JNMS Journal of Nippon Medical School https://www.nms.ac.jp/sh/jnms/

Original

Development and Clinical Application of a Deep Learning–Based AI Support Model for Endometrial Cancer Cytology

Ichito Shimokawa[1,2], Mika Terasaki[1], Shun Tanaka[1,3], Etsuko Toda[1,4], Shoichiro Takakuma[1], Yusuke Kajimoto[5], Shinobu Kunugi[1], Akira Shimizu[1] and Yasuhiro Terasaki[6]

[1]Department of Analytic Human Pathology, Nippon Medical School, Tokyo, Japan
[2]Faculty of Medicine, Nippon Medical School, Tokyo, Japan
[3]Jichi Medical University Saitama Medical Center, Saitama, Japan
[4]Laboratory for Morphological and Biomolecular Imaging, Nippon Medical School, Tokyo, Japan
[5]Division of Pathology, Nippon Medical School Musashikosugi Hospital, Kanagawa, Japan
[6]Division of Pathology, Nippon Medical School Hospital, Tokyo, Japan

Background: The global increase in endometrial cancer, including in Japan, and a shortage of pathologists and cytotechnologists have increased the diagnostic burden, emphasizing the need for an AI-based diagnostic support model that uses deep learning. We evaluated the clinical application of an improved AI-supported endometrial cytology model.

Methods: Using YOLOv5x and YOLOv7 models evaluated by mean average precision (mAP), we compared two datasets—one annotated for both benign and malignant cell clusters, and one for malignant only. In addition, using the Two One-Sided Tests (TOST) procedure, we assessed the correlation between AI diagnostic accuracy and the level of difficulty perceived by human diagnosticians. Finally, we used Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize and enhance the interpretability of the AI model's decision-making process.

Results: The YOLOv5x model with both benign and malignant annotations had the highest malignant mAP, 0.798, as compared with YOLOv7. The TOST analysis showed no significant difference in perceived diagnostic difficulty between cases that were correctly and incorrectly diagnosed by the AI model, indicating consistent AI accuracy regardless of case difficulty. Grad-CAM visualizations clarified the AI model's decision-making basis; in some cases, the model appeared to focus on regions that differed from those typically attended to by human diagnosticians.

Conclusion: The AI support model showed high and consistent accuracy in endometrial cytological analysis, regardless of diagnostic difficulty as perceived by human diagnosticians. Grad-CAM visualizations revealed diagnostic patterns, and the AI occasionally focused on regions different from those emphasized by human diagnosticians. This study advanced a real-time microscope-integrated AI system toward clinical application.

(J Nippon Med Sch 2026; 93 (1): 80–94. https://doi.org/10.1272/jnms.JNMS.2026_93-115)

Keywords: artificial intelligence, endometrial cancer, YOLO networks, deep learning-based object detection algorithms, cytology

Correspondence to Mika Terasaki, mterasaki@nms.ac.jp
https://doi.org/10.1272/jnms.JNMS.2026_93-115
Received: November 4, 2025; Accepted: December 22, 2025

Introduction

Incidence rates for endometrial cancer, a gynecological malignancy, are increasing worldwide[1] and particularly in developed countries, where it is most prevalent. Most cases occur in postmenopausal women; mean age is 60 years at presentation. Early detection and appropriate treatment have a profound impact on patient survival and quality of life. The 5-year survival rate exceeds 90% for stage I endometrial cancer but drops to approximately 70% for stage III disease[2]. This highlights the importance of early diagnosis and the need for rapid and accurate diagnostic methods.

Endometrial cytology is a crucial diagnostic method for early detection of endometrial cancer[3]. Because abnormal uterine bleeding is often the first symptom, cytology should be performed promptly when such bleeding is observed[4]. This test is widely used because it can be easily performed in an outpatient setting, with minimal burden on the patient. In endometrial cytology, cells directly collected from the endometrium with a specialized instrument are examined under a microscope to detect cancer cells at an early stage. However, this test presents certain diagnostic challenges. The morphology of cells can easily change under the influence of hormones[5], and the thickness and overlap of cell clusters can make diagnosis difficult, requiring advanced expertise and technical skills.

Deep learning, a subset of AI, has gained significant attention in recent years for its potential in multiple fields, particularly in medical image analysis. It has been used for detecting lung cancer in chest X-ray images[6], colon cancer[7] in endoscopic live images, and diabetic retinopathy[8] in retinal images. The potential for AI to streamline workflow and support decision-making extends beyond specific diseases, potentially benefiting a wide range of clinical applications, from early screening to complex case management. Deep learning has also been used in various fields in pathology[9]; however, applying it specifically to endometrial cytology presents unique challenges. One such challenge is obtaining clear digital images of tissue samples, known as Whole-Slide Imaging (WSI)[10]. In endometrial histology, WSI has been successfully applied to AI models for diagnostic purposes; however, cytology faces additional challenges due to the thickness and overlap of cell clusters[11]. WSI involves scanning entire glass slides to create high-resolution digital images, but achieving consistent focus across thick or overlapping cell clusters can be difficult when dealing with directly smeared cells on the glass slide, as opposed to uniformly thin sections from formalin-fixed, paraffin-embedded specimens. This inherent complexity has slowed the development of AI models in endometrial cytology.

To address challenges in obtaining digital images from endometrial cytological slides, we developed an AI-assisted system for detecting endometrial cancer cells. The system operates in real-time under a microscope without requiring WSI scans, as detailed in our previous research[12]. We used a pre-existing deep learning model known as You Only Look Once version 5x (YOLOv5x), which is specifically designed for rapid object detection, and customized the model to suit our specific dataset. This system is designed to seamlessly integrate with existing microscope workflows, allowing clinicians to verify AI-generated assessments as they view the slides. Our earlier studies demonstrated that this system could significantly reduce diagnostic time across evaluators with varying levels of experience, from skilled pathologists to medical students, while also alleviating the workload of diagnosticians and improving diagnostic accuracy.

Our previous research demonstrated the usefulness of an AI-assisted system in endometrial cytology but also revealed several significant limitations. Training an AI model typically requires a large amount of labeled data to teach it what to recognize. For instance, to enable an AI to differentiate cancerous from normal cells, regions containing cancer cells are labeled “malignant”, while the labeling strategy for normal cell clusters is often not well defined. Moreover, it remains unclear which annotation approach yields the best performance. This uncertainty affects the reliability of AI decisions and warrants further investigation.

Next, we chose to use the YOLOv5x model. This model is well-suited for our task because it can rapidly identify specific areas in microscope images where cancerous cells might be present. However, we have not yet compared the performance of YOLOv5x with that of other AI models to determine if it is truly the best option for this type of work. Evaluating and comparing different models is important because it helps ensure that we are using the most effective tool for accurately detecting cancer cells. This is an issue that warrants further investigation in future research.

Many studies of medical AI models emphasize quantitative metrics such as object or abnormality detection accuracy. While such metrics are useful for evaluating technical performance, they often fail to capture the true clinical value of AI systems. One method of addressing this limitation is to compare the diagnostic performance

of AI-assisted and AI-unassisted workflows to assess clinical applicability[13]. In our previous research, we responded to this need by conducting a comparative analysis of diagnostic performance with and without AI support[12]. However, we did not examine the relationship between diagnosticians' perceived case difficulty and whether the AI model made a correct or incorrect prediction. This omission may have limited our understanding of the model's practical utility in real-world settings, and ignoring this aspect could lead to a misleading interpretation of its usefulness in clinical practice.

