0.0, and a standard deviation of 0.1) [38].

We utilized a 3-D DenseNet-121 model from the MONAI framework (https://monai.io/) for training. DenseNet, a type of CNN, establishes direct connections between each layer and every subsequent layer in a feed-forward manner, efficiently extracting key features through convolution processes. It incorporates a 1×1 convolutional bottleneck layer before each 3-D convolutional layer, aimed at reducing the volume of feature maps. This architecture includes dense blocks that feature this bottleneck design and implement dimensionality reduction in transition layers, thereby enhancing parameter efficiency and diminishing the overall complexity of the model. The Adam optimization algorithm was employed with a batch size of 8 and an initial learning rate of 1e-5 for 200 epochs. The model was trained using a setup consisting of two clusters of four NVIDIA-A100 GPUs each. Due to the absence of significant class imbalance between endometriosis cases and controls, we opted not to implement specific sampling strategies. Model performance was assessed based on F1, area under the ROC curve (AUROCC), sensitivity, and specificity.

Our approach involved training individual models across seven folds for each specified set of image sequences, subsequently aggregating the best-performing models (selected based on AUROCC) from each fold into an ensemble. This ensemble was then applied to conduct inference on the independent test set (Fig. 1).

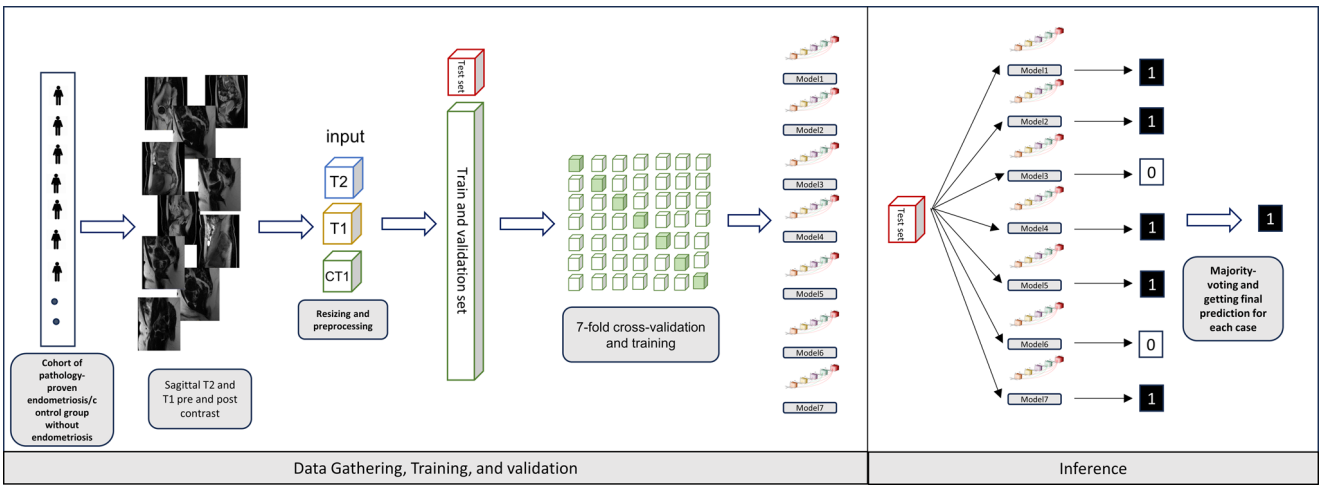


Fig. 1 Study pipeline

Generating occlusion maps

Following the development of our models, we generated occlusion maps to identify which regions contributed most significantly to the AI's predictions. Occlusion maps are a visualization technique used to understand which parts of an input image are most important for a model's predictions. In our study, we first segmented our MRI images into small patches measuring $4 \times 4 \times 4$. We then processed the complete MRI images through our trained models to obtain a baseline prediction. Subsequently, we processed the images again, each time omitting one of the patches, and noted the changes in the model's predictions. The difference in the logit outputs (prediction scores before the final activation function) of the model with and without each patch was calculated. These differences helped us generate occlusion maps, indicating the significance of each patch in influencing the model's decision-making process. This method allows us to visually interpret the predictive features recognized by the AI, enhancing our understanding of its operational mechanics [39].

Reader section

We engaged 8 readers, each at different experience levels, to evaluate a randomly selected set of 10 to 20 cases from our test set. The readers comprised one woman's imaging fellow (LX) with specific training in MRI for endometriosis and seven abdominal radiologists with 2 to more than 10 years of experience. Initially, each reader assessed the 10 or 20 cases blind to the AI predictions, and we recorded the time taken for each reading. After a washout period of 2 months, the same readers re-evaluated the same MRIs, this time with knowledge of the AI model predictions (without the occlusion maps) for each case. We again recorded the reading times to assess any differences. One of our readers participated only once in the reading for the 'without AI' part due to his time availability. Among 4 radiologists who read exactly 20 similar cases with and without knowing AI predictions, we evaluated the inter-rater agreement using Fleiss' kappa. Fleiss' kappa a statistical measure that quantifies the level of agreement beyond chance. This was calculated both before and after the introduction of AI assistance to assess its impact on diagnostic consistency. We calculated the p-value of the agreement scores using a bootstrapping technique to determine whether there was a statistically significant difference between the results with and without AI.

