Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative performance of four deep learning object detection models in static images and microscopic videos of effusion cytology

Katsuhide Ikeda

ikeda.katsuhide.y1@f.mail.nagoya-u.ac.jp

Nagoya University Graduate Scholl of Medicine

Nanako Sakabe

Nagoya University Graduate Scholl of Medicine

Kenta Fukuda

Nagoya University Graduate Scholl of Medicine

Kota Maruta

Nagoya University Graduate Scholl of Medicine

Mio Masuda

Nagoya University Graduate Scholl of Medicine

Shouichi Sato

Juntendo University

Toshiaki Hara

Nagoya University Hospital

Harumi Kobayashi

Nagoya University Hospital

Masato Nakaguro

Nagoya University Graduate Scholl of Medicine

Kennosuke Karube

Nagoya University Graduate Scholl of Medicine

Article

Keywords: Effusion cytology, Deep learning, Object detection, Malignant cell detection, Microscopic video analysis, Artificial intelligence

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11670228/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title

Comparative performance of four deep learning object detection models in static images and microscopic videos of effusion cytology

Author names

Katsuhide Ikeda, Ph.D.,[a*] Nanako Sakabe, Ph.D.,[a] Kenta Fukuda,[a] Kota Maruta,[a] Mio Masuda,[a] Shouichi Sato, Ph.D.,[b] Toshiaki Hara,[c] Harumi Kobayashi,[c] Masato Nakaguro, Ph.D.,[c,d] Kennosuke Karube, Ph.D.,[c,d]

Institutional affiliation

[a] Pathophysiology Sciences, Department of Integrated Health Sciences, Nagoya University Graduate School of Medicine, Nagoya, Aichi, Japan

[b] Clinical Engineering, Faculty of medical sciences, Juntendo University, Urayasu, Chiba, Japan

[c] Department of Medical Technique, Division of Pathology, Nagoya University Hospital, Nagoya, Aichi, Japan

[d] Department of Pathology and Laboratory Medicine, Nagoya University Graduate School of Medicine, Nagoya, Aichi, Japan

***Corresponding author**

Katsuhide Ikeda, Ph.D.

Pathophysiology Sciences, Department of Integrated Health Sciences

Nagoya University Graduate School of Medicine

1-1-20 Daiko-Minami, Higashi-ku, Nagoya 461-8673, JAPAN

Tel: +81-52-719-3152

Fax no.: +81-52-719-3152

E-mail: ikeda.katsuhide.y1@f.mail.nagoya-u.ac.jp

Abstract

Selecting deep learning models for effusion cytology requires evaluating their performance beyond static images. This study compared four object detection models—YOLO26x, RF-DETR-L, RT-DETRv2 (R50vd), and Faster R-CNN V2—using patient-independent image and microscopic video datasets. Bounding box-level detection was evaluated on 5136 images from 168 cases containing 7645 ground-truth annotations of malignant cells and cell clusters. Case-level classification was assessed using 80 videos (40 malignant and 40 negative cases), defining an AI-positive case by persistent detection across at least three consecutive frames. At validation-derived F1-optimized thresholds, image-based F1-scores were 0.795, 0.755, 0.808, and 0.767, respectively, with no pairwise F1-score differences clearly established by bootstrap confidence intervals. Video sensitivities were 100.0%, 100.0%, 100.0%, and 82.5%, whereas specificities were 32.5%, 22.5%, 0.0%, and 90.0%, respectively. Median false-positive alert frequencies in negative videos were 3.0, 3.0, 12.5, and 0.0 events/min, respectively. YOLO26x achieved the highest effective processing speed (35.57 frames/s). These findings demonstrate that static-image detection performance does not adequately characterize video-based classification and alert behavior, highlighting the need for application-specific, multidimensional evaluation before clinical implementation.

Keywords

Effusion cytology; Deep learning; Object detection; Malignant cell detection; Microscopic video analysis; Artificial intelligence

INTRODUCTION

Effusion cytology, including the examination of pleural, peritoneal, and pericardial fluids, plays an important role in the diagnosis, staging, and treatment planning of malignant neoplasms. In cytological examination, malignant cells must be identified within the specimen and their morphological features evaluated. However, detecting malignant cells can be challenging when they are present in small numbers or when numerous non-malignant cells, such as reactive mesothelial cells and histiocytes, are present. Microscopic evaluation is further complicated by cellular overlap, degenerative changes, and variations in focal planes. Detecting malignant cells and determining their histological type or primary site of origin are distinct diagnostic tasks. Reducing the likelihood of overlooking malignant cells is a prerequisite for subsequent diagnostic characterization. In this context, artificial intelligence (AI)–assisted detection systems that support microscopic examination by cytotechnologists and cytopathologists and highlight the locations of suspected malignant cells are expected to facilitate the identification of malignant cells.

In the field of cytology, deep learning research has evolved from image-level classification of benign and malignant specimens to object detection for localizing individual cells and cell clusters, as well as whole-slide image (WSI) analysis. In effusion cytology, deep learning–based approaches have also been investigated for distinguishing malignant from benign cases [1-3]. Studies using WSIs of pleural effusion cytology specimens [1,2] have demonstrated the potential to identify malignant features in previously acquired digital images. Furthermore, Ikeda et al. reported a precision of 0.909, recall of 0.911, and F1-score of 0.910 on an independent test image dataset using You Only Look Once v8x (YOLOv8x) for object detection of adenocarcinoma cells in effusion cytology, demonstrating the potential to localize suspected malignant cells [4]. Tian et al. developed a deep learning system called Tumor Origin Differentiation using Cytological Histology (TORCH), demonstrating that the integration of effusion cytology images and clinical information enables benign–malignant classification and prediction of the primary organ of origin [5]. Furthermore, comparative studies of

multiple object detection architectures, including Faster R-CNN, YOLO, and DETR-based models, have been reported in cervical cytology [6,7] and thyroid cytology [8]. However, image-level classification performance, cell-level localization performance, and case-level classification performance represent distinct evaluation metrics. Therefore, performance obtained from previously acquired static images alone may not be sufficient to assess the utility of these models during continuous microscopic examination.

In recent years, AI-assisted diagnostic systems that overlay AI-detected candidate regions onto microscopic images have been investigated, extending beyond the analysis of previously acquired images. Terasaki et al. reported a YOLO-based real-time detection system for endometrial cytology [9], while Tang et al. demonstrated the potential of AI-assisted microscopic examination in cervical cytology [10]. However, several challenges in the implementation of digital cytology have been identified, including three-dimensional cellular overlap, variations in focal planes, and image acquisition conditions [11,12]. In particular, microscopic videos introduce additional temporal factors, such as microscope stage movement and stationary periods, changes in focus, motion blur, and the repeated appearance of the same cells across consecutive frames. Therefore, simply counting individual bounding boxes (BBoxes) generated across video frames is insufficient to adequately characterize the persistence of detected candidates, the frequency and temporal characteristics of alerts, or case-level diagnostic performance. Accordingly, previous studies have highlighted the importance of separately evaluating object-level detection and case-level diagnostic performance in microscopic images [13], as well as incorporating the temporal distribution of detections into performance evaluation [14]. When evaluating AI performance in microscopic videos, it is necessary to assess not only precision and recall in static images but also case-level sensitivity and specificity, temporal persistence of detections, false-positive (FP) alert frequency in negative cases, and processing speed.

Although previous studies have investigated AI-based analysis of static images and WSIs in

effusion cytology, as well as AI-assisted microscopic examination using specific algorithms, few studies have systematically compared different object detection architectures using the same independent image and video datasets without patient overlap. In particular, evidence remains limited regarding comparisons of cell-level localization performance, case-level classification performance in videos, FP alert frequency, and processing speed under standardized evaluation criteria. In this study, four deep learning–based object detection models were developed using YOLO26x, Roboflow Detection Transformer Large (RF-DETR-L), Real-Time Detection Transformer version 2 (RT-DETRv2; R50vd backbone), and Faster Region-based Convolutional Neural Network (Faster R-CNN V2), and their performance in detecting malignant cells was compared using independent effusion cytology image and microscopic video datasets without patient overlap. Confidence thresholds for each model were determined in advance using the validation dataset. BBox-level detection performance was evaluated in static images, and case-level classification performance, FP alert frequency, and processing speed were evaluated in videos. This study aimed to provide insights into selecting appropriate models for different applications, with a view toward their potential use in AI-assisted microscopic examination. An overview of the study design and dataset allocation is shown in **Fig. 1**.

RESULTS

Model training results and selection of confidence thresholds

Changes in detection performance on the validation dataset during training for the four object detection algorithms are shown in **Supplementary Fig. S1**, and the training loss curves for each model are shown in **Supplementary Fig. S2**. The final training epoch, epoch of the best checkpoint, and validation performance for each model are summarized in **Supplementary Table S1**. The mean average precision (AP) across intersection over union (IoU) thresholds ranging from 0.50 to 0.95 in increments of 0.05 (AP50:95) on the validation dataset was used as the primary metric for model

selection. For each algorithm, the checkpoint achieving the highest validation AP50:95 was selected and fixed for subsequent evaluation on the independent test image and video datasets. YOLO26x training was completed after 69 epochs, with the highest validation AP50:95 of 0.8154 achieved at epoch 39. RF-DETR-L training was completed after 22 epochs, with the highest validation AP50:95 of 0.8207 achieved at epoch 10. RT-DETRv2 (R50vd) completed 72 training epochs, achieving its highest validation AP50:95 of 0.8222 at epoch 34. In contrast, Faster R-CNN V2 training was completed after 18 epochs, with the highest validation AP50:95 of 0.7030 achieved at epoch 3.

Changes in precision, recall, F1-score, and FP per image (FP/image) across confidence thresholds for each algorithm are shown in **Supplementary Fig. S3**. Confidence thresholds were evaluated from 0.05 to 0.95 in increments of 0.05 using the validation dataset. The confidence thresholds determined for each model before evaluation on the independent test datasets are summarized in **Supplementary Table S2**. The F1-optimized thresholds, which maximized the F1-score, were 0.45 for YOLO26x and RF-DETR-L, 0.65 for RT-DETRv2 (R50vd), and 0.90 for Faster R-CNN V2. The corresponding maximum F1-scores were 0.8335, 0.8405, 0.8415, and 0.7634, respectively, with RT-DETRv2 (R50vd) achieving the numerically highest value. Furthermore, the high-recall thresholds, defined as the thresholds yielding the lowest FP/image among those achieving a BBox-level recall of at least 0.95 for malignant cells, were 0.05 for YOLO26x, 0.20 for RF-DETR-L, 0.35 for RT-DETRv2 (R50vd), and 0.15 for Faster R-CNN V2.