Additionally, AI models that use deep learning face what is known as the "black-box problem"[14]. Deep learning–based models consist of multiple computational layers that automatically extract hierarchical features from the input data. While this allows for high performance, the depth and complexity of these layers make it difficult to understand the decision-making process. This opacity, termed the black-box problem, highlights the challenge of interpreting how the AI model arrives at its diagnostic conclusions. If we cannot identify the aspects that the AI focused on and why it reached a particular conclusion, it raises concerns about the reliability of the diagnosis. To address this issue, it is essential to develop "explainable AI (XAI)" that can clarify how the AI makes its decisions.

Using a technique called Gradient-weighted Class Activation Mapping (Grad-CAM), which visualizes the regions that most strongly influence the model's output by utilizing the gradients of the final convolutional layer, we can visualize which parts of an image the AI focused on during a specific diagnosis[15]. This allows us to visually confirm how the AI arrived at its decision, making the rationale behind the diagnosis more understandable. It is crucial to utilize such techniques to clarify the criteria that an AI uses to diagnose benign and malignant cell clusters.

To address the limitations identified in our previous research, this study had four key objectives:

1. Annotation Strategy: We investigated whether labeling both benign and malignant cell clusters or malignant cell clusters only provided better diagnostic accuracy. This helps us determine the most effective annotation strategy for training AI models.
2. Model Comparison: We compared the performance of the YOLOv5x model with other AI models to determine which is most effective for detecting cancer cells in endometrial cytology. This comparison is crucial to ensure we are using the best possible tool.
3. Integrating the AI Model into Routine Diagnostic Practice: We examined the correlation between the difficulty of diagnosis, as perceived by pathologists and cytotechnologists, and the accuracy of the AI model. This analysis helps us align AI models with the practical challenges faced in clinical settings.
4. Explainability and Transparency: We explored the use of techniques such as Grad-CAM to enhance the explainability of AI models. By visualizing the AI's decision-making process, we can improve the transparency and trustworthiness of AI-assisted diagnoses in clinical practice.

These objectives will help in refining our AI-assisted system for endometrial cytology by enhancing its accuracy, reliability, and clinical suitability. Moreover, these areas of investigation hold significance beyond this model, as they support the broader development of AI applications for medical imaging and healthcare innovation.

Methods

Ethics

The study was conducted using an opt-out method for informed consent and was approved by the Institutional Review Board of Nippon Medical School (approval number: O-2023-742).

Patient Data Collection

Patient selection and cytology preparation followed the methods described in our previous study[12], which detailed the collection and classification of endometrial cytology cases.

From April 2017 to March 2023, a total of 96 cytology slides were collected at Nippon Medical School Hospital: 47 cases were diagnosed as having "Malignant" endometrial cancer, and the remaining 49 cases were classified as "Benign", including non-malignant endometrial conditions such as leiomyoma. **Table 1** shows the case distribution and median age for each category. All diagnoses were pathologically confirmed with hysterectomy specimens. The endometrial cytology samples were prepared by smearing and stained using the standard Papanicolaou method. Specifically, the protocol included nuclear staining with Carazzi's Hematoxylin for 1 minute, followed by cytoplasmic staining with OG-6 for 2 minutes and EA-50 (containing <0.1% Light Green) for 4 minutes. Dehydration and clearing were performed according to standard procedures.

Table 1 Characteristics of Cases

	Object detection for images
	Training, validation, and test cases (n=96)
Malignant	
Median age (range)	57 (31–82)
Number of cases	47
Endometrioid carcinoma	
Grade 1	24
Grade 2	17
Grade 3	5
Serous carcinoma	1
Benign	
Median age (range)	47 (37–73)
Number of cases	49
Leiomyoma	49

Table 2 Number of malignant and benign images in each dataset

Dataset	Training Data	Validation Data	Test Data
Malignant images	1,234	154	148
Benign images	1,819	231	228
total	3,053	385	376

Acquisition of Digital Images with a Smartphone Diagnostic Device

Digital images were acquired with a smartphone-based system, as described in our earlier work[12]. The procedures, including device specifications and imaging conditions, are detailed in that report.

Digital images were obtained with a smartphone-enabled diagnostic setup. An iPhone SE (Apple Inc., Cupertino, CA, USA) was attached to an Olympus BX53 microscope (EVIDENT, Olympus, Tokyo, Japan) via a specialized adapter (i-NTER LENS; Micronet Co., Ltd., Kawaguchi, Saitama, Japan). The images were captured at a resolution of 4,032 × 3,024 pixels. The focus was manually adjusted during direct observation of the cytology slides. The fine adjustment knob was used to achieve the best possible clarity across the cell cluster, particularly focusing on diagnostically important features such as nuclear chromatin patterns and irregularities. Imaging was performed with a 20× objective lens. For malignant cases, images were centered on clusters of abnormal cells identified by a gynecological pathologist. In benign cases, normal cell clusters were randomly selected for imaging.

Creation of Image Datasets

For deep learning model training, 1,536 malignant cell cluster images and 2,278 benign cell cluster images were divided into training data, validation data, and test data in an 8:1:1 ratio. Crucially, this data splitting was performed at the patient (case) level. All images derived from a single patient were strictly assigned to only one of the three datasets, to prevent data leakage and ensure that the AI model was evaluated on completely unseen cases. Consequently, there is no overlap of patients between the training and test datasets. The breakdown is shown in **Table 2**.

Image Preprocessing and Annotation

The collected images, originally 4,032 × 3,024 pixels, were resized to 800 × 600 pixels for training. Annotation was performed using the Python-based software LabelImg (version 1.8.6). Bounding boxes were drawn as rectangles to encompass the entire extent of each cell cluster, which was identified based on standard cytological morphology without a strict numerical cell count cutoff. Handling of overlapping clusters was performed based on visual separability: distinctly separable clusters were annotated individually, while heavily overlapping or aggregated