Statistical analysis

We conducted the Kolmogorov-Smirnov Test of Normality to assess age distribution across all datasets, ensuring adherence to the normality assumption. For continuous variables such as age, we applied an unpaired t-test. These analyses were conducted using the SPSS 28.0 software (SPSS Inc., Chicago, IL, USA). A P value below 0.05 was considered indicative of statistical significance. An independent two-sample t-test was used to compare the means of different sets of numbers to assess the reader's reading time difference with and without AI assistance. We evaluated the inter-rater agreement using Fleiss' kappa.

Results

Database characteristics

We identified 395 patients with a diagnosis of endometriosis based on their pathology reports, including any type of endometriosis. Additionally, we identified 356 patients whose pathology reports did not contain any keywords related to endometriosis such as endometrioma in their descriptions, findings, or diagnosis sections.

We then identified MRI scans in all these subjects that had sagittal T2W, T1W FS pre-contrast, and post-contrast MRI acquisitions. The number of patients meeting the criteria for each group and median age are provided in the accompanying table (Table 1). The age range in our endometriosis group was from 12 to 55 years, so we limited the control group to the same age range.

Model results

We trained models using three different image sets: T2W images alone, T1W FS pre- and post-contrast images, and T2W, T1W FS pre- and post-contrast images. The results from the 7-fold cross-validation for each input set are presented in Table 2. Additionally, the table includes results on the test sets derived from an ensemble of the seven model decisions.

Our model using the combination of T2W, T1W FS, and CE-T1W FS images demonstrated the best performance, with a sensitivity of 97.6% and a specificity of 72% on the hold-out test fold. In our test set, the False Positive rate of 10.45% compared to the False Negative rate of 1.49% indicates that the model tends to overdiagnose cases of endometriosis. The model using only T1W FS pre- and post-contrast images showed a sensitivity of 90.9% and a specificity of 56.6%. The model using only T2W images had the lowest performance, with a sensitivity of 82.1% and a specificity of 51.3%. Figure 2 illustrates the ROC curve for the seven folds and the final ensemble results, based on T1W FS pre- and post-contrast, as well as T2W images. The 95% confidence interval (CI) for the final AUC value was calculated

Table 1 The table provides the number of patients with identified sequences and their median age

	T2 Available	T1 precontrast available	T1-post contrast available	T1 Pre and post available	T2 and T1 Pre available	T2, T1 Pre and post available	Total number of cases Found in the database
No. of patients							
With endometriosis	389	377	389	349	365	343	395
Without endometriosis	293	287	250	246	264	199	356
Median age							
With endometriosis	38	33	35	35	34	35	36
Without endometriosis	33	39	38	37	38	38	38
Standard Deviation							
With endometriosis	7.35	7.34	8.11	7.71	7.65	7.35	7.47
Without endometriosis	12.20	9.44	7.67	10.42	8.51	9.29	9.88
P-value for age difference between with and without endometriosis group	*0.631*	*0.601*	*0.766*	*0.870*	*0.621*	*0.610*	*0.726*

The p-values higher than 0.05 show that there was no significant age difference between the case and control groups

using the bootstrapping technique, with average final model scores for the cases.

Figure 3 presents one example of a successful prediction and one example of a failure by our AI model.

Reader results

We provided 10 to 20 cases to our readers for evaluation, with half cases having endometriosis and half cases without. Readers were blinded to the diagnosis and were asked to determine the presence of endometriosis and to report the duration taken to read each case. We provided our readers with only binary predictions (yes or no) for endometriosis, without including the computed maps. The AI prediction was accurate for 28 out of 30 selected cases. Since this cohort was drawn from the test set, we were confident that the AI was unaware of their endometriosis labels. Table 3 presents a comprehensive summary of the reader study results, highlighting key findings and statistical analyses. Our results showed that agreement with and without AI had the p-value of 0.503 which is statistically not significant.

The agreement among radiologists when predicting labels for endometriosis without AI assistance yielded a Fleiss' kappa of **0.5718**, indicating a moderate level of agreement. After providing them with AI predictions, the inter-rater agreement increased to a Fleiss' kappa of **0.6839**, reflecting a higher level of agreement.

Occlusion maps

We provided occlusion maps to highlight the pixels in the input MR images that were most influential in the AI model's predictions. As there is currently no evidence that these maps can definitively serve as explainability tools, we chose to present only a few cases as a sanity check to ensure that our model is not biased toward specific irrelevant regions of the MRI. We emphasize that these maps are not generated for the purpose of segmenting the lesions but rather to highlight influential points both within and outside the lesion area in the pelvic cavity. We hypothesize that the models, through delicate and numerical analysis, are extracting features in the MRI that may indicate chronic inflammation, which are not visible to the human eye. This hypothesis necessitates further analysis and studies to establish a connection between the disease pathophysiology and AI image analysis patterns. Figure 4 presents one example each of a correctly predicted positive case and a correctly predicted negative case.