Object detection performance on the independent test image dataset

The detection performance of four object detection algorithms on 5136 images containing 7645 ground truth BBoxes (GT BBoxes), obtained from 168 patients (149 malignant and 19 negative cases) in the independent test image dataset, is summarized in **Table 1**, and the precision–recall curves are shown in **Fig. 2**. When a fixed confidence threshold of 0.50 was applied to all models, RT-DETRv2 (R50vd) achieved the numerically highest F1-score of 0.809 (95% confidence interval [CI]: 0.762–0.851) and the highest recall of 0.851. The F1-score of YOLO26x was 0.796, exceeding those of RF-

DETR-L (0.755) and Faster R-CNN V2 (0.750). In contrast, YOLO26x achieved the highest precision of 0.850 and the lowest FP/image of 0.196. When the F1-optimized confidence thresholds selected using the validation dataset were applied, RT-DETRv2 (R50vd) achieved the numerically highest F1-score of 0.808 (95% CI: 0.758–0.854). The F1-scores of YOLO26x, RF-DETR-L, and Faster R-CNN V2 were 0.795, 0.755, and 0.767, respectively. Faster R-CNN V2 achieved the highest precision of 0.905 (95% CI: 0.885–0.922) and the lowest FP/image of 0.104 at a confidence threshold of 0.90. However, its recall was 0.665, the lowest among the four models. When the high-recall thresholds selected to achieve a malignant cell BBox-level recall of at least 0.95 on the validation dataset were applied, recall on the independent test dataset was 0.930 for YOLO26x, 0.956 for RF-DETR-L, 0.911 for RT-DETRv2 (R50vd), and 0.922 for Faster R-CNN V2. RF-DETR-L was the only model that maintained a recall of at least 0.95 on the independent test dataset. In contrast, RT-DETRv2 (R50vd) achieved the lowest FP/image of 0.746 and the highest F1-score of 0.755, whereas Faster R-CNN V2 showed the highest FP/image of 1.860.

Furthermore, to assess object detection performance across confidence thresholds, a Common Objects in Context (COCO)-based evaluation was performed on the independent test dataset (**Supplementary Table S3**). The evaluation metrics included AP at an IoU threshold of 0.50 (AP50), AP at an IoU threshold of 0.75 (AP75), and average recall across the same IoU thresholds with a maximum of 100 detections per image (AR100). RT-DETRv2 (R50vd) achieved the numerically highest values among the four models for AP50:95 (0.8410), AP50 (0.8947), AP75 (0.8757), and AR100 (0.9544). YOLO26x and RF-DETR-L showed similar AP50:95 values of 0.8243 and 0.8207, respectively, whereas Faster R-CNN V2 achieved an AP50:95 of 0.7621. The AR100 of RF-DETR-L was 0.9498, close to that of RT-DETRv2 (R50vd) (0.9544).

Pairwise comparison of image-based detection performance

Differences in detection performance among the four object detection algorithms on the independent test dataset were evaluated using patient-level cluster bootstrapping, preserving the paired structure of

the data across models (**Supplementary Table S4**). When the F1-optimized thresholds determined using the validation dataset were applied, RT-DETRv2 (R50vd) achieved the numerically highest F1-score among the four models. However, the 95% CIs for the differences in F1-score included zero in all six pairwise comparisons. In contrast, recall and FP/image showed different patterns across models. Compared with Faster R-CNN V2, RF-DETR-L had a higher recall by 0.1721 (95% CI: 0.1161–0.2250) and a higher FP/image by 0.4622 (95% CI: 0.0639–1.0701). Similarly, RT-DETRv2 (R50vd) had a higher recall than Faster R-CNN V2 by 0.1058 (95% CI: 0.0644–0.1486) and a higher FP/image by 0.0993 (95% CI: 0.0033–0.2362). These findings indicate that the balance between malignant cell detection and FP detections differed among the algorithms. Representative detection results from the four models applied to the same image in the independent test dataset are shown in **Fig. 3**. This figure visually illustrates differences in detection locations and patterns among the models but does not represent their overall performance across the dataset.

Case-level detection performance on the independent test video dataset

The case-level detection performance of the four deep learning models was evaluated using the F1-optimized confidence thresholds determined from the validation dataset on 80 independent test videos (40 malignant and 40 negative cases), and the results are presented in **Table 2**. YOLO26x, RF-DETR-L, and RT-DETRv2 (R50vd) each classified all 40 malignant cases as AI-positive, yielding a sensitivity of 100.0% (95% CI: 91.2–100.0%). In contrast, Faster R-CNN V2 classified 33 of the 40 malignant cases as AI-positive, with a sensitivity of 82.5% (95% CI: 68.1–91.3%). Among the 40 negative cases, specificity was 32.5% (95% CI: 20.1–48.0%) for YOLO26x, 22.5% (95% CI: 12.3–37.5%) for RF-DETR-L, and 0.0% (95% CI: 0.0–8.8%) for RT-DETRv2 (R50vd). In contrast, Faster R-CNN V2 classified 36 of the 40 negative cases as AI-negative, achieving the highest specificity of 90.0% (95% CI: 76.9–96.0%) among the four models. Overall accuracy across all 80 cases was 66.25% (95% CI: 55.4–75.7%) for YOLO26x, 61.25% (95% CI: 50.3–71.2%) for RF-DETR-L, 50.0% (95% CI: 39.3–60.7%) for RT-DETRv2 (R50vd), and 86.25% (95% CI: 77.0–92.1%) for Faster R-

CNN V2, with Faster R-CNN V2 achieving the highest accuracy. Under the F1-optimized threshold condition, at least one predefined qualified AI alert event occurred in all 40 malignant cases evaluated using YOLO26x, RF-DETR-L, and RT-DETRv2 (R50vd), resulting in AI-positive classifications. However, these models also produced frequent FP classifications in negative cases. Although Faster R-CNN V2 failed to classify some malignant cases as AI-positive, it achieved the highest proportion of correctly classified AI-negative cases among the four models.

Comparison of case-level detection performance among four models in videos

The case-level detection performance of the four object detection models was compared using 80 independent test videos, accounting for the paired structure of the data across models for the same cases (**Supplementary Table S5**). Cochran's Q tests, which accounted for the paired observations across the four models, revealed statistically significant differences among the models in both sensitivity for the 40 malignant cases ($Q = 21.00$, degrees of freedom = 3, $P < 0.001$) and specificity for the 40 negative cases ($Q = 74.21$, degrees of freedom = 3, $P < 0.001$). For sensitivity in the 40 malignant cases, no differences were observed among YOLO26x, RF-DETR-L, and RT-DETRv2 (R50vd). Each of these three models showed a sensitivity 17.5 percentage points (pp) higher than that of Faster R-CNN V2 (95% CI: 7.5–30.0 pp). However, the Holm-adjusted P values from the two-sided exact McNemar tests were 0.094 for all three comparisons, indicating that these differences were not statistically significant. For specificity in the 40 negative cases, no statistically significant difference was observed between YOLO26x and RF-DETR-L (difference: 10.0 pp, 95% CI: 0.0–22.5 pp, adjusted $P = 0.219$). In contrast, Faster R-CNN V2 showed significantly higher specificity than YOLO26x, RF-DETR-L, and RT-DETRv2 (R50vd), with differences of 57.5 pp, 67.5 pp, and 90.0 pp, respectively (all adjusted $P < 0.001$). Furthermore, YOLO26x and RF-DETR-L showed significantly higher specificity than RT-DETRv2 (R50vd), with differences of 32.5 pp (95% CI: 17.5–47.5 pp, adjusted $P < 0.001$) and 22.5 pp (95% CI: 10.0–35.0 pp, adjusted $P = 0.008$), respectively. In the case-level evaluation using F1-optimized confidence thresholds, no statistically significant pairwise

differences in sensitivity were identified among the four models after adjustment for multiple comparisons. In contrast, significant differences in specificity were observed, with Faster R-CNN V2 achieving the highest specificity.

False-positive alert frequency and video processing speed

FP alert event frequencies were evaluated for the four models using the F1-optimized confidence thresholds determined from the validation dataset in 40 negative cases from the independent test video dataset (**Table 3**, **Fig. 4**). The median (IQR) number of FP alert events per minute (FP/min) was 3.0 (0.0–8.0) for YOLO26x, 3.0 (1.0–11.3) for RF-DETR-L, 12.5 (6.0–24.0) for RT-DETRv2 (R50vd), and 0.0 (0.0–0.0) for Faster R-CNN V2. The Friedman test revealed a statistically significant difference among the four models ($\chi^2(3) = 84.34$, $P < 0.001$). Post hoc pairwise comparisons using the Wilcoxon signed-rank test showed no statistically significant difference between YOLO26x and RF-DETR-L (Holm-adjusted $P = 0.213$). In contrast, all five remaining pairwise comparisons yielded Holm-adjusted P values < 0.001. As secondary analyses, FP alert event frequencies were evaluated using a fixed confidence threshold of 0.50 for all models and the high-recall thresholds determined from the validation dataset (**Supplementary Table S6**). Under the fixed-threshold condition, the median (IQR) frequencies were 2.5 (0.0–6.0), 2.5 (0.8–9.3), 24.0 (10.8–41.5), and 2.0 (0.0–3.3) FP/min for YOLO26x, RF-DETR-L, RT-DETRv2 (R50vd), and Faster R-CNN V2, respectively. Under the high-recall condition, the corresponding frequencies were 16.5 (6.0–38.0), 15.5 (5.8–38.5), 47.0 (26.3–73.8), and 8.0 (1.8–16.0) FP/min, respectively. FP alert event frequencies differed significantly among the four models under both the fixed-threshold condition ($\chi^2(3) = 78.03$, $P < 0.001$) and the high-recall condition ($\chi^2(3) = 63.58$, $P < 0.001$). The median (IQR) effective processing speeds across the 80 independent test videos were 35.57 (34.03–36.20) frames per second (FPS) for YOLO26x, 22.82 (22.26–23.03) FPS for RF-DETR-L, 19.55 (19.35–19.84) FPS for RT-DETRv2 (R50vd), and 19.85 (19.69–20.10) FPS for Faster R-CNN V2. YOLO26x showed the numerically highest processing speed (**Table 3**). Furthermore, the detection results of the four models applied to an

intraoperative ascites cytology video from a patient with ovarian clear cell carcinoma, selected from the independent test video dataset, are presented in **Supplementary Video S1**. The detection results were displayed synchronously using the same video and the model-specific F1-optimized confidence thresholds, allowing visual comparison of differences in the locations and frequency of detected malignant cell candidates and frame-to-frame detection behavior among the models.