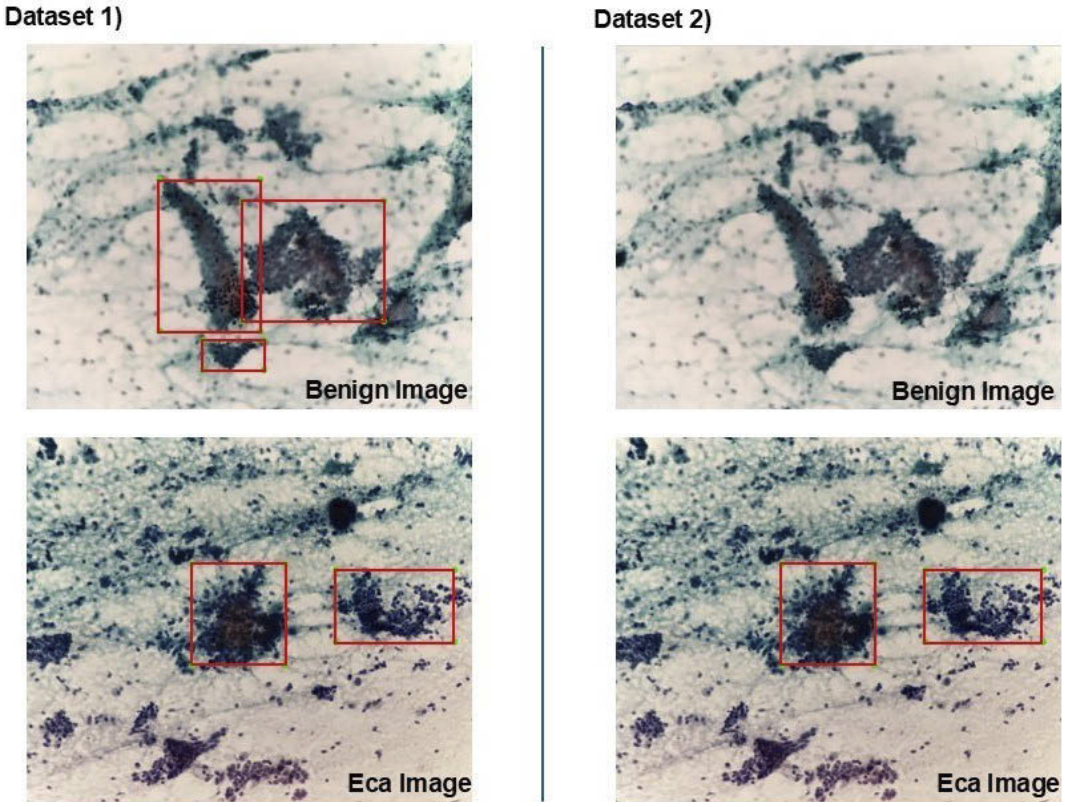

Figure 1 The process of creating datasets
Rectangular bounding boxes were drawn to encompass cell clusters based on their morphological extent. In Dataset 1, both benign and malignant cell clusters are annotated. In Dataset 2, only malignant cell clusters are annotated.

clusters were enclosed in a single bounding box to preserve the structural context. The annotation workflow followed a two-step verification process to ensure accuracy. First, initial annotations were performed by a medical student who had received intensive training in endometrial cytology from a board-certified pathologist. Subsequently, every annotated image was reviewed and validated by a board-certified pathologist. Any discrepancies or ambiguously annotated clusters were corrected by the pathologist, whose judgment served as the final ground truth.

Two datasets were created: one where only malignant cell clusters were annotated, treating benign cell clusters as background, and another where both malignant and benign cell clusters were annotated. This process is illustrated in **Figure 1**.

Models and Their Architecture

Because of its high accuracy and real-time performance, the YOLO series is widely used in medical image analysis for object detection tasks[16]. Specifically, YOLOv5x and YOLOv7 are suited for clinical environments that require high precision and fast processing capabilities for real-time diagnosis.

In medical image analysis, rapid and accurate diagnosis are essential, and the high frame rate capability of YOLO series is crucial[17]. These models efficiently process high-resolution images, reducing diagnostic delays. Moreover, YOLOv5x and YOLOv7 offer advanced feature extraction to detect complex and subtle abnormalities, which is crucial for medical images. This enables precise detection of small and ambiguous boundary lesions that might otherwise go unnoticed. Therefore, this study compared the accuracy of YOLOv5x and YOLOv7.

YOLOv5x provides high performance and flexibility in object detection. Its main features include a three-layer structure (backbone, neck, and head) that enables it to build robust models for a wide range of applications. It excels in real-time, high-precision object detection required in medical image analysis. Additionally, YOLOv5x achieves a lightweight model design and faster processing speed, making it effective for use on edge devices and in resource-limited environments.

YOLOv7, an improved version of the YOLOv5x series, utilizes Extended Efficient Layer Aggregation Networks (E-ELAN) to enhance feature learning ability without

changing the gradient propagation path and to improve parameter and computational efficiency[18].

The selected YOLO models were pretrained on the Microsoft COCO dataset and fine-tuned using our medical dataset through transfer learning. By transfer learning these pretrained models, the YOLO series can adapt effectively to the specific challenges of cytology images in medical image analysis.

Deep Learning Model Training

We adopted two pretrained models, YOLOv5x and YOLOv7, and experiments were conducted over 100 epochs with a batch size of four, and hyperparameters were manually tuned. To ensure reproducibility, a fixed random seed was applied, and a single training run was performed for each model configuration. Standard data augmentation techniques inherent to the YOLO framework, such as Mosaic augmentation and HSV color-space adjustments, were used during training to enhance model robustness. For full reproducibility, detailed hyperparameters and training configurations specific to the best-performing model (YOLOv5x) are comprehensively summarized in **Supplementary Table 1**.

The primary evaluation metric used to assess AI model performance in this study was mean average precision (mAP). mAP is widely used to assess model accuracy, especially in classification and object detection tasks, by calculating the average precision (AP) for each class and taking the mean of these values. AP evaluates the relationship between precision and recall, thus providing a comprehensive accuracy metric independent of specific thresholds.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

AP is obtained by plotting the precision-recall curve and calculating its area under the curve (AUC). Specifically, precision and recall are calculated for different thresholds for each class, and the area under the precision-recall curve is computed as AP. The mAP is obtained by averaging the AP for all classes.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

where N is the total number of classes, and AP_i is the average precision for class i. mAP is a comprehensive metric that evaluates a model's prediction accuracy across all classes. It is especially useful in multi-class classification and object detection tasks. High mAP values indicate that the model balances high precision and recall, demonstrating its utility as a reliable diagnostic support tool.

Execution Environment

An Iiyama Sense 15F161 notebook PC with an Intel Core i7 CPU and an NVIDIA GeForce RTX3060 GPU was utilized. The software environment consisted of the Anaconda distribution (version 2022.10) of Python 3.9.16 as the programming language and PyTorch 1.13.1 as the deep learning framework.

Comparison with Perceived Diagnostic Difficulty by Cytotechnologists and Pathologists

As shown in **Table 2**, the created dataset was divided into training, validation, and test data in an 8:1:1 ratio. During the model training of YOLOv5x and YOLOv7, training data were used for training, and validation data were used for inference regarding model accuracy metrics. Finally, test data were read by YOLOv5x to investigate the correlation between the diagnosticians' perceived difficulty and the AI model's accuracy in determining malignant and benign cases. This was because the YOLOv5x model produced the best results, as shown below. From the test set, 20 images were extracted—five each from the following four categories: (1) benign cases misclassified as malignant, (2) benign cases correctly classified, (3) malignant cases misclassified as benign, and (4) malignant cases correctly classified. These were grouped into two sets: 10 images correctly classified by the AI and 10 misclassified.

Eight cytotechnologists and three pathologists who did not participate in image capture or model creation evaluated the difficulty of each image on a scale of 1 (easiest) to 5 (most difficult). The evaluators were not informed of whether the images were benign or malignant. This process is summarized in the flowchart in **Figure 2**.

Calculation of Mean Difficulty Scores and Box Plot Creation

To quantify the perceived difficulty of each image, we calculated the mean difficulty scores provided by the diagnosticians for correctly and incorrectly classified image groups. The results were visualized using box plots, which illustrate the central tendency and variability of the difficulty scores in each category. This analysis aimed to investigate whether cases misclassified by the AI were