Discussion

In this study, we developed a DL model for the classification of MRIs to detect the presence of endometriosis. Our ensemble techniques, combining output from seven models, demonstrated excellent performance in detecting endometriosis, achieving a sensitivity of 97.6% and an AUROCC of 0.911 on the unseen test set using T2W, T1W FS precontrast, and T1W FS post-contrast images. These results suggest that DL is a promising approach for the automated detection and evaluation of endometriosis in MRI, offering a valuable tool for diagnosing this widespread and complex disease in the pelvic and abdominal area. Our current model does not explicitly locate or segment lesions; however, the generated maps reveal that the model captures small-sized signals within the pelvic area. We hypothesize that these signals may correspond to inflammation, which is commonly present in patients with endometriosis. In contrast, voxel-by-voxel analysis performed by CNN models can

Table 2 Presents the results of the 7-fold cross-validation and the test set evaluations

	T2 (training set size = 596, test set size = 86)				T2 and T1 Pre-Contrast (training set size = 551, test set size = 78)				T1 Pre and Post Contrast (training set size = 521, test set size = 74)				T2, T1 Pre and Post Contrast (training set size = 68, test set size = 474)			
Fold	AUROCC	F1	Sensitivity	Specificity	AUROCC	F1	Sensitivity	Specificity	AUROCC	F1	Sensitivity	Specificity	AUROCC	F1	Sensitivity	Specificity
1	0.791	0.772	0.776	0.549	0.790	0.672	0.667	0.761	0.832	0.800	0.909	0.645	0.912	0.897	0.953	0.800
2	0.921	0.892	0.895	0.661	0.833	0.667	0.714	0.864	0.922	0.880	0.886	0.871	0.814	0.794	0.837	0.720
3	0.703	0.650	0.885	0.772	0.882	0.781	0.769	0.813	0.908	0.824	0.767	0.903	0.843	0.823	0.837	0.800
4	0.781	0.750	0.917	0.776	0.859	0.601	0.615	0.629	0.882	0.864	0.860	0.871	0.955	0.912	0.884	0.960
5	0.802	0.772	0.885	0.442	0.828	0.690	0.681	0.758	0.918	0.905	0.930	0.871	0.888	0.882	0.930	0.800
6	0.840	0.792	0.776	0.791	0.911	0.813	0.834	0.812	0.918	0.864	0.814	0.935	0.834	0.779	0.791	0.760
7	0.828	0.772	0.669	0.648	0.868	0.847	0.850	0.850	0.912	0.864	0.818	0.933	0.807	0.776	0.698	0.917
Mean	0.809	0.771	0.829	0.662	0.853	0.724	0.733	0.783	**0.898**	**0.792**	**0.854**	**0.861**	0.865	0.754	0.847	0.822
STD	0.066	0.070	0.090	0.131	0.039	0.089	0.088	0.079	0.032	0.041	0.058	0.099	0.055	0.055	0.087	0.085
Test Set	0.746	0.704	0.821	0.513	0.856	0.738	0.824	0.713	0.835	0.825	0.909	0.566	**0.911**	**0.881**	**0.976**	**0.720**

area under the ROC curve (AUROCC)

detect subtle inflammation-related signal patterns, offering AI a distinct advantage over human observation.

It is worth noting that a subset of our models using T2W-weighted and T1W FS pre-contrast MRI sequences only achieved a sensitivity of 82.4%, a specificity of 71.3%, and an AUC of 0.856. Although this performance is lower compared to models adding contrast-enhanced sequences, the use of only T2W and precontrast T1W FS sequences presents significant advantages. These sequences do not require contrast agents, reducing both risk and acquisition time, making them an excellent initial choice for screening and AI detection with MRI.

Detecting and accurately localizing endometriosis lesions on MRI requires specialized expertise in abdominal imaging and many years of experience [43]. This level of expertise is necessary due to the subtle and complex nature of the imaging features associated with endometriosis, which can be easily missed or misinterpreted by even seasoned radiologists. Our results showed an overall sensitivity of 84.48% (range 54.5-100%) and a specificity of 69.64% (range 30-85%) for radiologists. When providing radiologists with the AI predictions, the overall sensitivity increased to 87.93% (range 72.7-100%) and specificity to 71.43% (range 65-85%), which is nearly equivalent. However, even within our cohort of experienced radiologists, the AI helped to normalize diagnostic performance. Although lesion-specific detection was not performed, a common pitfall of endometriosis MRI is the threshold of when to diagnose a potential superficial disease or mild peritoneal thickening without pathognomic T1W FS hyperintensity. The normalization of interpretation is a tremendous advantage, particularly when looking at the generalization of the model into non-specialized centers. Additionally, the utility of a test such as this with a high sensitivity to exclude disease may potentially prevent unnecessary interventions such as diagnostic laparoscopy, the current gold standard. Our results showed an increase in the level of agreement among radiologists after learning AI prediction although it showed no statistical significance This slight improvement -although not statistically significant- suggests that AI assistance provided a decision-support tool, enabling radiologists to align their interpretations more closely. By offering standardized and consistent predictions, the AI likely reduced variability in subjective assessments, which is often a challenge in complex cases like endometriosis.