DISCUSSION

In this study, four object detection models were evaluated using independent effusion cytology image and microscopic video datasets without patient overlap to assess BBox-level detection performance for malignant cells and case-level AI classification performance. The results revealed differences among the models in the relationships between F1-score, recall, precision, and FP/image in static images, as well as substantial differences in the balance between sensitivity and specificity in videos. Under the F1-optimized threshold condition, RT-DETRv2 (R50vd), which achieved the highest F1-score in static images, showed a sensitivity of 100.0% but a specificity of 0.0% in videos. In contrast, Faster R-CNN V2 achieved a video-based sensitivity of 82.5% and specificity of 90.0%. These findings suggest that model selection based solely on a single static-image performance metric is insufficient to characterize classification performance in microscopic videos and that multiple evaluation metrics tailored to the intended application are necessary.

Under the F1-optimized threshold condition, the F1-scores on the independent test dataset comprising 5136 images and 7645 GT BBoxes of malignant cells and cell clusters were 0.795 for YOLO26x, 0.755 for RF-DETR-L, 0.808 for RT-DETRv2 (R50vd), and 0.767 for Faster R-CNN V2. However, the 95% CIs for all six pairwise differences in F1-score, estimated using patient-level stratified cluster bootstrapping, included 0, indicating that superiority cannot be established solely from the rankings of point estimates. RF-DETR-L achieved the highest recall (0.838) but also the highest FP/image (0.567), whereas Faster R-CNN V2 achieved the highest precision (0.905) and

lowest FP/image (0.104), but had a recall of 0.665. Zhang et al. reported that the rankings of models based on AP and F1-score did not necessarily agree in cervical cytology [6]. Similarly, Leordean et al. [15] and Li et al. [16] demonstrated that detection performance in pathology and cytology images varies according to the detection targets and evaluation metrics. In the present study, evaluating both missed detections of malignant cells and the occurrence of FP BBoxes, rather than relying solely on aggregate detection performance, was important for understanding the characteristics of each model. The representative detection results of the four models on the same ascites cytology image from a patient with pancreatic adenocarcinoma, shown in **Fig. 3**, provide a visual illustration of these differences; however, the relative performance of the models across the entire dataset should not be inferred from a single image.

A key feature of this study is that the same four models were evaluated separately for BBox-level detection performance in static images and case-level classification performance in continuous microscopic videos. In static images, true positives (TP), FP, and false negatives (FN) were defined based on the IoU between predicted and GT BBoxes. In contrast, frame-level GT BBoxes were not established for videos because of challenges associated with changes in focus, cellular overlap, and microscope stage movement. Instead, a qualified AI alert event was defined as the detection of the same candidate in at least three consecutive frames within a stationary-stage interval. Furthermore, a case was classified as AI-positive if at least one qualified AI alert event occurred. Therefore, case-level sensitivity in videos does not represent the detection rate of individual malignant cells, and the F1-score from static-image analysis cannot be directly compared with case-level sensitivity or specificity from video analysis. Guemas et al. distinguished between object-level and patient-level evaluations in microscopic images of malaria parasites [13], while Rad et al. highlighted the importance of the temporal distribution of detections in videos [14]. In the present study, under the F1-optimized threshold condition, YOLO26x, RF-DETR-L, and RT-DETRv2 (R50vd) each achieved a case-level sensitivity of 100.0% in videos, but their specificities were low at 32.5%, 22.5%, and 0.0%,

respectively. In contrast, Faster R-CNN V2 achieved a sensitivity of 82.5% and a specificity of 90.0%. These findings highlight the need to evaluate model performance using video-specific classification criteria in addition to localization performance in static images.

FP detections in videos have different implications from individual FP BBoxes in static images. In this study, qualified AI alert events, defined as detections of the same candidate cell in at least three consecutive frames during stationary-stage intervals, were tracked in 40 negative cases, and their frequency per minute was calculated. Under the F1-optimized threshold condition, the median FP/min frequencies were 3.0 for both YOLO26x and RF-DETR-L, 12.5 for RT-DETRv2 (R50vd), and 0.0 for Faster R-CNN V2. Previous studies in effusion cytology have identified the potential for non-malignant cells, such as reactive mesothelial cells and histiocytes, to cause FP detections [2-4]. However, because cell-level GT BBoxes were not annotated for individual detection candidates in the videos, the morphological causes of individual alerts could not be determined. Furthermore, FP/image in static images and FP/min in videos represent different evaluation metrics. Frequent FP alerts may increase the workload associated with reviewing AI-detected candidates and affect users' trust in AI-generated alerts; however, these effects on users were not evaluated in this study. **Supplementary Video S1** illustrates the detection behavior of the four models on the same intraoperative ascites cytology video from a patient with ovarian clear cell carcinoma, providing a visual supplement to the quantitative results. Before clinical implementation, further studies are needed to identify conditions that balance the reduction of missed detections with the control of alert frequency and to evaluate the workload imposed on microscopic observers and the reliability of AI-assisted detection.

In this study, the best checkpoints and confidence thresholds were selected using the validation dataset, without further adjustment based on the independent test results. The F1-optimized thresholds used in the primary analysis differed among models: 0.45 for YOLO26x and RF-DETR-L, 0.65 for RT-DETRv2 (R50vd), and 0.90 for Faster R-CNN V2. Moreover, threshold dependence was evaluated through secondary analyses using a fixed threshold of 0.50 and high-recall thresholds.

Sevilla et al. reported an approach for selecting confidence thresholds using validation data according to the intended application [17], while Xu et al. [18] reported that pretrained weights, multi-scale training, and fine-tuning can influence detection performance in cytology. Furthermore, YOLO-based models, which predict candidate boxes from numerous spatial locations in image feature maps, and Transformer-based models, which use a limited number of object queries to predict objects, may differ in their sensitivity to post-processing conditions [19]. In the present study, the performance of four object detection models was compared using the same patient-level dataset partitioning and evaluation procedures while retaining the model-specific training and inference settings of each algorithm. Therefore, the observed differences in performance cannot be attributed solely to architectural differences, and confidence scores may not have equivalent interpretations across models. Moreover, because confidence thresholds were selected using the validation image dataset, they were not directly optimized for case-level specificity or alert frequency in videos. Optimizing confidence thresholds and temporal alert criteria for video-based applications will require an independent video validation cohort, followed by further evaluation using a separate test cohort.

For applications in AI-assisted microscopic examination, evaluating processing speed is essential in addition to missed detections of malignant cells and FP alerts. In this study, the median effective processing speeds, including video loading and model inference, were 35.57 FPS for YOLO26x, 22.82 FPS for RF-DETR-L, 19.55 FPS for RT-DETRv2 (R50vd), and 19.85 FPS for Faster R-CNN V2. Although YOLO26x achieved the highest processing speed, Faster R-CNN V2 showed the highest specificity in videos, indicating that processing speed alone is insufficient for selecting a model for practical applications. Terasaki et al. reported a real-time AI-assisted cytology system using an optical microscope and camera; however, their system and the present study differed in imaging setup, hardware, input conditions, and the scope of processing time measurements, limiting direct comparisons of processing speed [9]. Furthermore, because software implementations and preprocessing and post-processing procedures differed among the models, FPS comparisons should be

interpreted descriptively. The effective FPS measured in this study reflects processing performance on prerecorded videos and does not establish end-to-end latency from image acquisition by the microscope camera to the display of detection BBoxes, or guarantee the responsiveness experienced by users during actual microscopic examination. Future studies should evaluate frame latency, image quality, detection stability, and usability in a live microscopy setting.

A major strength of this study is that four different models were evaluated on the same test datasets, with no patient overlap among the training, validation, independent image test, and independent video test datasets and with separate specimen collection periods. Furthermore, the independent image test dataset comprised 168 cases, 5136 images, and 7645 GT BBoxes of malignant cells. Uncertainty in the performance estimates was assessed using patient-level stratified cluster bootstrapping to account for the clustering of multiple images within individual patients. For video evaluation, the same classification criteria were applied to 40 malignant and 40 negative cases, allowing assessment of not only case-level sensitivity and specificity but also FP alert event frequency. Nevertheless, this study had several limitations. First, all data were obtained from a single institution, and the generalizability of the findings to other institutions with different specimen preparation methods, staining procedures, imaging equipment, and patient populations remains unverified. Second, all videos were recorded for 60 seconds per case by a single cytotechnologist with 26 years of experience. Consequently, interobserver variability in scanning speed, field selection, and focus adjustment could not be evaluated. Third, because cell-level GT BBoxes were not available for the videos, AI-positive classifications in malignant cases could not be reliably distinguished as arising from correct localization of malignant cells or FP detections of non-malignant cells. Finally, the requirement for detection in three consecutive frames and the stage-stationarity threshold were predefined for this study, and their robustness under alternative settings has not been evaluated. Considering these limitations, the present findings should be interpreted as measures of technical performance under the specific implementation and observation conditions used in this study.