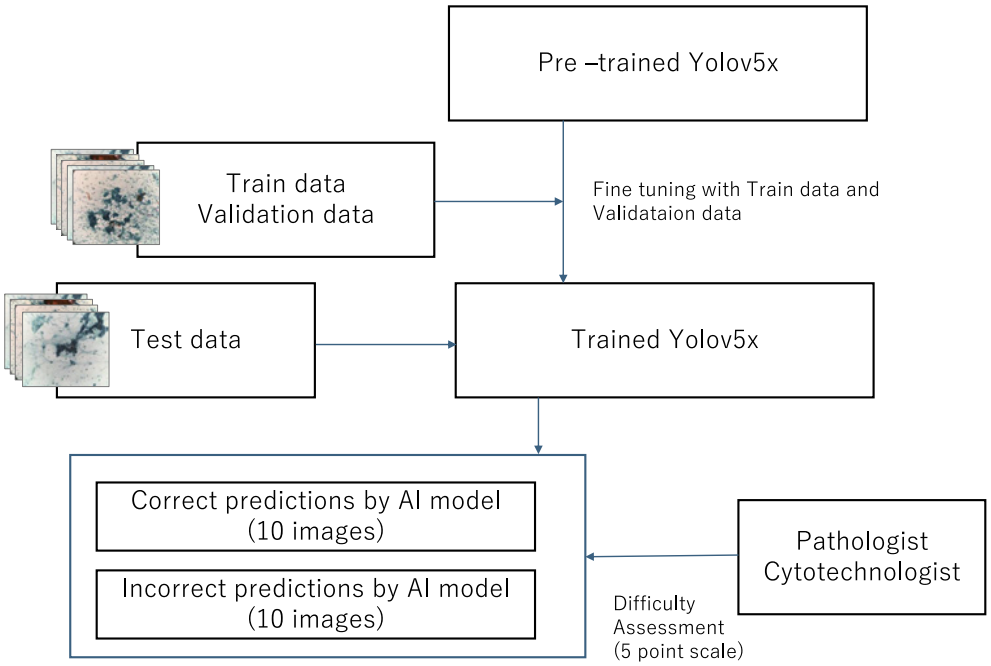

Figure 2 Flowchart of model training, testing, and evaluation of diagnostic difficulty perceived by pathologists and cytotechnologists

systematically associated with higher perceived difficulty by human diagnosticians.

Statistical Analysis

To examine whether the difficulty scores for images correctly classified and misclassified by AI were equivalent, we conducted an equivalence test (two one-sided tests, TOST)[19]. The equivalence range was set at ±0.5. This value was selected based on the 5-point Likert scale used for difficulty rating; a difference of less than 0.5 units was considered clinically negligible as it is less than half of a single scale step and within the expected variability of subjective human assessment. The significance level was set at 0.05 for each of the two one-sided tests, which corresponds to calculating a 90% confidence interval for the mean difference. The standard error (SE) of the sample difference is calculated as follows:

$$SE = \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

where s_1 and s_2 are the standard deviations of each group. The t-values for the lower and upper bounds are calculated using:

$$t_{\text{lower}} = \frac{\bar{d} - (-\Delta)}{SE}$$

$$t_{\text{upper}} = \frac{\bar{d} - \Delta}{SE}$$

The cumulative distribution function is then used to determine the p-values. If the p-values are below 0.05, the two groups are considered statistically equivalent within the specified range.

Visualization of Diagnostic Basis for Benign and Malignant Diagnosis by YOLOv5x

In this study, a model combining YOLOv5x and Grad-CAM was used to visualize the diagnostic basis of the AI model. Grad-CAM is a technique that visualizes the areas of interest in the model for a specific input image, helping to identify the parts of the image the model focuses on. Specifically, Grad-CAM was applied to the final convolutional layer of the YOLOv5x model to reveal the regions of the image that the model focused on when rendering a diagnosis of benign or malignant. The images generated by this technique highlight the model's reasoning for specific diagnostic outcomes. The heatmaps created by Grad-CAM emphasize key regions in the image, clarifying the features the model relied on to make its diagnosis. For example, if the model diagnosed malignancy, the regions contributing to this decision, such as abnormal cell clusters, are shown in high-intensity colors like red or yellow. Conversely, if the model rendered a diagnosis of benign, normal tissue structures or non-abnormal cell clusters are highlighted.

Table 3 The mAP of models of YOLOv5x and YOLOv7

		Benign and Eca annot	Only Eca annot
YOLOv5x	Eca	0.798	0.769
	Benign	0.676	
YOLOv7	Eca	0.781	0.647
	Benign	0.663	

mAP, mean Average Precision (IoU threshold = 0.5); Eca, Endometrial Cancer; IoU, Intersection over Union.

Results

Training and Accuracy of the Benign and Malignant Mixed Models

Two types of training models were developed using both YOLOv5x and YOLOv7: one where both malignant and benign cell clusters were annotated, and another where only malignant cell clusters were annotated. The accuracy metrics for these models are as follows:

- YOLOv5x (mixed annotation): A malignant mAP of 0.798 and a benign mAP of 0.676.
- YOLOv7 (mixed annotation): A malignant mAP of 0.781 and a benign mAP of 0.663.
- YOLOv5x (malignant-only annotation): A malignant mAP = 0.769.
- YOLOv7 (malignant-only annotation): A malignant mAP = 0.647.

These results are summarized in **Table 3**.

In this study, the YOLOv5x model with mixed benign and malignant annotations achieved the highest accuracy; the mAP was 0.798 in detecting malignant cell clusters. YOLOv5x outperformed YOLOv7 in the "Benign and Malignant" (mAP 0.798 vs. 0.781) and "Malignant-only" (mAP 0.798 vs. 0.647) annotation settings, demonstrating its superior performance in identifying malignant cell clusters (**Supplementary Figures 1** and **2**). **Figure 3** illustrates representative detection outputs from each model. **Figure 3** illustrates representative detection outputs from each model. To provide a more comprehensive clinical evaluation beyond mAP, we further assessed the diagnostic performance of the best-performing model, YOLOv5x (mixed annotation), on the test dataset. The model achieved a recall of 0.820, a precision of 0.653, and an F1-score of 0.727 for malignant cell cluster detection. These metrics are summarized in **Table 4**.

Distribution of Perceived Diagnostic Difficulty in AI Model Correct and Incorrect Groups

In this study, 20 images—10 correctly diagnosed by the AI model (Correct Group) and 10 incorrectly diagnosed (Incorrect Group)—were evaluated by three pathologists and eight cytotechnologists.

The median difficulty score for the Correct Group was 2.775, with an interquartile range (IQR) of 0.570, while the Incorrect Group had a median score of 2.910 and an IQR of 0.503. These distributions are shown in the box plot in **Figure 4a**, explaining how diagnosticians perceived the difficulty of cases that were classified correctly or incorrectly by the AI model.

Equivalence Test (TOST) for Difficulty Assessment in AI Model Correct and Incorrect Groups

The results of the equivalence test (TOST) showed a lower bound p-value of 0.044, an upper bound p-value of 0.009, and an overall equivalence p-value of 0.044. The mean difference in difficulty scores was −0.090, with a 90% confidence interval of [−0.485, 0.305]. These results indicate that the two groups are statistically equivalent ($p < 0.05$), as shown in **Figure 4b**. These statistical tests suggest that there is no significant variation in diagnostic difficulty between the Correct and Incorrect Groups, supporting the potential applicability of the AI model in clinical cytology settings.

Visualization of AI Model Diagnostic Basis Using Grad-CAM

Grad-CAM was used to visualize the AI model's diagnostic reasoning by highlighting the regions that the model focused on during its prediction. The resulting heatmaps represent the level of attention or importance assigned by the AI model, with warmer colors such as red indicating higher relevance and cooler colors like blue indicating lower relevance. This visualization helps cytology specialists and cytotechnologists identify the features used by the AI for diagnosis.

Figure 5 presents examples of Grad-CAM visualizations for both correct and incorrect identifications of benign and malignant cell clusters. In several cases, the AI model appeared to focus on regions not typically emphasized by human diagnosticians, eg, areas lacking prominent nuclear atypia or irregular cluster margins, which are often key features for human interpretation. Instead, the model sometimes extracted features from the central portion of loosely aggregated cell clusters or from peripheral cytoplasmic regions.