Our results indicated that while the accuracy and sensitivity of radiologists improved with AI assistance, the difference in diagnostic performance was not statistically significant. However, the time required for readings significantly decreased with AI assistance. This reduction in reading time could be biased by the rereading of images, but we adhered to the standard protocol of more than

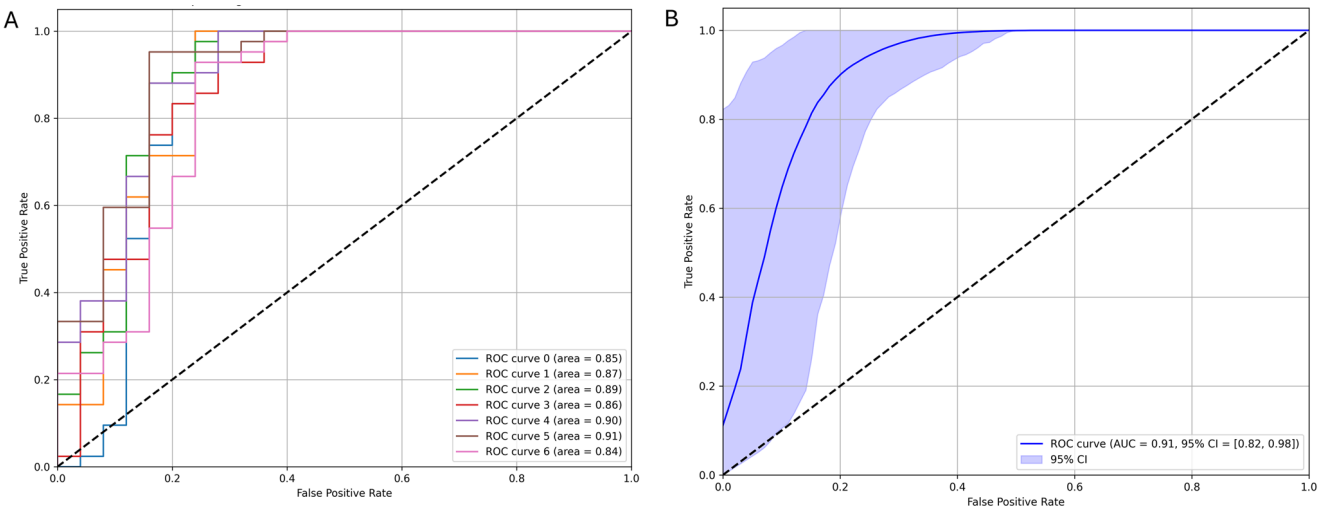


Fig. 2 Illustrates the ROC curve for (**A**) the seven folds and (**B**) the final ensemble results, based on fat-saturated T1W pre- and post-contrast, as well as T2W images. The 95% confidence interval for the AUC values of the final predictions ranges from 82–98%