Future research should focus on optimizing alert criteria using a dedicated video validation dataset, conducting external validation with multicenter data, and analyzing FP detections according to cell size, single cells versus cell clusters, and the types of benign cells. Furthermore, the present study evaluated the detection and case-level classification performance of AI alone. It remains unclear whether cytotechnologists reviewing AI-identified candidates can reduce missed detections of malignant cells or whether FP alerts may instead increase interpretation time and workload. Although previous studies have demonstrated the potential of AI-assisted cytology [9,10,20,21], the effectiveness of AI assistance may also depend on users and clinical workflows. Therefore, future studies should compare AI-assisted and unassisted video-based and microscopic examinations in terms of malignant cell detection, case-level classification accuracy, examination time, workload, and usability. The present study demonstrated that four object detection models for malignant cells in effusion cytology exhibited distinct performance characteristics not only in static-image detection but also in case-level classification performance, FP alert frequency, and processing speed in videos. These findings suggest that selecting an appropriate AI model for microscopic examination support requires a multidimensional performance evaluation tailored to the intended application rather than ranking models according to a single metric.

METHODS

Study design

This study was approved by the Bioethics Review Committee, Nagoya University Graduate School of Medicine and Nagoya University Hospital (approval number: 2023-0096-2; June 17, 2025). The requirement for written informed consent was waived by the Bioethics Review Committee because information about the study was publicly disclosed and patients were provided with an opportunity to opt out of participation. All research procedures were conducted in accordance with the Declaration of Helsinki and the Ethical Guidelines for Medical and Biological Research Involving Human Subjects

of Japan. Effusion specimens, including pleural effusions, ascites, pericardial effusions, and intraoperatively collected effusion and lavage fluids, obtained at Nagoya University Hospital between January 1, 2018, and December 31, 2025, were identified and retrieved through a search of the electronic pathology database. Specimens collected from patients aged <18 years were excluded. Cytological diagnoses of malignancy or benignity were established by at least one cytotechnologist and at least one cytopathologist. Pathological diagnoses were based on histopathological examination of surgical biopsy or resection specimens from the primary tumor site or analysis of cell blocks prepared for immunohistochemical staining. Cases with a pathological diagnosis of adenocarcinoma confirmed before August 2026 were included. For tumors of ovarian origin, eligible histological diagnoses comprised serous carcinoma, endometrioid carcinoma, clear cell carcinoma, and mucinous carcinoma. When multiple effusion specimens had been collected from the same patient during the study period, only the first submitted specimen was included, and subsequent specimens were excluded. The dataset used for deep learning model development comprised effusion specimens collected between January 1, 2018, and December 31, 2022. The independent image test dataset comprised specimens collected between January 1, 2023, and December 31, 2024, whereas the independent video test dataset comprised specimens collected between January 1, 2025, and December 31, 2025. These three datasets were derived from distinct collection periods and had no patient overlap. The model development dataset was further divided into training and validation datasets at the patient level, ensuring that no patients were shared between these subsets. This study was designed and reported with reference to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) 2024 Update [22] and the Standards for Reporting Diagnostic Accuracy Studies Using Artificial Intelligence (STARD-AI) [23].

Preparation of image and video datasets

Cytology specimens were prepared using a cytocentrifugation method (Sakura Finetek Japan Co., Ltd., Japan), fixed in 95% ethanol, and subsequently stained using the Papanicolaou method. Cytology

images were acquired using a Basler Microscopy Pulse 5.0 MP camera (Basler AG, Germany) at magnifications of ×200 or ×400 and saved as JPEG files with a resolution of 2592 × 1944 pixels.

For the model development and independent image test datasets, individual malignant cells and malignant cell clusters were each counted as one detection target. When fewer than 30 targets were present on a glass slide, all targets were selected for imaging. When 30 or more targets were present, at least 30 targets were randomly selected and imaged. All malignant cells and/or cell clusters within the acquired images were annotated using the open-source image annotation tool labelImg (version 1.8.6). Annotations were performed by two cytotechnologists and subjected to double-checking. The numbers of cases, images, and BBoxes, as well as patient characteristics and specimen types, in the static-image datasets used in this study are summarized in **Table 4**. The training dataset comprised 314 cases and 19,657 images; the validation dataset comprised 78 cases and 5201 images; and the independent image test dataset comprised 168 cases and 5136 images. In the model development dataset, the distributions of primary tumor sites were balanced between the training and validation datasets, with no patient overlap between them. The numbers of BBoxes per image are presented as counts or as medians with interquartile ranges (IQRs; first quartile [Q1] to third quartile [Q3]). BBox counts were calculated from the annotation files in YOLO format. Images without malignant cell annotations were assigned a BBox count of 0 when calculating medians and quartiles.

The specimen types, patient characteristics, and primary tumor sites of malignant cases in the video test dataset are summarized in Table 4. A cytotechnologist with 26 years of professional experience recorded one video for each of 80 cases (40 malignant and 40 negative cases). Videos were recorded using a Basler pulse puA2500-14ucBAS camera (Basler AG, Germany) at ×200 magnification and an original frame rate of 12 FPS. The first 60 seconds of each video were extracted using Adobe Premiere Pro 2026 (Adobe Inc., San Jose, CA, USA), without altering the playback speed or frame rate. All videos used for analysis consisted of 720 frames recorded over 60 seconds. Each video was recorded continuously while the microscope stage was moved and stopped, resulting in

overlapping fields of view. In microscopic videos, changes in focus, cellular overlap, partial visualization of cells at the edges of the field of view, and motion blur caused by microscope stage movement can make morphological assessment of individual cells and consistent assignment of GT BBoxes difficult. To avoid the potential effects of uncertainty in GT annotations on performance evaluation, cell-level GT BBoxes were not created for video frames. Instead, the reference standard was the case-level cytological diagnosis of malignancy or negativity. The consistency of all videos with the reference standard was verified by two experts. During microscope stage movement, the positions of BBoxes change between consecutive frames because of cell displacement and motion blur. Therefore, predefined criteria for stationary fields of view and persistent detection of the same candidate cell were applied to case-level AI-positive classification. Each case was evaluated to determine whether at least one detection event satisfying the predefined AI alert criteria occurred.

Deep learning models and confidence thresholds

This study was conducted using a workstation running Windows 11 (Microsoft Corporation, Redmond, WA, USA), equipped with an AMD Ryzen 9 9950X3D central processing unit (CPU; Advanced Micro Devices, Inc., Santa Clara, CA, USA), an NVIDIA RTX 5090 graphics processing unit (GPU; 32 GB; NVIDIA Corporation, Santa Clara, CA, USA), and 128 GB of system memory. Four object detection architectures—YOLO26x [24], RF-DETR-L [25], RT-DETRv2 (R50vd) [26], and Faster R-CNN V2 [27,28]—were used to develop separate single-class deep learning models for malignant cell detection in Anaconda virtual environments. The models were implemented using the respective frameworks provided by Ultralytics, Roboflow, the official RT-DETR repository, and PyTorch torchvision. Transfer learning was performed using pretrained weights specific to each model as initialization. The model architectures, backbones, pretrained weights, and training configurations are summarized in **Supplementary Table S7**. For comparisons among models, identical patient-level dataset partitions were used, while the training procedures and architecture-specific preprocessing and post-processing methods were retained.

The best checkpoint selected for each algorithm was used to evaluate the trained models. For RF-DETR-L and RT-DETRv2 (R50vd), weights obtained using an exponential moving average (EMA) of the model parameters during training were used for inference. For YOLO26x and Faster R-CNN V2, checkpoints saved by their respective training frameworks were used. Algorithm-specific image preprocessing and post-processing procedures were retained during inference. For YOLO26x, the one-to-many prediction pathway was used with non-maximum suppression (NMS), an IoU threshold of 0.70, and a maximum of 300 detections per image. RF-DETR-L used the standard post-processing procedures implemented in the RF-DETR library, whereas RT-DETRv2 (R50vd) used the post-processing procedures provided by the official implementation, including Top-K selection. No additional NMS was applied to RT-DETRv2 (R50vd). For Faster R-CNN V2, NMS implemented in torchvision was applied with an IoU threshold of 0.50 and a maximum of 300 detections per image. Detection results from all models were saved in a common format containing BBox coordinates, class IDs, and confidence scores and were analyzed using the same evaluation program. However, model-specific post-processing procedures differed in candidate BBox generation, removal of duplicate predictions, and limitations on the number of output detections. Consequently, applying the same confidence threshold across models did not necessarily result in identical BBox output conditions. Therefore, detection performance was compared using common evaluation datasets, confidence threshold selection procedures, and evaluation metrics while retaining the model-specific post-processing procedures. For the primary analysis, F1-optimized confidence thresholds that maximized the F1-score were used. Two additional threshold conditions were evaluated in secondary analyses: a fixed confidence threshold of 0.50 for all models and high-recall thresholds selected to minimize FP/image while maintaining BBox-level recall for malignant cells and cell clusters of at least 95%. All thresholds were determined using the validation dataset and fixed before evaluation of the independent test datasets. No further adjustments were made based on the results obtained from the independent test images or videos.

Criteria for case-level AI classification in videos

To identify stationary-stage intervals, each video frame was resized to a width of 640 pixels, converted to grayscale, and subjected to Gaussian smoothing. Global image displacement between consecutive frames was estimated using a feature-based image registration method [29]. When sufficient feature tracking could not be achieved, displacement was estimated using a Fourier transform-based phase correlation method [30]. Predefined thresholds for interframe displacement were applied to identify stationary-stage intervals. Consecutive intervals with displacement of 1.5 pixels or less were considered candidate stationary intervals if they contained at least one interframe displacement of 0.6 pixels or less. When two stationary intervals were separated by an unstable transition corresponding to a single frame, they were merged into one stationary interval if the displacement was 2.25 pixels or less. Intervals indicating clear microscope stage movement were not merged. All displacement thresholds were applied to consecutive frames resized to a width of 640 pixels. Within stationary-stage intervals, BBoxes were tracked using IoU-based association, as commonly employed in video object tracking [31]. BBoxes in consecutive frames with an IoU of at least 0.50 were considered potential matches. When multiple matching candidates were present, one-to-one associations that maximized the sum of IoU values were determined using the Hungarian algorithm [32]. No interpolation was performed to bridge missed detections, and tracking across stationary-stage intervals was not permitted. To suppress transient detections, a qualified AI alert event was defined as the detection of the same candidate cell in at least three consecutive frames within a stationary-stage interval. At 12 FPS, detection in three consecutive frames corresponds to a time interval of approximately 0.167 seconds between the first and third detection frames. This criterion was a study-specific classification rule informed by the concept of temporal confirmation of detections and has not been established as the optimal number of consecutive detection frames for cytology. A video was classified as AI-positive if at least one qualified AI alert event occurred during the recording and as AI-negative if no such event occurred. These classification criteria were applied consistently across all four deep learning

models. AI-based case-level classifications were compared with the reference diagnoses to determine TP, FN, true negatives (TN), and FP. Case-level sensitivity, specificity, and accuracy were then calculated from these values. For the 40 negative cases, the total number of qualified AI alert events was counted in each video. Because all videos were 60 seconds long, the number of qualified AI alert events per video was expressed as FP/min. Moreover, FP alert event frequency in negative videos and effective processing speed, including video loading and model inference (effective processing FPS), were evaluated.