These observations suggest that the AI may be recognizing patterns not conventionally prioritized by experts, raising important considerations regarding the interpretability, reliability, and complementary potential of AI in

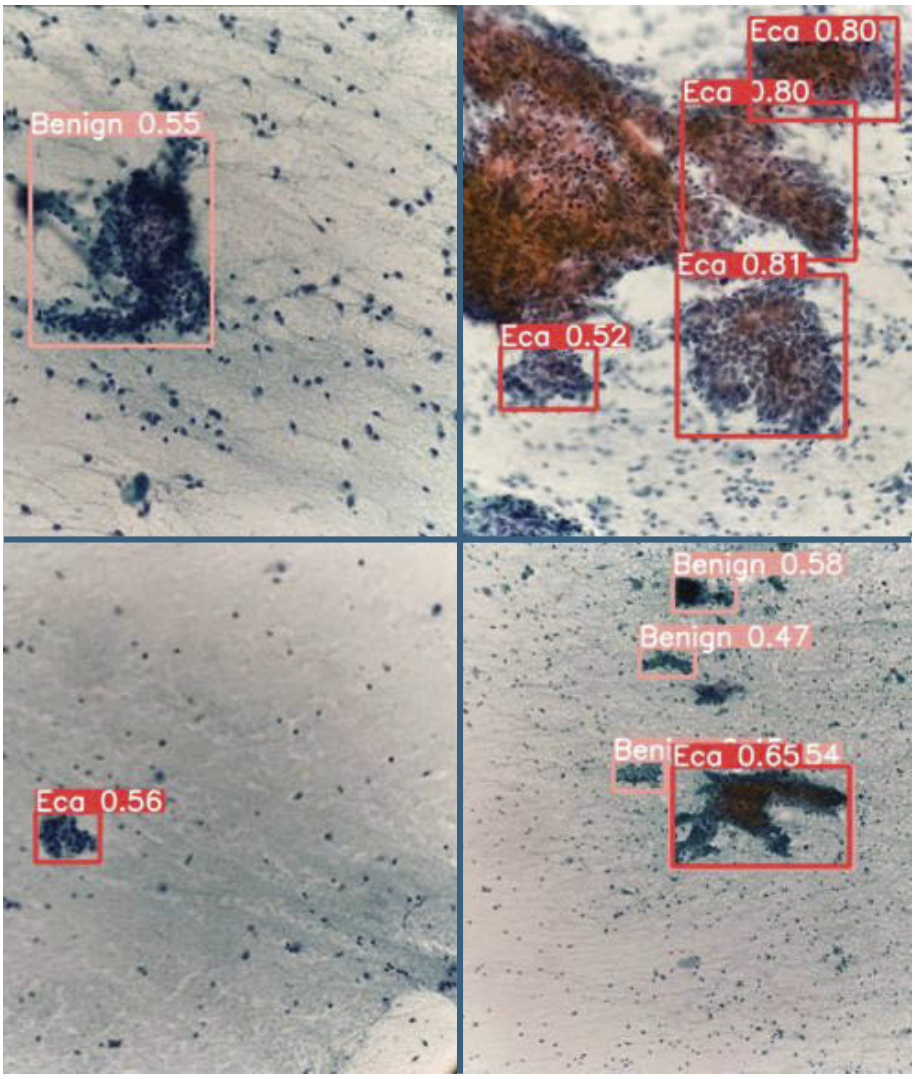

Figure 3 Detection results from the model trained on both benign and malignant cases
The top left shows an image correctly identified as benign, the top right shows an image correctly identified as malignant, the bottom left shows an image incorrectly identified as malignant, and the bottom right shows an image incorrectly identified as benign.

Table 4 Diagnostic performance metrics of the best-performing model (YOLOv5x with mixed annotation) on the test dataset

Class	Precision	Recall	F1-score
Eca	0.653	0.820	0.727
Benign	0.810	0.491	0.611

F1-score is calculated as the harmonic mean of precision and recall.
Eca, endometrial cancer.

cytological diagnosis. This discrepancy between AI attention and human reasoning warrants further investigation and is discussed in detail below.

Discussion

Summary of Results

In this study, we developed and evaluated AI models focusing on four key objectives: (1) optimizing the annotation strategy, (2) comparing model performance, (3) integrating the AI model into routine diagnostic workflows, and (4) enhancing model explainability and transparency. Our findings address technical and clinical aspects of endometrial cytology AI implementation.

Annotation of Benign and Malignant Cell Clusters

In histological studies utilizing WSI, it is standard practice to annotate both benign and malignant cell clusters[20]. In contrast, our prior cytological research focused exclusively on annotating malignant cell clusters, operating under the hypothesis that including benign cell clusters

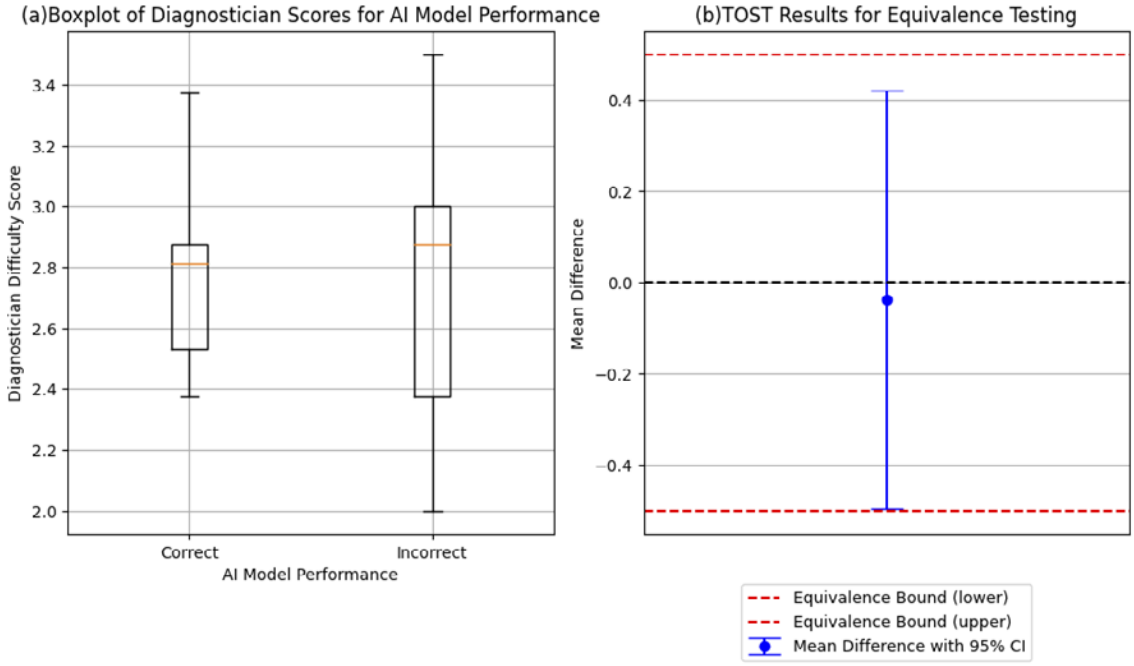

Figure 4 Diagnostician scores for AI model performance and TOST results for equivalence testing
(a) Box plot of difficulty ratings given by diagnostic professionals for each AI model correctness. (b) TOST results showing equivalence between the difficulty assessments of correctly and incorrectly diagnosed images.