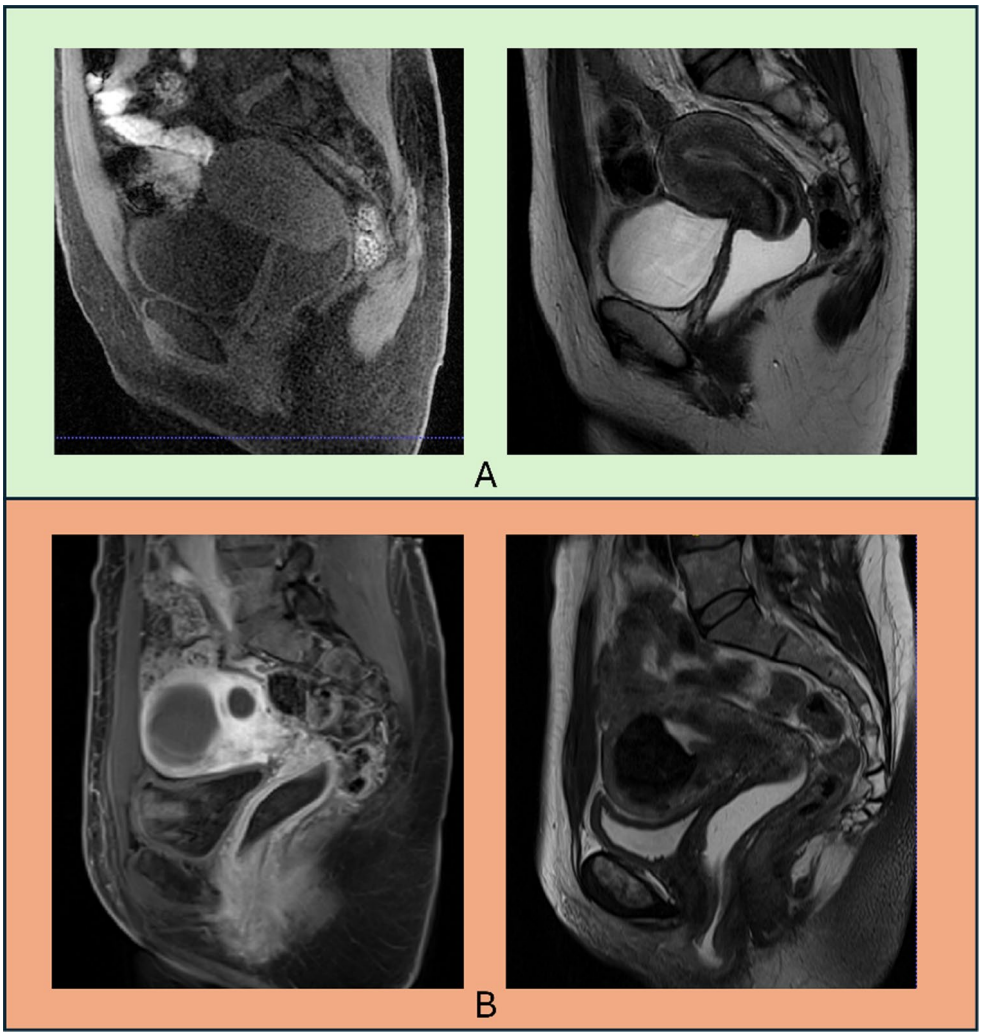


Fig. 3 **A** represents a successful case where the AI correctly identified endometriosis. This was a subtle case that was also incorrectly classified as negative by three radiologists. **B** illustrates a failure case in which the AI misclassified a negative case as endometriosis due to the presence of blood products; pathology later confirmed the absence of endometriosis

Table 3 Presents the readers' results with and without AI assistance. One of the readers only participated in one stage of study with AI assistance

Reader #	Correctly Predicted Cases by Readers Without AI Assistance	Correctly Predicted Cases by Readers with AI Assistance	Reader Case Count Per Session	Mean Time for Reading without AI Assistance per Case(seconds)	Mean Time for Reading with AI Assistance per Case(seconds)	Reader Time Difference pValue
Reader 1	10(100%)	9(90%)	10	690.0	282.0	0.0001> *
Reader 2	7(70%)	9(90%)	10	966.0	348.0	0.0001> *
Reader 3	9(90%)	9(90%)	10	360.0	326.67	0.708
Reader 4	16(80%)	15(75%)	20	715.62	443.76	0.0001> *
Reader 5	16(80%)	17(85%)	20	77.67	68.62	0.1581
Reader 6	9(45%)	18(90%)	20	397.14	214.29	0.0001> *
Reader 7	17(85%)	17(85%)	20	196.14	150.52	0.0051*
Reader 8	15(75%)	-	20	142.86	-	N/A
Mean Correct Reader's Prediction Percent	76.15%	85.45%				0.798
Standard Deviation Correct Reader's Prediction Percent	15.18%	5.12%				
Mean Reading Time per Case				443.18	261.98	0.0001> *
Standard Deviation for Reading Time per Case				296.30	117.79	0.0001> *

one-month washout interval to ensure that the readers did not remember the specific MR characteristics for each case. They confirmed not remembering the details of any cases and accurately read each case as if it were the first time. We acknowledge that a potential limitation is readers recalling cases or rereading, which could reduce reading time. To mitigate this, we implemented a standard washout period. We did not directly compare the AI alone versus radiologists without endometriosis-specific training, but presumably would show measurably improved detection.

Currently, cases of superficial endometriosis have a very low detection rate, even among experienced radiologists. Our endometriosis cohort included superficial lesion cases confirmed by surgery and pathology. The fact that our AI tools demonstrated a higher sensitivity rate compared to the readers suggests that AI *could* be a valuable tool for detecting superficial lesions. One of our future directions is to enhance our model to categorize different types of endometriosis lesions.

To the best of our knowledge, this is the first study to utilize DL models for detecting endometriosis using MRI. A previous study used DL techniques on US videos and achieved 96.5% (95% CI: 90.8–100.0%), sensitivity of 88.6% (95% CI: 83.0-92.9%), specificity of 90.0% (95% CI: 68.3–98.8%) [44]. Another study using feature selection and machine learning on US showed a sensitivity of 0.72 and a specificity of 0.73 on 333 patients [45]. Our multi-sequence model utilized MRI (T1W FS, T2W, and CE-T1W FS sequences), which is not operator-dependent, and demonstrated higher performance in detecting endometriosis than these prior publications. This was validated with a large patient cohort as well as a separate test set, proving the model's robust prediction. The use of multiple MRI sequences allows for comprehensive imaging that enhances the model's accuracy and robustness, making it a reliable tool for clinical application across different centers.

Conclusion

By applying DL to MR imaging, our study opens new avenues for enhancing diagnostic accuracy and efficiency in detecting endometriosis. This approach not only supports radiologists by providing a powerful diagnostic tool but also holds promise for improving patient outcomes through earlier and more reliable detection. Because there is often a very long delay in diagnosis, and because there are significant consequences such as organ damage, chronic pain, infertility and increased risk of malignancy, the high sensitivity presents the possibility of a screening exam that was previously not possible due to limitations in visual perception. Future work will focus on the external validation of the model, including collaboration with other centers using a similar MR protocol. Additionally, we will work on