Statistical analysis of static-image detection performance

The detection performance of the four deep learning models was evaluated using the independent test dataset comprising 168 cases, 5136 images, and 7645 GT BBoxes of malignant cells and cell clusters. Predicted BBoxes were matched to GT BBoxes based on IoU, with an IoU threshold of 0.50 used to identify potential matches. Predicted BBoxes were processed in descending order of confidence score and matched one-to-one to the unmatched GT BBox with the highest IoU, provided that the IoU was at least 0.50. Matched predicted BBoxes were classified as TP, unmatched predicted BBoxes as FP, and GT BBoxes that were not matched to any predicted BBox as FN. Multiple predicted BBoxes were not permitted to match the same GT BBox. Precision, recall, and F1-score were calculated from TP, FP, and FN for each model. FP/image was calculated by dividing the total number of FP detections by the total number of evaluated images. These metrics were calculated using TP, FP, and FN aggregated across the entire independent test dataset. The 95% CIs for each performance metric were estimated using nonparametric patient-level stratified cluster bootstrapping [33,34]. To account for dependencies among multiple images and BBoxes derived from the same patient, patients were used as the resampling units. To preserve the original distribution of 149 malignant and 19 negative cases, patients were stratified into malignant and negative groups, and resampling with replacement was performed separately within each group. In total, 2000 bootstrap replicates were generated, with all images and BBoxes belonging to each resampled patient included in the analysis. In each replicate, TP, FP, FN,

and image counts were aggregated across the resampled patients, and precision, recall, F1-score, and FP/image were recalculated. The 2.5th and 97.5th percentiles of the bootstrap distributions were used as the 95% confidence limits [35]. For pairwise comparisons among the four models, the same resampled patient sets were used across all models in each bootstrap replicate to preserve the paired structure of the data. Differences in precision, recall, F1-score, and FP/image between models were calculated for each replicate, and their 95% CIs were derived from the corresponding bootstrap distributions. These between-model differences and their 95% CIs were treated as estimation results without adjustment for multiple comparisons. Furthermore, AP was calculated according to the COCO evaluation protocol as a measure of object detection performance across confidence thresholds. The primary AP metrics were AP50:95 and AP50. AP was calculated using pycocotools, with a maximum of 100 detections per image. AP values were reported as point estimates calculated across the entire independent test dataset and were not included in the patient-level bootstrap analyses for estimating 95% CIs. Statistical analyses were performed using Python and NumPy. The random seed for bootstrapping was set to 20261014.

Statistical analysis of video-based detection performance

The 95% CIs for case-level sensitivity, specificity, and accuracy were calculated using the Wilson score method. Sensitivity and specificity were calculated using the 40 malignant cases and 40 negative cases as their respective denominators, whereas accuracy was calculated using all 80 cases. To compare case-level detection performance among the four models while accounting for paired observations from the same cases, Cochran's Q tests were performed separately for sensitivity and specificity. Pairwise comparisons between models were performed using two-sided exact McNemar tests, and P values were adjusted using the Holm method for the six pairwise comparisons within each performance metric. Comparisons of sensitivity and specificity were treated as separate families of multiple comparisons. Moreover, between-model differences in sensitivity and specificity were estimated in pp using 20,000 nonparametric bootstrap replicates while preserving the paired structure

of the four models within each case. The 40 malignant cases and 40 negative cases were resampled with replacement separately. The 95% CIs for between-model differences were determined using the percentile method, based on the 2.5th and 97.5th percentiles of the bootstrap distributions. FP alert event frequencies in the 40 negative cases were summarized as medians and IQRs, as well as means and standard deviations (SDs). Differences among the four models were evaluated using the Friedman test, followed by pairwise comparisons using two-sided Wilcoxon signed-rank tests. Cases with zero between-model differences were excluded from the ranking procedure in the Wilcoxon tests. Comparisons of FP alert event frequencies were conducted separately for each confidence threshold condition, with multiple-comparison adjustment applied to the six pairwise comparisons within each condition. Effective processing FPS was summarized for each model across the 80 cases using the median (IQR) and mean ± SD. Because the processing environments and software implementations differed among the models, processing speed comparisons were limited to descriptive statistics. All statistical tests were two-sided, with $P < 0.05$ considered significant. Statistical analyses were performed using Python with NumPy, SciPy, statsmodels, and pandas.

Use of generative AI in manuscript preparation

ChatGPT (OpenAI) was used to assist with manuscript organization, language editing, and refinement of scientific writing. The authors reviewed and verified the final manuscript and take full responsibility for its content.

Data availability

The patient-derived images and videos analyzed in this study are not publicly available due to patient privacy and ethical restrictions. Sharing these data with third parties was not covered by the approved research protocol. Aggregated results supporting the findings of this study are provided in the article and its Supplementary Information.

Code availability

The custom code developed for model training, inference, and performance evaluation in this study is not publicly available to protect intellectual property associated with the research. Relevant code may be made available by the corresponding author upon reasonable request for non-commercial academic research and verification purposes, subject to applicable access conditions.

References

[1] Wang, C. W., Muzakky, H., Chung, Y. P., Lai, P. J. & Chao, T. K. Interpretable multi-scale deep learning to detect malignancy in cell blocks and cytological smears of pleural effusion and identify aggressive endometrial cancer. *Med. Image Anal.* **106,** 103742 (2025). https://doi.org/10.1016/j.media.2025.103742

[2] Xie, X. et al. Deep convolutional neural network-based classification of cancer cells on cytological pleural effusion images. *Mod. Pathol.* **35,** 609–614 (2022). https://doi.org/10.1038/s41379-021-00987-4

[3] Sanyal, P. & Dey, P. Using a deep learning neural network for the identification of malignant cells in effusion cytology material. *Cytopathology* **34,** 466–471 (2023). https://doi.org/10.1111/cyt.13260

[4] Ikeda, K. et al. Deep learning neural network of adenocarcinoma detection in effusion cytology. *Am. J. Clin. Pathol.* **164,** 415–423 (2025). https://doi.org/10.1093/ajcp/aqaf067

[5] Tian, F. et al. Prediction of tumor origin in cancers of unknown primary origin with cytology-based deep learning. *Nat. Med.* **30,** 1309–1319 (2024). https://doi.org/10.1038/s41591-024-02915-w

[6] Zhang, X. et al. A large annotated cervical cytology images dataset for AI models to aid cervical cancer screening. *Sci. Data* **12,** 23 (2025). https://doi.org/10.1038/s41597-025-04374-5

[7] Niu, S., Zhang, L., Wang, L., Zhang, X. & Liu, E. Hybrid feature fusion in cervical cancer cytology: a novel dual-module approach framework for lesion detection and classification using radiomics, deep learning, and reproducibility. *Front. Oncol.* **15,** 1595980 (2025). https://doi.org/10.3389/fonc.2025.1595980

[8] Sun, M., Zhang, J., Zhao, S., Li, X. & Zhuo, L. Cell cluster detection of thyroid FNAB-WSI via deformable convolution with frequency channel attention. *Biomed. Signal Process. Control* **100,** 107023 (2025). https://doi.org/10.1016/j.bspc.2024.107023

[9] Terasaki, M. et al. An integrated system from microscopy to AI for real-time object detection in endometrial cytology. *J. Pathol. Inform.* **20,** 100541 (2026).

https://doi.org/10.1016/j.jpi.2025.100541

[10] Tang, H. P. et al. Cervical cytology screening facilitated by an artificial intelligence microscope: a preliminary study. *Cancer Cytopathol.* **129,** 693–700 (2021). https://doi.org/10.1002/cncy.22425

[11] Satturwar, S., Parwani, A. V. & Li, Z. Advances in digital cytopathology and artificial intelligence applications. *Adv. Anat. Pathol.* **33,** 190–200 (2026). https://doi.org/10.1097/PAP.0000000000000529

[12] Hang, J. F. et al. Evaluating urine cytology slide digitization efficiency: a comparative study using an artificial intelligence-based heuristic scanning simulation and multiple Z-plane scanning. *Acta Cytol.* **68,** 342–350 (2024). https://doi.org/10.1159/000538985

[13] Guemas, E. et al. Automatic patient-level recognition of four Plasmodium species on thin blood smear by a real-time detection transformer (RT-DETR) object detection algorithm: a proof-of-concept and evaluation. *Microbiol. Spectr.* **12,** e01440-23 (2024). https://doi.org/10.1128/spectrum.01440-23

[14] Rad, I. H. & Taghizadeh, S. S-MEOD: a novel evaluation metric for frame-based medical object detection. *J. Imaging Inform. Med.* (2026). https://doi.org/10.1007/s10278-026-01957-5

[15] Leordean, D. V. & Ardelean, E. R. Object detection in histology: a multi-dataset benchmark and test-time inference. *PLoS ONE* **21,** e0354618 (2026). https://doi.org/10.1371/journal.pone.0354618

[16] Li, G. et al. Detection of cervical cell based on multi-scale spatial information. *Sci. Rep.* **15,** 3117 (2025). https://doi.org/10.1038/s41598-025-87165-7

[17] Fernández de Sevilla, M. T. et al. Comparative evaluation of YOLOv8, YOLOv11, and RT-DETR for automated microcarrier colonization assessment in bioreactor cell cultures. *Front. Bioeng. Biotechnol.* **14,** 1849366 (2026). https://doi.org/10.3389/fbioe.2026.1849366

[18] Xu, C. et al. Cervical cell/clumps detection in cytology images using transfer learning. *Diagnostics* **12,** 2477 (2022). https://doi.org/10.3390/diagnostics12102477

[19] Xie, L. et al. Performance trade-offs between dense prediction and sparse query mechanisms for

brain tumor MRI detection: a comparative study of YOLOv8 and RT-DETR. *Front. Med. (Lausanne)* **13,** 1905061 (2026). https://doi.org/10.3389/fmed.2026.1905061

[20] Xue, P. et al. Deep learning-assisted versus manual reading in routine cervical cytopathology: a multicentre randomised crossover trial. *npj Digit. Med.* **9,** 422 (2026). https://doi.org/10.1038/s41746-026-02703-5

[21] Bao, H. et al. Artificial intelligence-assisted cytology for detection of cervical intraepithelial neoplasia or invasive cancer: a multicenter, clinical-based, observational study. *Gynecol. Oncol.* **159,** 171–178 (2020). https://doi.org/10.1016/j.ygyno.2020.07.099

[22] Tejani, A. S. et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiol. Artif. Intell.* **6,** e240300 (2024). https://doi.org/10.1148/ryai.240300

[23] Sounderajah, V. et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* **31,** 3283–3289 (2025). https://doi.org/10.1038/s41591-025-03953-8

[24] Jocher, G. et al. Ultralytics YOLO26: unified real-time end-to-end vision models. Preprint at https://arxiv.org/abs/2606.03748 (2026).