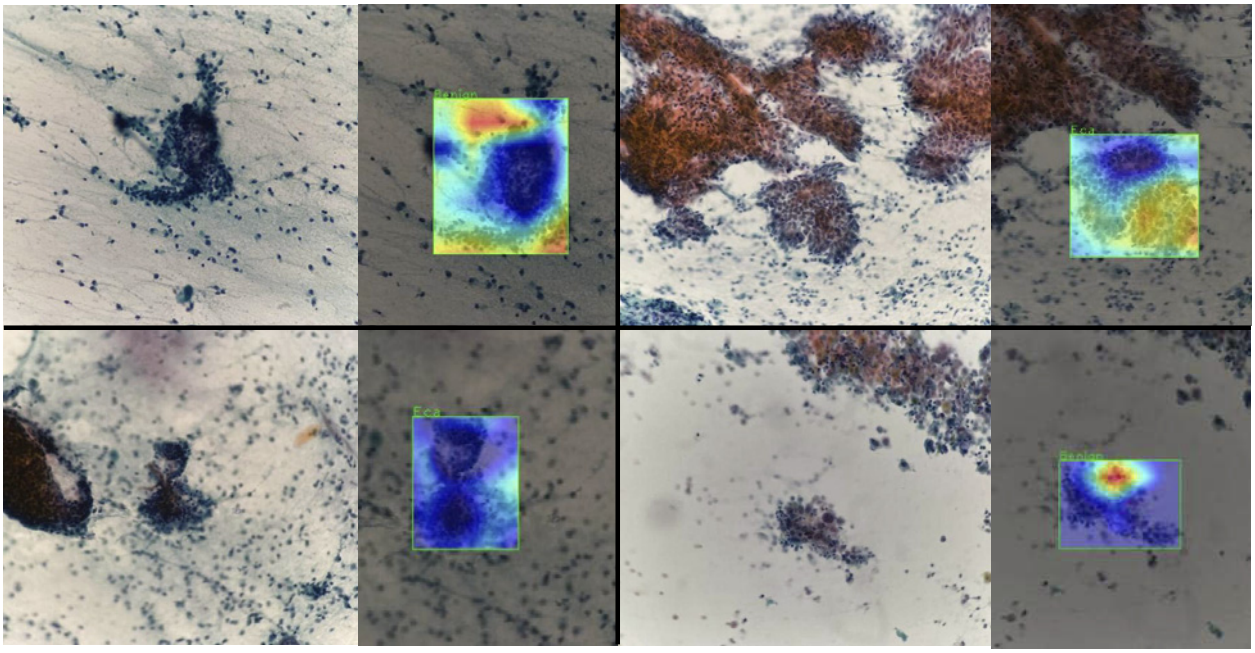

Figure 5 Results of Grad-CAM visualization
The top left image correctly identifies benign cell clusters as benign (True Negative). The top right image correctly identifies malignant cell clusters as malignant (True Positive). The bottom left image incorrectly detects benign cell clusters as malignant (False Positive). The bottom right image incorrectly detects malignant cell clusters as benign (False Negative).

would introduce noise, thereby adversely affecting the detection of clinically significant malignant cells. To test this hypothesis, we conducted an experiment comparing both annotation strategies, initially presuming that annotating only malignant cell clusters would enhance detection accuracy by reducing noise interference. Contrary to our expectations, the results revealed that annotating both benign and malignant cell clusters significantly improved the model's ability to detect malignant cells. Specifically, the YOLOv5x model achieved a mAP of 0.798

when both benign and malignant cell clusters were annotated, as compared with a mAP of 0.769 when only malignant cell clusters were annotated. These findings suggest that including benign clusters in the annotation process enhances malignant cell detection accuracy. While these results are promising preliminary evidence, further validation with larger, multicenter datasets is necessary to confirm whether this insight can serve as general guidance for annotation strategies in AI model development across various pathological domains, particularly in contexts where both benign and malignant cell clusters coexist or where there may be ambiguity in subtype labeling.

Comparison of YOLOv5x and YOLOv7 Accuracy

To compare the diagnostic performance of YOLOv5x and YOLOv7, hyperparameter tuning was performed under identical conditions in our study, with 100 training epochs and a batch size of four. YOLOv5x achieved the highest mAP for both benign and malignant cell clusters, outperforming YOLOv7. No signs of overfitting were observed on the loss curves, suggesting that the inclusion of both benign and malignant annotations improved generalizability. Interestingly, similar findings have been reported in recent studies involving medical image analysis. Prisilla et al.[21] demonstrated that YOLOv5x outperformed YOLOv7 in detecting lumbar disc herniation on MRI. While Sadhin et al.[22] reported that YOLOv7 had a higher mAP in road damage detection, the YOLOv5 series, including YOLOv5x, still exhibited strong detection precision and utility.

These results suggest that the optimal model architecture may depend heavily on task-specific image characteristics. In cytological images, accurate detection often requires interpreting not only localized features but also the broader context of cell clusters, such as irregular margins, structural overlap, and peripheral cytoplasmic morphology. YOLOv5x, with its larger parameter capacity and expressive power, may be more capable of capturing these global morphological features than more compact architectures like YOLOv7. In contrast, YOLOv7's architectural efficiencies may be better suited to large-scale natural image datasets. These differences highlight the importance of evaluating model suitability within the specific context of clinical applications.

Correlation between Model and Diagnostician Diagnoses

Several previous studies examined how AI affects human diagnostic performance, particularly in cases perceived as difficult. For example, in lung cancer detection, high-accuracy AI improved radiologist performance in difficult cases, while incorrect AI predictions had adverse effects[23]. Similarly, Kiani et al.[24] demonstrated that AI assistance significantly influenced pathologists' decisions in the histopathological classification of liver cancer: when the AI model's prediction was correct, diagnostic accuracy improved, but when incorrect, accuracy decreased—even across different levels of expertise and case difficulty. These studies suggest that the impact of AI in clinical settings is not solely dependent on its accuracy. How its outputs interact with human perception and decision-making is also important.

However, rather than focusing on how AI influences human performance, our study investigated how human perception of case difficulty relates to the diagnostic accuracy of the AI model itself. Interestingly, we found no significant difference in perceived difficulty between cases the AI correctly and incorrectly diagnosed. This suggests that the AI model's performance was robust regardless of diagnosticians' perceived difficulty and that AI and human experts may be evaluating complexity through fundamentally different frameworks.

Furthermore, recent studies in computational pathology emphasize a complementary relationship between AI and human diagnosticians. Rather than replacing pathologists, AI is increasingly seen as an assistive tool capable of performing standardized and repetitive tasks with high precision, thereby freeing pathologists to focus on complex cases and integrative decision-making[25,26]. In particular, AI excels at tasks requiring quantitative consistency, such as biomarker assessment or routine classification, while pathologists provide contextual interpretation based on rare patterns or clinical complexity. This evolving role of AI as a diagnostic assistant aligns with our findings, which suggest that the AI model maintained stable diagnostic accuracy regardless of human-perceived difficulty. Understanding how AI and humans perceive diagnostic complexity differently—and where their assessments diverge—is critical for designing systems that are truly integrated with the pathology workflow.

To deepen this understanding, it is essential to visualize how the AI model makes its diagnostic decisions and whether those decisions rely on features similar to those used by human diagnosticians. In our study, we used Grad-CAM to visualize the regions the AI model focused on during diagnosis, thereby offering interpretable in-

sights into the model's reasoning process.

Visualization of AI Diagnostic Basis

To enhance the interpretability of the AI model's diagnostic decisions, we used Grad-CAM to visualize the specific regions within cell clusters that influenced the model's predictions. Grad-CAM calculates the gradient of the output with respect to the final convolutional layer and combines it with the associated feature maps, producing heatmaps that highlight areas most relevant to the AI's decision-making process[15]. This architecture-agnostic technique can be readily applied to convolutional neural networks without requiring structural modifications, making it particularly well-suited for real-time cytological applications.