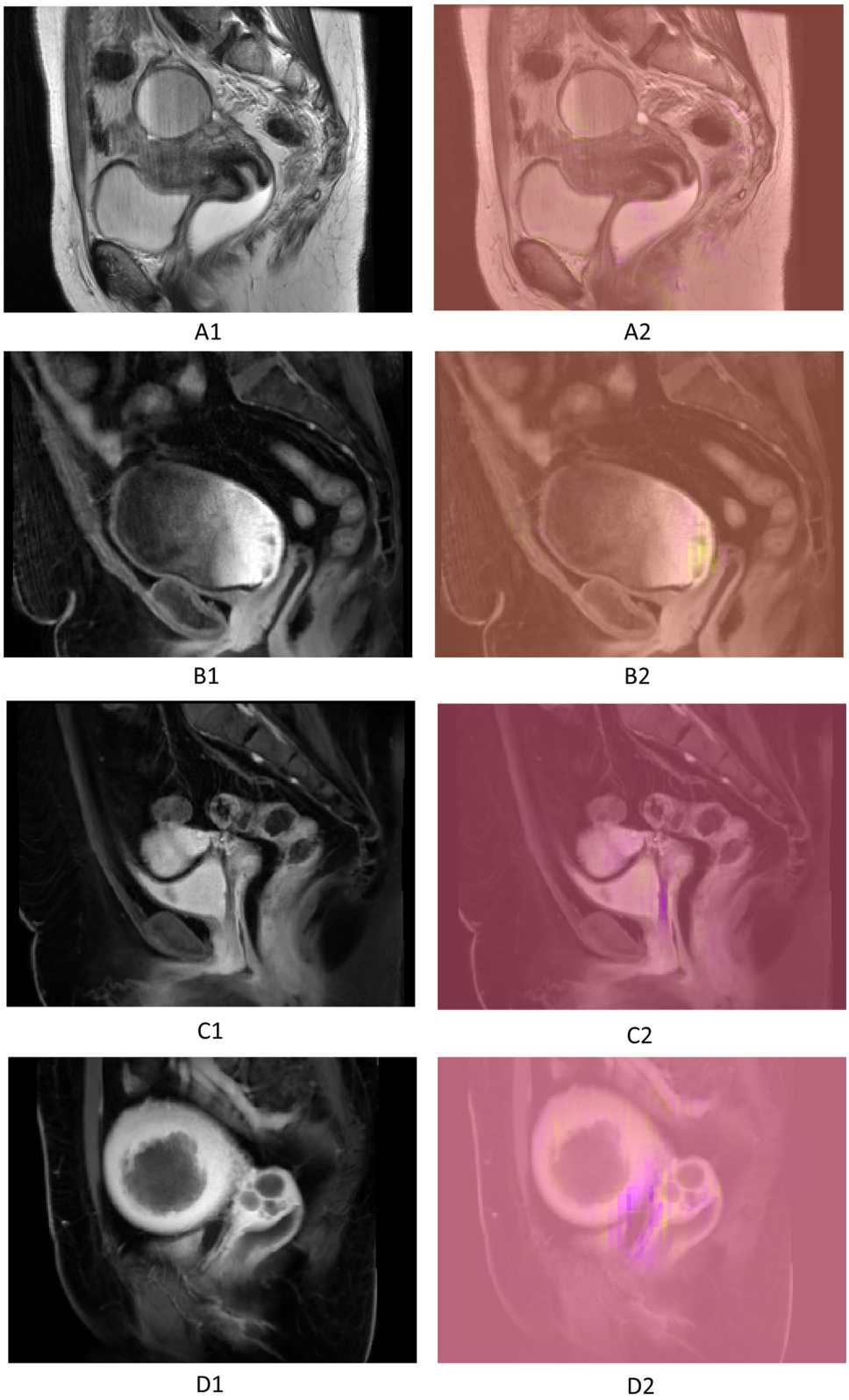


Fig. 4 **A1** and **B1** are sagittal T2W and fat-saturated T1W post-gadolinium slices of two different patients with endometriosis, and **A2** and **B2** are the same slices with overlaid occlusion maps. The model correctly predicted these cases as endometriosis positive. The yellow areas in the maps represent pixels that influenced the model's decision toward diagnosing endometriosis. **C1** and **D1** are sagittal T1W FS post-gadolinium slices of two different patients without endometriosis, and **C2** and **D2** are the same slices with overlaid occlusion maps. The model correctly predicted this case as endometriosis negative. The purple areas in the maps represent pixels that influenced the model's decision toward diagnosing the absence of endometriosis. The focus area can vary since there are different cases in the non-endometriosis cohort of our study

providing more detailed information about the subtypes of endometriosis affecting patients.

Acknowledgements Funding: None.Statement and Declaration: Competing Interest: Dr. VanBuren has received funding from the following societies to support travel for educational lectures at: the American Association of Gynecological Laparoscopists (AAGL), the American Society of Reproductive Medicine (ASRM), the American Roentgen Ray Society (ARRS), the World Endometriosis Congress (WEC) and from the Society for Women's Health Research (SWHR) to advocate for endometriosis on Capitol Hill. She is the founder of the International Endometriosis Imaging Congress.