[25] Robinson, I., Robicheaux, P., Popov, M., Ramanan, D. & Peri, N. RF-DETR: neural architecture search for real-time detection transformers. *In: International Conference on Learning Representations* (2026).

[26] Lv, W. et al. RT-DETRv2: improved baseline with bag-of-freebies for real-time detection transformer. Preprint at https://arxiv.org/abs/2407.17140 (2024).

[27] Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39,** 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031

[28] Li, Y. et al. Benchmarking detection transfer learning with vision transformers. Preprint at https://arxiv.org/abs/2111.11429 (2021).

[29] Lucas, B. D. & Kanade, T. An iterative image registration technique with an application to stereo vision. *In: Proc. 7th Int. Joint Conf. Artif. Intell.*, 674–679 (1981).

[30] Reddy, B. S. & Chatterji, B. N. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Trans. Image Process.* **5,** 1266–1271 (1996). https://doi.org/10.1109/83.506761

[31] Bewley, A., Ge, Z., Ott, L., Ramos, F. & Upcroft, B. Simple online and realtime tracking. *In: Proc. IEEE Int. Conf. Image Process.*, 3464–3468 (2016). https://doi.org/10.1109/ICIP.2016.7533003

[32] Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2,** 83–97 (1955). https://doi.org/10.1002/nav.3800020109

[33] Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7,** 1–26 (1979). https://doi.org/10.1214/aos/1176344552

[34] Field, C. A. & Welsh, A. H. Bootstrapping clustered data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69,** 369–390 (2007).

[35] Efron, B. Nonparametric standard errors and confidence intervals. *Can. J. Stat.* **9,** 139–172 (1981). https://doi.org/10.2307/3314608

Acknowledgements

We thank Editage (www.editage.jp) for assistance with English language editing.

Author contributions

K.I. conceptualized and designed the study; K.I. wrote the manuscript; K.I. collected specimens and analyzed the data; K.I. and K.F. annotated the data; K.I. contributed to algorithm development and programming; N.S., K.M., M.M., S.S., T.H., H.K., M.N., and K.K. reviewed and edited the manuscript. All the authors read and approved the final manuscript.

Ethics approval

This study was approved by the Institutional Ethics Committee of Nagoya University Hospital (approval number 2023-0096-2, June 17, 2025).

Informed consent

The requirement for informed consent was waived by the institutional ethics committee because information about the study was publicly disclosed and patients were provided with an opportunity to opt out of participation.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [Grant Number JP26K15660] and by a research grant from Takahashi Industrial and Economic Research Foundation.

724

725 **Supplementary Information**

726 Supplementary Information accompanies this manuscript and includes Supplementary Figs. S1–S3,

727 Supplementary Tables S1–S7, and Supplementary Video S1.

728

Figure

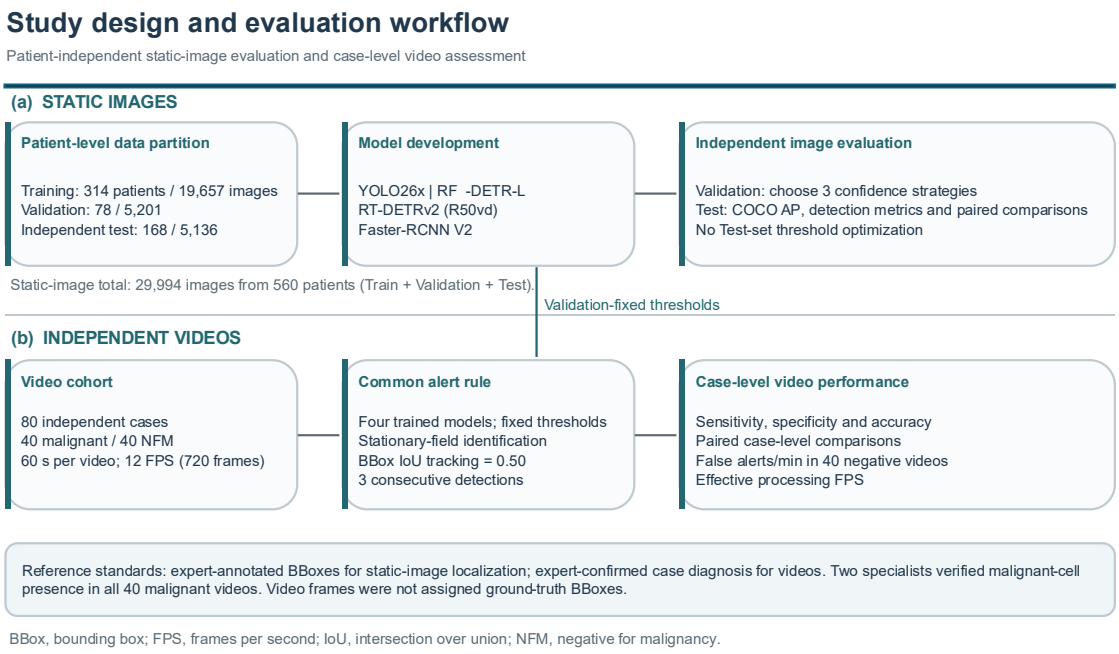

Figure 1. Dataset composition and evaluation workflow for deep learning models in this study. This study compared the performance of four object detection algorithms—YOLO26x, RF-DETR-L, RT-DETRv2 (R50vd), and Faster R-CNN V2—for detecting malignant cells in cytopathology images and videos. **(a) Image-based analysis:** A total of 29,994 effusion cytology images from 560 patients were divided at the patient level into training (314 patients, 19,657 images), validation (78 patients, 5201 images), and independent test (168 patients, 5136 images) datasets. The four models were trained using the training dataset, and three confidence thresholds (fixed, F1-optimized, and high-recall) were determined using the validation dataset. Detection performance was evaluated on the independent test dataset. **(b) Video-based analysis:** Videos from 80 patients (40 malignant and 40 negative), independent of the image datasets, were analyzed. Each video was 60 seconds long and recorded at 12 frames per second (FPS). The confidence thresholds determined from the validation dataset were applied to all four models. A case was classified as positive when the same candidate cell was detected in at least three consecutive frames within a stationary-stage interval. An intersection over union (IoU) threshold of 0.50 was used for bounding box (BBox) tracking. Reference diagnoses were established

by at least one cytotechnologist and one cytopathologist, and all videos and their reference diagnoses were reviewed by at least two experts. Because individual video frames were not annotated with ground truth BBoxes, evaluation was performed at the case level using sensitivity, specificity, accuracy, false-positive alert frequency in negative videos, and effective processing speed.

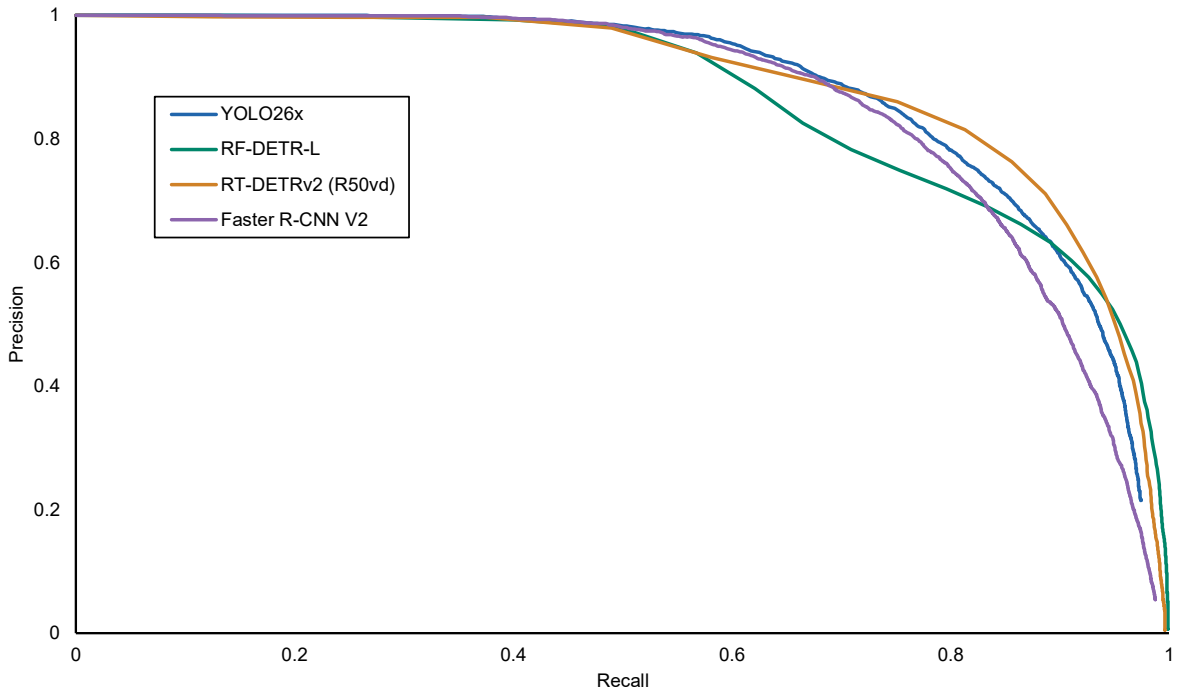

Figure 2. Precision–Recall curves of four deep learning–based object detection models on the independent test image dataset.