Grad-CAM has shown promise in various medical imaging domains. Musthafa et al.[27] demonstrated its utility in brain tumor detection using MRI, while Wei et al.[28] employed Grad-CAM to visualize prognostic features in colorectal cancer pathology. In cytological diagnosis, where real-time feedback and intuitive interpretation are crucial, the ability of Grad-CAM to directly map influential regions onto diagnostic images offers practical advantages for clinical integration.

However, Grad-CAM does not fully resolve the black-box issue inherent in deep learning. Grad-CAM provides a coarse, architecture-dependent localization of discriminatory regions but does not necessarily reflect a single, interpretable decision rule. Therefore, these visualizations should be interpreted as highlighting statistical saliency rather than revealing a definitive causal reasoning process. In our study, there were instances in which the model's highlighted regions diverged from areas typically emphasized by pathologists—such as nuclear atypia or irregular cluster margins—and instead focused on peripheral cytoplasmic zones or loosely aggregated cell areas. This discrepancy reinforces the understanding that Grad-CAM visualizations may not always faithfully represent human-like diagnostic reasoning but rather reflect statistical saliency at the final convolutional layer.

This concern aligns with findings by Saporta et al.[29], who reported that Grad-CAM failed to reliably localize diagnostically critical regions in chest X-ray analysis, particularly for pathologies that were small, multifocal, or morphologically complex. They also reported a positive correlation between model confidence and Grad-CAM reliability, suggesting that saliency maps should be interpreted cautiously in cases where model certainty is low.

In summary, Grad-CAM is a valuable tool for enhancing transparency in AI-assisted cytological diagnosis, particularly in settings requiring real-time visual feedback. Nevertheless, its limitations, especially the potential misalignment with expert-derived diagnostic features, underscore the need for continued investigation of the fidelity of interpretability tools and their integration into clinical decision-making frameworks.

Limitations

This study has several limitations. First, the dataset size was relatively small, particularly for rare malignant cases and high-grade tumors, which limits the generalizability of the model. Specifically, because of the limited number of cases in the test set, a subgroup analysis comparing performance across different histological grades (e.g., Grade 1 vs. Grade 3) could not be performed. Verifying the model's accuracy on high-grade carcinomas remains an important task for future research using larger datasets. This constraint reflects a broader challenge in AI cytopathology development, where the acquisition of rare cases and accurate annotations is inherently resource-intensive and time-consuming. In addition, the limited availability of publicly accessible annotated medical image datasets is a major barrier to innovation in computational research and education. Second, all data were collected from a single facility, requiring external validation across multiple institutions and using different imaging equipment. This introduces the risk of domain shift, where a model trained on data from one facility or device may perform poorly when applied to data from different settings. Variations in imaging conditions, such as magnification, lighting, and even patient demographics, can significantly impact the model's ability to generalize. Therefore, it is crucial to test the model across diverse institutions and imaging modalities to ensure broader applicability.

Additionally, as discussed with Grad-CAM, the visualized diagnostic basis was sometimes unclear, further highlighting unresolved aspects of the AI model's black-box nature. Lastly, the model's performance is dependent on hardware constraints and computational resources, which may limit its scalability in clinical practice. Addressing these limitations will require larger datasets, collaboration with multiple institutions, and continued development of methods to improve explainability and robustness. In particular, non-visual XAI methods—such as auxiliary explanations, case-based explanations, and textual explanations—may serve as promising alternatives or complements to visual techniques like Grad-CAM in

resolving the black-box challenges of deep learning models[30].

Significance of the Developed Model

Although this study focuses on cell clusters, slide-level performance is crucial for clinical use. Our AI model integrates with a microscope and a CCD camera, allowing diagnosticians to observe AI-generated judgments and explanations in real time on a nearby screen. While 30 frames per second (FPS) is generally considered sufficient for real-time detection, our previous research showed that our model can achieve 60 FPS. This higher frame rate ensures smooth operation, minimizing any potential delays or stress for diagnosticians as they move the slide. The ability to provide real-time feedback at this speed enhances the clinical workflow by reducing diagnostic time and promoting workstyle reform in medical settings, such as decreasing work hours and improving overall efficiency. Furthermore, many studies of medical AI focus solely on accuracy metrics, and few have investigated the correlation between AI model performance and the assessments made by the diagnosticians who intend to use these models, as the present study does. By incorporating this perspective, our work not only contributes to technical advancement but also aligns with the practical realities of clinical deployment. Moreover, in the development of medical AI models, many healthcare professionals feel that AI might threaten their job security[31], necessitating development approaches that are more aligned with the needs of healthcare providers. This study proposes a new direction for the practical implementation of medical AI models by evaluating the correlation between perceived diagnostic difficulty and the AI model's accuracy, ensuring that the model is developed not only with consideration for the diagnosticians' role and perspective but also as a tool to support and enhance their diagnostic capabilities.

Furthermore, the challenges encountered in this study, such as acquiring rare cases, annotating cytological images, and ensuring real-time compatibility, are not unique to endometrial cytology. They reflect broader issues in implementing AI in pathology. Therefore, our strategy, including microscope-integrated inference and explainability visualization, may be extended to other cytological and pathological domains, such as the analysis of bone marrow aspirates[32], where real-time decision support and interpretable results are equally critical. Ultimately, we hope that this model will serve as a foundation for broader applications, such as automated screening support in resource-limited settings and educational tools for training cytotechnologists and pathologists, thereby further expanding the utility of AI in medicine.

Conclusion

In this study, we developed and evaluated AI models based on YOLOv5x and YOLOv7. YOLOv5x had the highest diagnostic accuracy (mAP 0.798) for both benign and malignant cell cluster annotations. The annotation strategy that targeted both benign and malignant cell clusters was a novel aspect in the field of pathology AI, offering preliminary evidence that this approach may improve detection performance. Analysis of the correlation between AI model performance and perceived diagnostic difficulty by cytology professionals revealed that the AI maintained consistent accuracy, regardless of case complexity. Additionally, the use of Grad-CAM for visualizing diagnostic reasoning significantly enhanced the transparency and interpretability of the model, making it more reliable for clinical use. These advances highlight the potential of integrating AI-assisted tools into pathological workflows to improve diagnostic precision and efficiency.

The high-performing YOLOv5x model, its robust performance across varying levels of case difficulty, and the enhanced model interpretability through Grad-CAM visualization emphasize the importance of AI–human collaboration in medical diagnosis. This study is a significant milestone in the development of AI-assisted pathology tools, offering valuable insights for future research and clinical use. Furthermore, the focus on model transparency and consistent performance strengthens the role of AI in healthcare, ultimately contributing to more accurate and more reliable diagnostic outcomes. Importantly, because this system operates without the need for whole-slide scanners and integrates directly with standard microscopes, it holds potential for broader applications beyond endometrial cytology, including other cytological and pathological workflows.

Author Contributions: I.S. and M.T. are responsible for the study concept and design, sample collection, development of methodology, and writing, review, and revision of the paper; S. Tanaka developed the methodology, and writing, review, and revision of the paper; E.T. and Y.T. acquired and analyzed the data and reviewed and revised the paper. S. Takakuma, Y. K., and S.K. provided technical and material support; Y.T. and A.S. reviewed the study concept and design, as well as the paper. All the authors read and approved the final manu-

script.

Acknowledgments: We extend our gratitude to Arimi Ishikawa, Naomi Kuwahara, and Masaya Okaizumi for their technical support throughout this study. We would also like to express our sincere gratitude to the cytotechnologists and the medical technologist, Yukihiro Murase, Harumi Kamaguchi, Sachiko Nagai, Ayako Hayama, Mika Sakata, Yuri Kimura, Marina Aida, Satoshi Suzuki, Nana Inoue, Rina Kodama, Satono Shima, Yasuhiko Watarai, and Marina Aoki, for their invaluable cooperation in this study. We also thank Kazuhiro Takeuchi, Hideaki Kuno, and Emi Sakamoto for their contributions to the education of medical students, and Kenta Tominaga for his unwavering support. Furthermore, we appreciate the development of YOLOv5x by Ultralytics and Mohammadi Kazaj, Pooya, who connects YOLOv5x and Grad-CAM.

