scientific reports

Check for updates

OPEN

Enhancing cervical cancer cytology screening via artificial intelligence innovation

Yuki Kurita[1]✉, Shiori Meguro[1]✉, Isao Kosugi[1], Yasunori Enomoto[1], Hideya Kawasaki[2], Tomoaki Kano[3], Takeji Saitoh[4], Kazuya Shinmura[5] & Toshihide Iwashita[1]

A double-check process helps prevent errors and ensures quality control. However, it may lead to decreased personal accountability, reduced effort, and declining quality checks. Introducing an artificial intelligence (AI)-based system in such scenarios could effectively address the risk of oversights. This study introduces an innovative AI-integrated workflow for cervical cytology screening that substantially improves efficiency and reduces the burden on cytologists. The AI model prioritizes cases for review based on anomaly scores and streamlines the first screening process to approximately 10 s per case. The model enhances the identification of high-risk cases via detailed microscopic observation, high anomaly scores cases, and a targeted review of low-score cases. The workflow highlights its capability for rapid, accurate, and less labor-intensive evaluations, demonstrating the potential to transform cervical cancer screening. This study highlights the importance of AI in modern medical diagnostics, particularly in areas with a high demand for accuracy and efficiency.

Cervical cancer is the fourth most common cancer among women worldwide, with approximately around 660,000 new cases and around 350,000 deaths reported in 2022, highlighting its substantial impact on global health[1]. Early detection of cervical cancer significantly affects survival rates, and screening based on the Bethesda System[2] is essential for early detection. This method is cost-effective and non-invasive. The use of a microscope plays a crucial role in identifying abnormal cells, thereby serving as a vital strategy to reduce the incidence and mortality rates of cervical cancer.

Improving the accuracy and efficiency of cytological examinations is crucial in medicine, especially because early detection and diagnostic accuracy are directly linked to successful cancer treatment. The introduction of liquid-based cytology (LBC) technology has improved the detection accuracy of abnormal cells, reducing the rate of inadequate samples[3–5]. However, cytological examination relies heavily on meticulous manual observations with the human eye by cytologists and cytopathologists. This process often involves double-checking by multiple people using a microscope, causing physical and mental strain. Such strain may result in delays in obtaining results and potential oversights, jeopardizing patient treatment opportunities[6,7]. A new approach is needed to improve the efficiency and diagnostic accuracy of cytological examinations simultaneously. In particular, reducing the burden on cytotechnologists and cytopathologists while expediting diagnosis is crucial for maximizing patient treatment opportunities.

Against this backdrop, the development of artificial intelligence (AI)-based diagnostic systems has progressed. AI technology in cytology has the potential for quantitative, objective, and reproducible examinations. However, systematic reviews indicate that although AI-based cytological research has shown promising results, much of it remains at an experimental stage, with limited implementation in clinical settings[8–15]. Specifically, a substantial portion of the research has not been validated with realistic clinical data, constraining its applicability in actual clinical scenarios. Moreover, the focus of most developments has been predominantly on diagnosis, neglecting the comprehensive needs and specific workflows of clinical environments. Additionally, research has often disregarded evaluation time, a pivotal factor for clinical deployment.

[1]Department of Regenerative and Infectious Pathology, Hamamatsu University School of Medicine, Hamamatsu, Shizuoka, Japan. [2]Institute for NanoSuit Research, Preeminent Medical Photonics Education and Research Center, Hamamatsu University School of Medicine, Hamamatsu, Japan. [3]Department of Obstetrics and Gynecology, JA Shizuoka Kohseiren Enshu Hospital, Hamamatsu, Shizuoka, Japan. [4]Next Generation Creative Education Center for Medicine, Engineering, and Informatics, Hamamatsu University School of Medicine, Hamamatsu, Shizuoka, Japan. [5]Department of Tumor Pathology, Hamamatsu University School of Medicine, Hamamatsu, Shizuoka, Japan. ✉email: kuri358@hama-med.ac.jp; megu.s@hama-med.ac.jp

Furthermore, convolutional neural network (CNN)-based models developed thus far rely on large volumes of labeled image data and require improvements to accommodate the diversity and complexity of actual clinical data. Our focus has been on developing models for screening rather than diagnosis and rapidly assessing low-magnification images. However, these models have not yet been used practically[16]. Therefore, in this study, we adopted the latest visual language models, using more advanced algorithms than traditional CNNs, and evaluated them using more realistic data.