Author contributions M.M. and S.F. authored the initial draft of the manuscript and revised it based on feedback provided by B.E. and W.V. The development of the AI model was conducted by M.M. and S.F., under the supervision of B.E. Data collection and curation were contributed by C.C., W.V., T.B., and S.G. All other authors participated in the radiological review of the imaging data and the detection of endometriosis. The final manuscript was reviewed and approved by all authors.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Bennett GL, Slywotzky CM, Cantera M, Hecht EM. Unusual manifestations and complications of endometriosis–spectrum of imaging findings: pictorial review. AJR Am J Roentgenol. 2010;194: WS34–46.
2. Lu PY, Ory SJ. Endometriosis: current management. Mayo Clin Proc. 1995;70: 453–463.
3. Heaps JM, Nieberg RK, Berek JS. Malignant neoplasms arising in endometriosis. Obstet Gynecol. 1990;75: 1023–1028.
4. Chamié LP, VanBuren WM, Xiao L, Pires Franco IV, Feldman MK, Causa Andrieu P, et al. Postmenopausal endometriosis: Clinical insights and imaging considerations. Radiographics. 2024;44: e240046.
5. Ozkan S, Murk W, Arici A. Endometriosis and infertility: epidemiology and evidence-based treatments. Ann N Y Acad Sci. 2008;1127: 92–100.
6. Meuleman C, Vandenabeele B, Fieuws S, Spiessens C, Timmerman D, D'Hooghe T. High prevalence of endometriosis in infertile women with normal ovulation and normospermic partners. Fertil Steril. 2009;92: 68–74.
7. Frankel LR. A 10-Year Journey to Diagnosis With Endometriosis: An Autobiographical Case Report. Cureus. 2022;14: e21329.
8. Practice Committee of the American Society for Reproductive Medicine. Endometriosis and infertility: a committee opinion. Fertil Steril. 2012;98: 591–598.
9. Coccia ME, Nardone L, Rizzello F. Endometriosis and Infertility: A Long-Life Approach to Preserve Reproductive Integrity. Int J Environ Res Public Health. 2022;19. doi:https://doi.org/10.3390/ijerph19106162
10. Simoens S, Dunselman G, Dirksen C, Hummelshoj L, Bokor A, Brandes I, et al. The burden of endometriosis: costs and quality of life of women with endometriosis and treated in referral centres. Hum Reprod. 2012;27: 1292–1299.
11. Bourgioti C, Preza O, Panourgias E, Chatoupis K, Antoniou A, Nikolaidou ME, et al. MR imaging of endometriosis: Spectrum of disease. Diagn Interv Imaging. 2017;98: 751–767.
12. Becker CM, Bokor A, Heikinheimo O, Horne A, Jansen F, Kiesel L, et al. ESHRE guideline: endometriosis. Hum Reprod Open. 2022;2022: hoac009.
13. Hirsch M, Begum MR, Paniz É, Barker C, Davis CJ, Duffy J. Diagnosis and management of endometriosis: a systematic review of international and national guidelines. BJOG. 2018;125: 556–564.
14. Agarwal SK, Chapron C, Giudice LC, Laufer MR, Leyland N, Missmer SA, et al. Clinical diagnosis of endometriosis: a call to action. Am J Obstet Gynecol. 2019;220: 354.e1–354.e12.
15. Practice bulletin no. 114: management of endometriosis. Obstet Gynecol. 2010;116: 223–236.
16. Young SW, Jha P, Chamié L, Rodgers S, Kho RM, Horrow MM, et al. Society of Radiologists in Ultrasound Consensus on Routine Pelvic US for Endometriosis. Radiology. 2024;311: e232191.
17. Tong A, VanBuren WM, Chamié L, Feldman M, Hindman N, Huang C, et al. Recommendations for MRI technique in the evaluation of pelvic endometriosis: consensus statement from the Society of Abdominal Radiology endometriosis disease-focused panel. Abdom Radiol (NY). 2020;45: 1569–1586.
18. Guerriero S, Condous G, van den Bosch T, Valentin L, Leone FPG, Van Schoubroeck D, et al. Systematic approach to sonographic evaluation of the pelvis in women with suspected endometriosis, including terms, definitions and measurements: a consensus opinion from the International Deep Endometriosis Analysis (IDEA) group. Ultrasound Obstet Gynecol. 2016;48: 318–332.
19. Bazot M, Daraï E. Diagnosis of deep endometriosis: clinical examination, ultrasonography, magnetic resonance imaging, and other techniques. Fertil Steril. 2017;108: 886–894.
20. Kido A, Himoto Y, Moribata Y, Kurata Y, Nakamoto Y. MRI in the diagnosis of endometriosis and related diseases. Korean J Radiol. 2022;23: 426–445.
21. Tong A, Cope AG, Waters TL, McDonald JS, VanBuren WM. Best practices: Ultrasound versus MRI in the assessment of pelvic endometriosis. AJR Am J Roentgenol. 2024; 1–16.
22. VanBuren W, Feldman M, Shenoy-Bhangle AS, Sakala MD, Young S, Chamie LP, et al. Radiology state-of-the-art review: Endometriosis imaging interpretation and reporting. Radiology. 2024;312: e233482.
23. Murphy AA, Santanam N, Parthasarathy S. Endometriosis: a disease of oxidative stress? Semin Reprod Endocrinol. 1998;16: 263–273.
24. Donnez J, Binda MM, Donnez O, Dolmans M-M. Oxidative stress in the pelvic cavity and its role in the pathogenesis of endometriosis. Fertil Steril. 2016;106: 1011–1017.
25. Van Langendonckt A, Casanas-Roux F, Donnez J. Oxidative stress and peritoneal endometriosis. Fertil Steril. 2002;77: 861–870.
26. Cacciottola L, Donnez J, Dolmans M-M. Can Endometriosis-Related Oxidative Stress Pave the Way for New Treatment Targets? Int J Mol Sci. 2021;22. doi:https://doi.org/10.3390/ijms22137138
27. Defrère S, Van Langendonckt A, Vaesen S, Jouret M, González Ramos R, Gonzalez D, et al. Iron overload enhances epithelial cell proliferation in endometriotic lesions induced in a murine model. Hum Reprod. 2006;21: 2810–2816.
28. Van Langendonckt A, Casanas-Roux F, Dolmans M-M, Donnez J. Potential involvement of hemoglobin and heme in the pathogenesis of peritoneal endometriosis. Fertil Steril. 2002;77: 561–570.
29. Van Langendonckt A, Casanas-Roux F, Eggermont J, Donnez J. Characterization of iron deposition in endometriotic lesions