Funding: This study was supported by a Grant-in-Aid (No. 23K08900) awarded to Mika Terasaki by the Japan Society for the Promotion of Science through the Diversity Women Leader Development Grant and Scientific Research (C). The funder had no role in the committee work, discussions, literature research, decision to publish, or manuscript preparation.

Conflict of Interest: The authors declare no competing interests.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this work, the authors used AI-assisted technologies (Chat GPT and Gemini) for tasks such as English academic writing assistance. After using these tools, the authors reviewed and edited all content and take full responsibility for the content of the published article.

Supplementary Material: Supplementary material associated with this article is available at https://doi.org/10.1272/jnms.JNMS.2026_93-115.

References

1. Crosbie EJ, Kitson SJ, McAlpine JN, Mukhopadhyay A, Powell ME, Singh N. Endometrial cancer. Lancet. 2022;399(10333):1412−28. doi: 10.1016/S0140-6736(22)00323-3
2. O'Flynn H, Ryan NAJ, Narine N, Shelton D, Rana D, Crosbie EJ. Diagnostic accuracy of cytology for the detection of endometrial cancer in urine and vaginal samples. Nat Commun. 2021;12(1):952. doi: 10.1038/s41467-021-21257-6
3. Wang Q, Wang Q, Zhao L, et al. Endometrial cytology as a method to improve the accuracy of diagnosis of endometrial cancer: case report and meta-analysis. Front Oncol. 2019;9:256. doi: 10.3389/fonc.2019.00256
4. Clarke MA, Long BJ, Del Mar Morillo A, Arbyn M, Bakkum-Gamez JN, Wentzensen N. Association of endometrial cancer risk with postmenopausal bleeding in women: a systematic review and meta-analysis. JAMA Intern Med. 2018;178(9):1210−22. doi: 10.1001/jamainternmed.2018.2820
5. Yanoh K, Hirai Y, Sakamoto A, et al. New terminology for intrauterine endometrial samples: a group study by the Japanese Society of Clinical Cytology. Acta Cytol. 2012;56(3):233−41. doi: 10.1159/000336258
6. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol. 2020;1213:3−21. doi: 10.1007/978-3-030-33128-3_1
7. Ben Hamida A, Devanne M, Weber J, et al. Deep learning for colon cancer histopathological images analysis. Comput Biol Med. 2021;136:104730. doi: 10.1016/j.compbiomed.2021.104730
8. Tsiknakis N, Theodoropoulos D, Manikis G, et al. Deep learning for diabetic retinopathy detection and classification based on fundus images: a review. Comput Biol Med. 2021;135:104599. doi: 10.1016/j.compbiomed.2021.104599
9. Serag A, Ion-Margineanu A, Qureshi H, et al. Translational AI and deep learning in diagnostic pathology. Front Med (Lausanne). 2019;6:185. doi: 10.3389/fmed.2019.00185
10. Zuraw A, Aeffner F. Whole-slide imaging, tissue image analysis, and artificial intelligence in veterinary pathology: an updated introduction and review. Vet Pathol. 2022;59(1):6−25. doi: 10.1177/03009858211040484
11. Zhang X, Ba W, Zhao X, et al. Clinical-grade endometrial cancer detection system via whole-slide images using deep learning. Front Oncol. 2022;12:1040238. doi: 10.3389/fonc.2022.1040238
12. Terasaki M, Tanaka S, Shimokawa I, et al. An integrated system from microscopy to AI for real-time object detection in endometrial cytology. J Pathol Inform. Forthcoming. doi: 10.1016/j.jpi.2025.100541
13. Park SH, Han K, Jang HY, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. Radiology. 2023;306(1):20−31. doi: 10.1148/radiol.220182
14. Gallee L, Kniesel H, Ropinski T, Gotz M. Artificial intelligence in radiology - beyond the black box. Rofo. 2023;195(9):797−803. doi: 10.1055/a-2076-6736
15. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128(2):336−59. doi: 10.1007/s11263-019-01228-7
16. Ragab MG, Abdulkader SJ, Muneer A, et al. A comprehensive systematic review of YOLO for medical object detection (2018-2023). IEEE Access. 2024;12:57815−36. doi: 10.1109/ACCESS.2024.3386826
17. Tulbure AA, Tulbure AA, Dulf EH. A review on modern defect detection models using DCNNs - deep convolutional neural networks. J Adv Res. 2021;35:33−48. doi: 10.1016/j.jare.2021.03.015
18. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detector. arXiv [preprint]. 2022 Jul 6 [Internet]. Available from: https://doi.org/10.48550/arXiv:2207.02696
19. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc Psychol Pers Sci. 2017;8(4):355−62. doi: 10.1177/1948550617697177
20. Fell C, Mohammadi M, Morrison D, et al. Detection of malignancy in whole slide images of endometrial cancer

biopsies using artificial intelligence. PLoS One. 2023;18(3): e0282577. doi: 10.1371/journal.pone.0282577
21. Prisilla AA, Guo YL, Jan YK, et al. An approach to the diagnosis of lumbar disc herniation using deep learning models. Front Bioeng Biotechnol. 2023;11:1247112. doi: 10.3389/fbioe.2023.1247112
22. Sadhin AH, Hashim SZM, Samma H, Khamis N. YOLO: a competitive analysis of modern object detection algorithms for road defects detection using drone images. Baghdad Sci J. 2024;21(6):Article 24. doi: 10.21123/bsj.2023.9027
23. Lee JH, Hong H, Nam G, Hwang EJ, Park CM. Effect of human-AI interaction on detection of malignant lung nodules on chest radiographs. Radiology. 2023;307(5):e222976. doi: 10.1148/radiol.222976
24. Kiani A, Uyumazturk B, Rajpurkar P, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPI Digit Med. 2020;3:23. doi: 10.1038/s41746-020-0232-8
25. Cifci D, Veldhuizen GP, Foersch S, Kather JN. AI in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? Annu Rev Cancer Biol. 2023;7:57–71. doi: 10.1146/annurev-cancerbio-061521-092038
26. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol. 2022;35(1):23–32. doi: 10.1038/s41379-021-00919-2
27. Mohamed Musthafa M, Mahesh TR, Vinoth Kumar V, Guluwadi S. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Med Imaging. 2024;24(1):107. doi: 10.1186/s12880-024-01292-7
28. Wei B, Li L, Feng Y, Liu S, Fu P, Tian L. Exploring prognostic biomarkers in pathological images of colorectal cancer patients via deep learning. J Pathol Clin Res. 2024;10(6):e70003. doi: 10.1002/2056-4538.70003
29. Saporta A, Gui X, Agrawal A, et al. Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell. 2022;4(10):867–78. doi: 10.1038/s42256-022-00536-x
30. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med. 2023;29(9):2307–16. doi: 10.1038/s41591-023-02504-3
31. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - beyond saliency-based XAI approaches. Eur J Radiol. 2023;162:110786. doi: 10.1016/j.ejrad.2023.110786
32. Bermejo-Pelaez D, Rueda Charro S, Garcia Roa M, et al. Digital microscopy augmented by artificial intelligence to interpret bone marrow samples for hematological diseases. Microsc Microanal. 2024;30(1):151–9. doi: 10.1093/micmic/ozad143