The new workflow proposed in this study incorporates AI. It aims to improve the traditional screening process, reduce the burden on cytologists and cytopathologists, and enhance the quality of cytological examinations. This approach is expected to substantially contribute to the early detection and treatment of cervical cancer by increasing the efficiency and quality of the cytological examination process. Integrating AI technology is anticipated to simultaneously increase the speed and accuracy of screening and reduce delays and oversights in test results, thereby maximizing patient treatment opportunities. Unlike typical AI-based direct diagnostic studies that only deal with images of already detected abnormalities, our approach targets all slides, including those with undetected anomalies, thereby enhancing the pre-diagnosis screening process. AI is believed to bring objectivity and reproducibility to cytological examinations, benefiting healthcare providers and patients.

Results

Time required for generating and evaluating whole-slide image and tile image

Converting a single specimen into a whole-slide image (WSI) required 210 s, whereas generating tile images required approximately 60 s. The tile images were evaluated using a single RTX A6000 GPU. Of the 938 cases in the test dataset, 896 (389,566 tile images; an average of 500 images per case) were analyzed. This excluded cases in which scanning was impossible (eight cases) and cases without generated tile images (34 cases). The total determination time was approximately 160 min, with an average of 10.7 s per case.

Sorting results of test cases

Among the 938 cases in the test dataset, 162 (17.3%) were deemed inadequate. The inadequate cases included those with fewer than 50 tile images (120 cases), cases where scanning was not possible (8 cases), and cases where no tile images were generated (34 cases). Among the inadequate cases, there was one case each of atypical squamous cells that could not exclude HSIL (ASC-H) and high-grade squamous intraepithelial lesions (HSIL).

Sorting all 938 cases based on anomaly scores and age revealed that among the low-grade squamous intraepithelial lesion (LSIL), squamous cell carcinoma (SCC), and adenocarcinoma (ADC) cases, 40 (76.9%) were sorted into the top 50% and 47 (90.4%) into the top 75%. For atypical squamous cells of undetermined significance (ASC-US) and ASC-H cases, 23 (56.1%) were in the top 50%, and 39 (95.1%) were in the top 75%. Overall, among the abnormal cases, 63 (67.7%) were sorted into the top 50% and 86 (92.5%) into the top 75% (see Supplementary Fig. S1 online).

Seven abnormal cases did not rank in the top 75%. After excluding inadequate specimens (two cases), only five cases remained. These included one case of ASC-US, two of LSIL, one of HSIL, and one of ADC. A common characteristic of these cases was a low number of atypical cells.

Sorting results of each bag

The test dataset was grouped (bags) by submission date, creating 42 bags. Among the 42 grouped bags, eight were only negative for intraepithelial lesion or malignancy (NILM), representing approximately 19% of all bags. The average number of cases per bag was 22.3, with a maximum of 42 cases and a minimum of 10 cases.

There were 52 cases of LSIL, HSIL, SCC, and ADC and 41 cases of ASC-US and ASC-H. Among the top five, 21 cases (40.4%) of LSIL, HSIL, SCC, and ADC were sorted, along with 15 cases (36.6%) of ASC-US and ASC-H. In total, 36 abnormal cases (38.7%) were sorted into the top five. When considering the top 50% in each bag, 41 cases (78.8%) of LSIL, HSIL, SCC, and ADC and 28 cases (68.3%) of ASC-US and ASC-H were sorted, resulting in 69 abnormal cases (74.2%) sorted overall. Among the top 75%, 49 cases (94.2%) of LSIL, HSIL, SCC, and ADC, 37 cases (90.2%) of ASC-US and ASC-H, and 86 abnormal cases (92.5%) were sorted in total.

A total of 174 NILM cases were sorted into the top five in each bag, the details of which are presented in Table 1. The analysis revealed that cases with overlapping cells or a high cell count (Fig. 1a) were more frequently sorted higher. Cases with inflammatory cells (Fig. 1b) or bacteria (Fig. 1c) present in large numbers, overlapping

	Number of cases	%
Overlap	120	69.0
High cell number	85	48.8
Inflammatory	53	30.5
Bacterial	43	24.7
Atrophy	43	24.7
Unknown	13	7.5
Glandular cell	7	4.0
Artifact	5	2.9

Table 1. Details of the top five NILM cases that are sorted in each bag.