Precision–Recall (PR) curves are shown for YOLO26x (blue), RF-DETR-L (green), RT-DETRv2 (R50vd) (orange), and Faster R-CNN V2 (purple) on the independent test dataset comprising 168 patients, 5136 images, and 7645 ground truth bounding boxes (GT BBoxes) of malignant cells. The horizontal and vertical axes represent recall and precision, respectively. An intersection over union (IoU) threshold of 0.50 was used to match predicted BBoxes with GT BBoxes. PR curves were generated by ranking detection candidates according to their confidence scores for each model. Model-specific inference post-processing procedures were retained, and all models were evaluated using the same test dataset and BBox matching criteria.

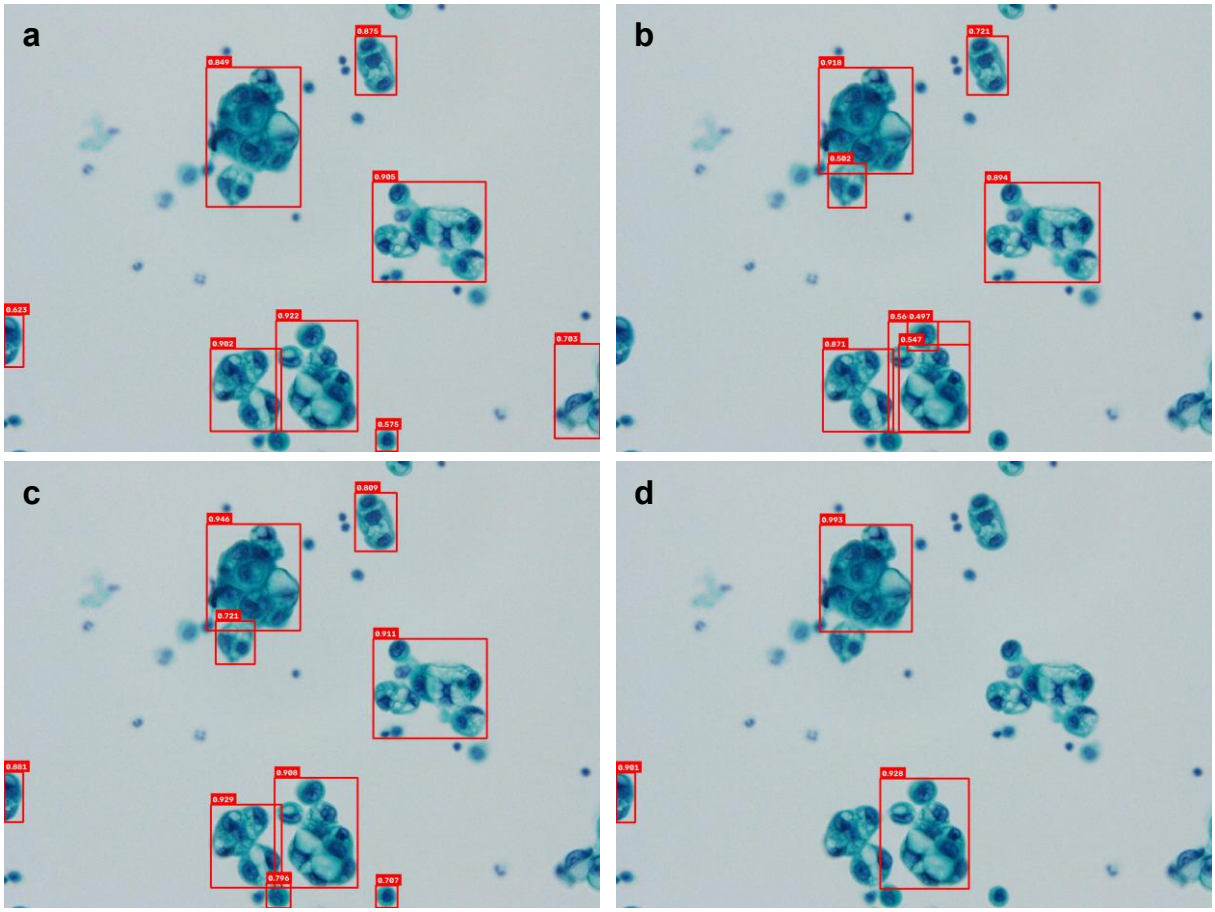

Figure 3. Representative detection results of four object detection algorithms on an image from the independent test dataset.

Detection results of four deep learning–based object detection algorithms on the same image selected from the independent test dataset are shown. The image was obtained from an ascites specimen from a patient with pancreatic adenocarcinoma. Panels show the results for (a) YOLO26x (threshold, 0.45), (b) RF-DETR-L (threshold, 0.45), (c) RT-DETRv2 (R50vd) (threshold, 0.65), and (d) Faster R-CNN V2 (threshold, 0.90). For each model, detections were visualized using the F1-optimized confidence threshold determined from the validation dataset. Regions detected as malignant cells are indicated by red bounding boxes. The number shown at the upper-left corner of each box represents the confidence score. Comparison of the detection results on the same image visually illustrates differences among the algorithms in the number and locations of detected malignant cells.

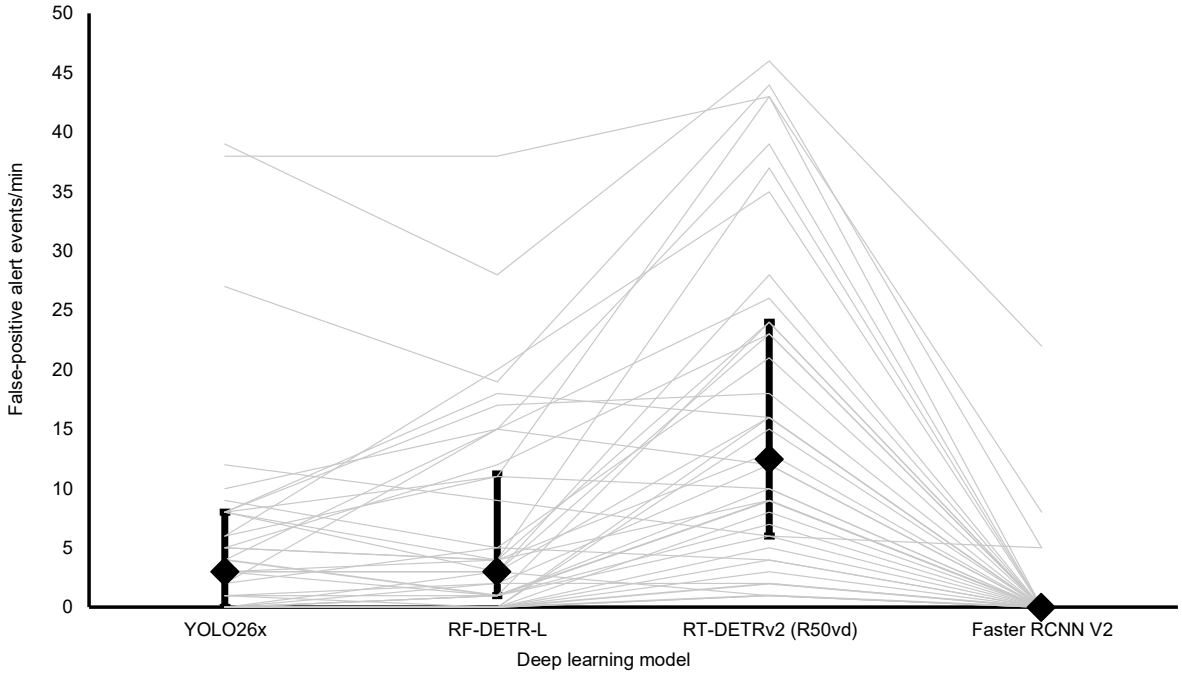

Figure 4. Comparison of false-positive alert event frequencies among four deep learning–based object detection models in negative cytology videos.

False-positive alert event frequencies were compared among YOLO26x, RF-DETR-L, RT-DETRv2 (R50vd), and Faster R-CNN V2 using 40 negative cases from the independent test video dataset. Each model was evaluated using its F1-optimized confidence threshold determined from the validation image dataset (YOLO26x: 0.45; RF-DETR-L: 0.45; RT-DETRv2 [R50vd]: 0.65; Faster R-CNN V2: 0.90). A false-positive alert event was defined as a tracked detection of the same candidate cell in at least three consecutive frames within a stationary-stage interval. An intersection over union (IoU) threshold of 0.50 was used for bounding box (BBox) tracking. As each video was 60 seconds long, the number of alert events was expressed as the frequency per minute. Gray lines connect the results of the four models for each individual negative case. Black diamonds indicate the medians, and black vertical lines represent the interquartile ranges (IQRs; first to third quartiles). The Friedman test revealed a statistically significant difference in false-positive alert event frequencies among the four models ($\chi^2(3) = 84.34$, $P < 0.001$).