induced in the nude mouse model. Hum Reprod. 2004;19: 1265–1271.
30. Wyatt J, Fernando SM, Powell SG, Hill CJ, Arshad I, Probert C, et al. The role of iron in the pathogenesis of endometriosis: a systematic review. Hum Reprod Open. 2023;2023: hoad033.
31. Van Langendonckt A, Casanas-Roux F, Donnez J. Iron overload in the peritoneal cavity of women with pelvic endometriosis. Fertil Steril. 2002;78: 712–718.
32. Muñoz M, Villar I, García-Erce JA. An update on iron physiology. World J Gastroenterol. 2009;15: 4617–4626.
33. Wood JC. Use of magnetic resonance imaging to monitor iron overload. Hematol Oncol Clin North Am. 2014;28: 747–64, vii.
34. Amin K, Mileto A, Kolokythas O. MRI for Liver Iron Quantification: Concepts and Current Methods. Semin Ultrasound CT MR. 2022;43: 364–370.
35. Ghugre NR, Wood JC. Relaxivity-iron calibration in hepatic iron overload: probing underlying biophysical mechanisms using a Monte Carlo model. Magn Reson Med. 2011;65: 837–847.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.
37. Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, et al. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. Radiol Artif Intell. 2022;4: e210290.
38. Zhang K, Khosravi B, Vahdati S, Faghani S, Nugen F, Rassoulinejad-Mousavi SM, et al. Mitigating Bias in Radiology Machine Learning: 2. Model Development. Radiol Artif Intell. 2022;4: e220010.
39. Faghani S, Moassefi M, Madhavan AA, Mark IT, Verdoorn JT, Erickson BJ, et al. Identifying Patients with CSF-Venous Fistula Using Brain MRI: A Deep Learning Approach. AJNR Am J Neuroradiol. 2024;45: 439–443.
40. Polak G, Barczyński B, Wertel I, Kwaśniewski W, Bednarek W, Derewianka-Polak M, et al. Disrupted iron metabolism in peritoneal fluid may induce oxidative stress in the peritoneal cavity of women with endometriosis. Ann Agric Environ Med. 2018;25: 587–592.
41. Ganz T. Macrophages and iron metabolism. Microbiol Spectr. 2016;4. doi:https://doi.org/10.1128/microbiolspec.MCHD-0037-2016
42. Allavena G, Carrarelli P, Del Bello B, Luisi S, Petraglia F, Maellaro E. Autophagy is upregulated in ovarian endometriosis: a possible interplay with p53 and heme oxygenase-1. Fertil Steril. 2015;103: 1244–51.e1.
43. Bruyere C, Maniou I, Habre C, Kalovidouri A, Pluchino N, Montet X, et al. Pelvic MRI for Endometriosis: A Diagnostic Challenge for the Inexperienced Radiologist. How Much Experience Is Enough? Acad Radiol. 2021;28: 345–353.
44. Maicas G, Leonardi M, Avery J, Panuccio C, Carneiro G, Hull ML, et al. Deep learning to diagnose pouch of Douglas obliteration with ultrasound sliding sign. Reprod Fertil Dev. 2021;2: 236–243.
45. Guerriero S, Pascual M, Ajossa S, Neri M, Musa E, Graupera B, et al. Artificial intelligence (AI) in the detection of rectosigmoid deep endometriosis. Eur J Obstet Gynecol Reprod Biol. 2021;261: 29–33.
46. Ganjizadeh A, Zawada S, Langer S, et al. Visualizing Clinical Data Retrieval and Curation in Multimodal Healthcare AI Research: A Technical Note on RIL-workflow. J Digit Imaging Inform Med. 2024;37(6):1239–1247. https://doi.org/10.1007/s10278-024-00977-3
47. Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, et al. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. Radiol Artif Intell. 2022;4: e210290.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.