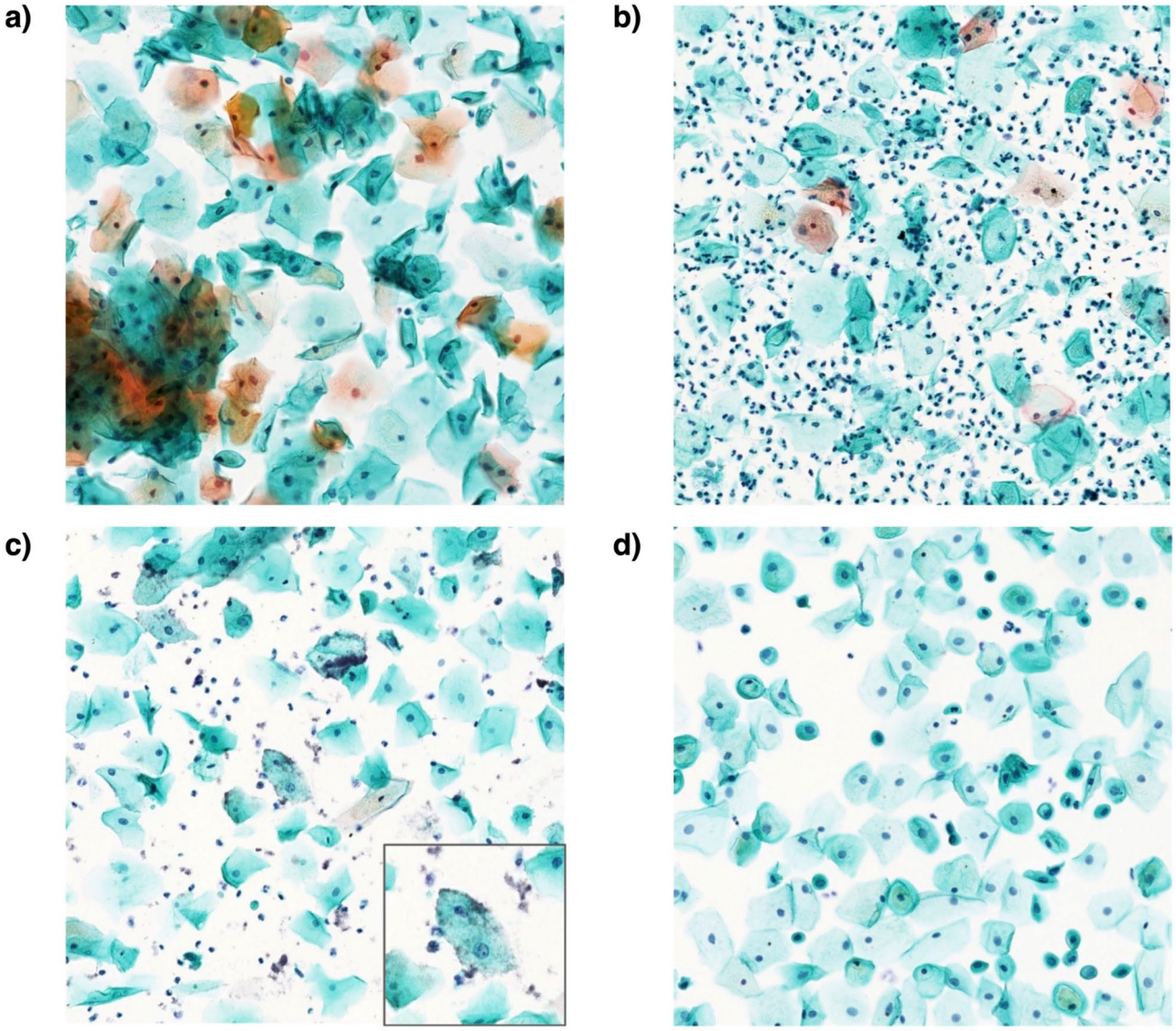

Fig. 1. Images of the top five NILM cases are sorted in each bag. (**a**) Examples with overlapping cells or a high cell count. (**b**) Cases where inflammatory cells are abundant in the specimen and overlapping with epithelial cells. (**c**) Cases where bacteria are abundant in the specimen and overlapping with epithelial cells. (**d**) Examples showing cell atrophy.

with epithelial cells, or showing cell atrophy (Fig. 1d) were also sorted higher. Regarding artifacts, notable cases included those with considerable bubble inclusions or poor encapsulation. In addition, 13 cases were ranked higher for unknown reasons. Some patients also exhibited overlapping factors. In age-based analysis, the sample was categorized into two groups: under 50 and over 50 years old, and the significance of the appearance frequency of each item was verified using the chi-square test. Consequently, bacteria, high cell numbers, and overlap appeared significantly more in the under 50 group (bacteria: $p = 0.0016$, high cell numbers: $p = 0.0384$, overlap: $p = 0.0003$). In contrast, atrophy appeared predominantly in the over-50 age group ($p < 0.001$) (Fig. 2).