Table 1. Independent test performance at predefined and validation-selected confidence thresholds

Model	Threshold method	TP	FP	FN	Precision (95% CI)	Recall (95% CI)	F1-score (95% CI)	FP/image (95% CI)
YOLO26x	Fixed	5715	1008	1930	0.850 (0.719–0.936)	0.748 (0.686–0.807)	0.796 (0.724–0.853)	0.196 (0.068–0.446)
YOLO26x	F1-optimized	5857	1226	1788	0.827 (0.684–0.923)	0.766 (0.706–0.824)	0.795 (0.716–0.857)	0.239 (0.085–0.539)
YOLO26x	High-recall	7111	6347	534	0.528 (0.393–0.646)	0.930 (0.893–0.961)	0.674 (0.551–0.763)	1.236 (0.680–2.188)
RF-DETR-L	Fixed	6194	2569	1451	0.707 (0.523–0.903)	0.810 (0.774–0.844)	0.755 (0.637–0.859)	0.500 (0.122–1.082)
RF-DETR-L	F1-optimized	6403	2910	1242	0.688 (0.509–0.879)	0.838 (0.805–0.868)	0.755 (0.634–0.862)	0.567 (0.161–1.180)
RF-DETR-L	High-recall	7306	7364	339	0.498 (0.381–0.629)	0.956 (0.939–0.968)	0.655 (0.545–0.759)	1.434 (0.802–2.320)
RT-DETRv2 (R50vd)	Fixed	6506	1938	1139	0.770 (0.701–0.832)	0.851 (0.803–0.895)	0.809 (0.762–0.851)	0.377 (0.244–0.555)
RT-DETRv2 (R50vd)	F1-optimized	5896	1046	1749	0.849 (0.768–0.915)	0.771 (0.713–0.824)	0.808 (0.758–0.854)	0.204 (0.102–0.348)
RT-DETRv2 (R50vd)	High-recall	6966	3831	679	0.645 (0.590–0.697)	0.911 (0.873–0.946)	0.755 (0.713–0.795)	0.746 (0.540–1.006)
Faster R-CNN V2	Fixed	6415	3036	1230	0.679 (0.641–0.716)	0.839 (0.794–0.885)	0.750 (0.722–0.777)	0.591 (0.471–0.731)
Faster R-CNN V2	F1-optimized	5087	536	2558	0.905 (0.885–0.922)	0.665 (0.602–0.735)	0.767 (0.722–0.813)	0.104 (0.081–0.133)
Faster R-CNN V2	High-recall	7050	9555	595	0.425 (0.380–0.472)	0.922 (0.890–0.951)	0.581 (0.538–0.624)	1.860 (1.463–2.352)

Abbreviations: TP, true positive; FP, false positive; FN, false negative; CI, confidence interval; FP/image, false positives per image.

Object detection performance was evaluated on an independent test dataset comprising 5136 images from 168 patients and 7645 ground-truth bounding boxes (GT BBox) of malignant cells. Predicted bounding boxes were matched one-to-one with GT BBoxes using an intersection over union (IoU) threshold of 0.50, processing predictions in descending order of confidence. Unmatched predictions were counted as false positives (FP), and unmatched GT BBoxes were counted as false negatives (FN). Precision, recall, and F1-score were calculated from the aggregate true positive (TP), FP, and FN counts across all test images. FP/image was calculated by dividing the total number of FP detections by 5136 images. Values in parentheses represent 95% confidence intervals estimated using a stratified patient-cluster percentile bootstrap with 2000 replicates, maintaining the malignant (n = 149) and negative for malignancy (NFM) (n = 19) patient strata. Confidence thresholds were selected from the validation dataset before independent testing. The high-recall criterion refers to validation recall ≥0.95; this constraint was not reimposed or optimized on the independent test dataset.

Table 2. Case-level diagnostic performance of four deep learning models on 80 independent cytology videos using validation-derived F1-optimized confidence thresholds

Model	Threshold	TP	FN	TN	FP	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Accuracy, % (95% CI)
YOLO26x	0.45	40	0	13	27	100.0 (91.2–100.0)	32.5 (20.1–48.0)	66.25 (55.4–75.7)
RF-DETR-L	0.45	40	0	9	31	100.0 (91.2–100.0)	22.5 (12.3–37.5)	61.25 (50.3–71.2)
RT-DETRv2 (R50vd)	0.65	40	0	0	40	100.0 (91.2–100.0)	0.0 (0.0–8.8)	50.00 (39.3–60.7)
Faster R-CNN V2	0.90	33	7	36	4	82.5 (68.1–91.3)	90.0 (76.9–96.0)	86.25 (77.0–92.1)

Abbreviations: TP, true positive; FN, false negative; TN, true negative; FP, false positive; CI, confidence interval.

Values are presented as numbers or percentages with 95% confidence intervals (95% CIs). The independent video test dataset comprised 40 malignant and 40 negative for malignancy (NFM) cases, with one video per patient. Model-specific F1-optimized confidence thresholds were selected using the validation image dataset and fixed before video evaluation. A case was classified as AI-positive if at least one qualified AI alert event occurred during the video, defined as detection of the same candidate object in at least three consecutive frames within a microscope stage-stop interval, using an intersection over union (IoU) threshold of 0.50 for bounding-box tracking. Otherwise, the case was classified as AI-negative. True positive (TP) and false negative (FN) were defined among malignant cases, whereas true negative (TN) and false positive (FP) were defined among NFM cases. Sensitivity was calculated as TP/(TP + FN), specificity as TN/(TN + FP), and accuracy as (TP + TN)/80. The 95% CIs were calculated using the Wilson score method. No frame-level ground-truth bounding boxes were assigned to the videos; therefore, performance measures reflect case-level classification rather than individual cell-level object detection accuracy.

Table 3. False-positive alert frequency and processing speed of four deep learning models in cytology video analysis

model	Confidence threshold	False-alert events/min, median (IQR)	False-alert events/min, mean ± SD	Effective processing FPS, median (IQR)	Effective processing FPS, mean ± SD
YOLO26x	0.45	3.0 (0.0–8.0)	5.72 ± 9.13	35.57 (34.03–36.20)	35.04 ± 1.63
RF-DETR-L	0.45	3.0 (1.0–11.3)	6.90 ± 8.78	22.82 (22.26–23.03)	22.58 ± 0.67
RT-DETRv2 (R50vd)	0.65	12.5 (6.0–24.0)	16.77 ± 13.59	19.55 (19.35–19.84)	19.59 ± 0.39
Faster R-CNN V2	0.90	0.0 (0.0–0.0)	1.00 ± 3.78	19.85 (19.69–20.10)	19.99 ± 0.52

Abbreviations: FPS, frames per second; IQR, interquartile range; IoU, intersection over union; SD, standard deviation.

False-positive alert frequency was evaluated using 40 negative-for-malignancy (NFM) videos, whereas effective processing speed was evaluated using all 80 independent test videos. Model-specific F1-optimized confidence thresholds were determined using the validation image dataset before testing. A qualified AI alert event was defined as detection of the same candidate object in at least three consecutive frames within a microscope stage-stop interval, using an intersection over union (IoU) threshold of 0.50 for bounding-box tracking. Because all videos were 60 seconds long, the number of qualified alert events per video was equivalent to the number of events per minute. False-positive alert events represent alerts generated in NFM videos, not individual false-positive bounding boxes. Values are presented as median (interquartile range [IQR]) and mean ± standard deviation (SD). Differences in false-positive alert frequency among the four models were assessed using the Friedman test, followed by two-sided paired Wilcoxon signed-rank tests with Holm adjustment for six pairwise comparisons. Effective processing speed, expressed in frames per second (FPS), included video decoding and model inference. Stage-motion analysis, bounding-box tracking, and annotated-video generation were excluded from the FPS measurement. Processing speeds were compared descriptively.

Table 4. Characteristics of the datasets used for model development and evaluation

	Training	Validation	Test (Images)	Total (static-image datasets)	Test (Videos)
Dataset composition					
Patients / Specimens, n	314	78	168	560	80
Malignant cases, n	274	68	149	491	40
NFM cases, n	40	10	19	69	40
Image characteristics					
Total images, n	19,657	5201	5136	29,994	–
Images with annotated malignant-cell BBoxes, n	12,173	3254	2772	18,199	–
Images without annotated malignant-cell BBoxes, n	7484	1947	2364	11,795	–
Annotated malignant-cell bounding boxes, n	30,840	7760	7645	46,245	–
Bounding boxes per image, median (Q1–Q3)	1 (0-2)	1 (0-2)	1 (0-1)	1 (0-2)	–
Primary tumor sites - malignant cases, n (%)					
Ovary	62 (22.6%)	15 (22.1%)	38 (25.5%)	115 (23.4%)	6 (15.0%)
Lung	58 (21.2%)	15 (22.1%)	33 (22.1%)	106 (21.6%)	8 (20.0%)
Pancreas	49 (17.9%)	12 (17.6%)	31 (20.8%)	92 (18.7%)	5 (12.5%)
Stomach	38 (13.9%)	9 (13.2%)	23 (15.4%)	70 (14.3%)	8 (20.0%)
Uterine body	28 (10.2%)	7 (10.3%)	11 (7.4%)	46 (9.4%)	3 (7.5%)
Gall bladder / Biliary tract	19 (6.9%)	5 (7.4%)	7 (4.7%)	31 (6.3%)	5 (12.5%)
Breast	12 (4.4%)	3 (4.4%)	3 (2.0%)	18 (3.7%)	4 (10.0%)
Colorectum	8 (2.9%)	2 (2.9%)	3 (2.0%)	13 (2.6%)	1 (2.5%)
Specimen types, n (%)					
Pleural fluid	71 (22.6%)	20 (25.6%)	42 (25.0%)	133 (23.8%)	23 (28.75%)
Intraoperative pleural fluid/washing specimens	25 (8.0%)	5 (6.4%)	14 (8.3%)	44 (7.9%)	4 (5.00%)
Ascitic fluid	76 (24.2%)	19 (24.4%)	30 (17.9%)	125 (22.3%)	11 (13.75%)
Intraoperative peritoneal fluid/washing specimens	132 (42.0%)	32 (41.0%)	78 (46.4%)	242 (43.2%)	38 (47.50%)
Pericardial fluid	10 (3.2%)	2 (2.6%)	4 (2.4%)	16 (2.9%)	4 (5.00%)

Abbreviations: NFM, negative for malignancy; BBox, bounding box; Q1, first quartile; Q3, third quartile.

Data are presented as counts, percentages, or medians (Q1–Q3). Percentages for primary tumor sites were calculated using the number of malignant cases in each dataset as the denominator, whereas percentages for specimen types were calculated using the total number of patients in each dataset. Only one specimen per patient was included, with no patient overlap among the training, validation, independent image test, and independent video test datasets. The number of bounding boxes (BBoxes) represents the number of annotated malignant cells or cell clusters. For the calculation of medians and quartiles, images without malignant cell annotations were assigned a BBox count of zero. One test video per patient was included, and no ground truth BBoxes were annotated for individual video frames.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesandFigurever3.0.docx
- SupplementaryVideoS1ver3.mp4