Description of representative bags

First, we analyzed two bags with many daily diagnostic cases: Bag10 and Bag25. Bag10 (Fig. 3a) included cases of LSIL, ASC abnormalities, and ADC, whereas Bag25 (Fig. 3b) contained LSIL and ASC-US. When screening in the order of specimen reception, ADC and LSIL in Bag10 were last in the screening order. However, all abnormal cases in Bag10 were placed in the top 50% of cases after sorting. In Bag25, although the sorting order of LSIL marginally decreased, all abnormal cases were sorted into the top 50%.

In Bag1 (Fig. 3c) and Bag4 (Fig. 3d), abnormal cases positioned lower were sorted higher. In Bag1, the lower-positioned LSIL was sorted into the top 1. In Bag4, despite one-third of the cases deemed inadequate, the lower-positioned HSIL and ASC-US were sorted to the top. These results are considered effective for improving screening efficiency and reducing the workload of cytologists.

However, some bags showed no change in sorting results. Bag11 and Bag28 included cases of LSIL, HSIL, and ADC. In Bag11 (Fig. 3e), although LSIL was sorted higher, three out of the four HSIL cases originally in

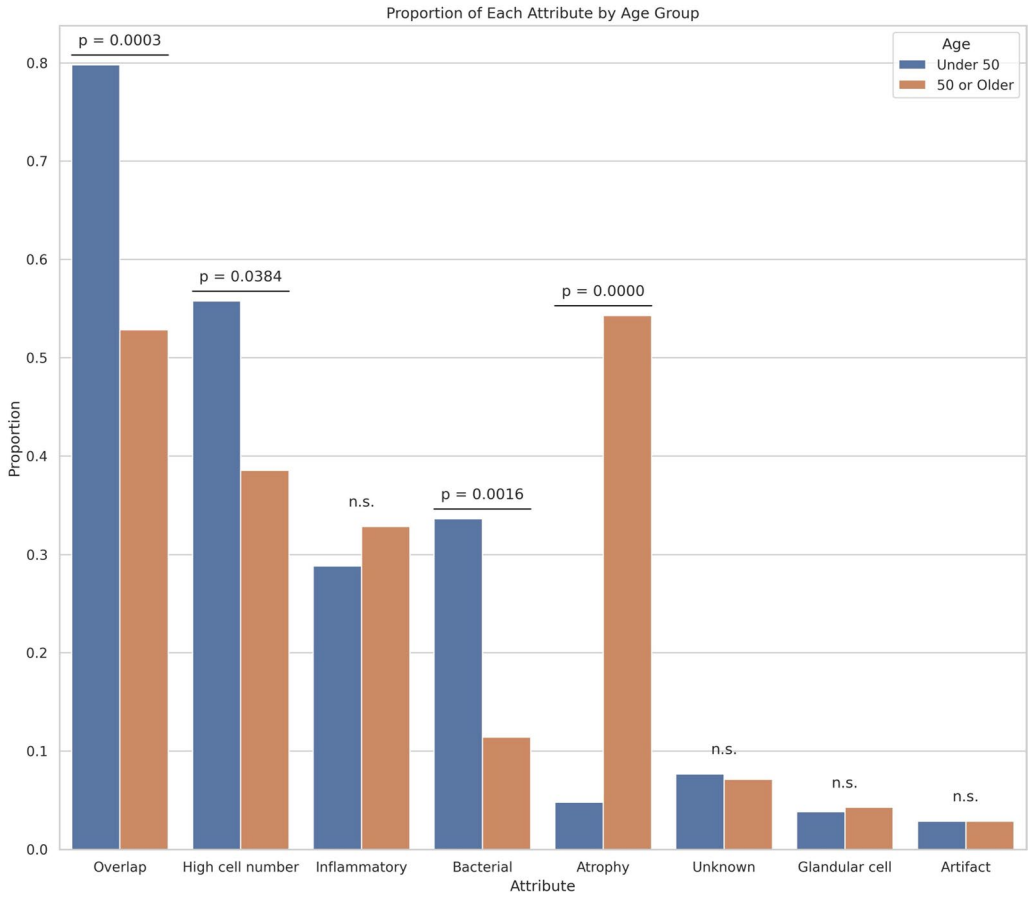

Fig. 2. Analysis of the top five NILM cases sorted within each bag. Cases with overlapping cells or a high cell count were sorted higher. In addition, cases with a large presence of inflammatory cells or bacteria overlapping with epithelial cells and those showing cell atrophy were also sorted higher. Regarding artifacts, notable examples included those with significant bubble inclusion or poor encapsulation. A chi-square test was used to conduct an age-based analysis, categorizing the sample into two groups: under 50 and over 50 years. The results indicated bacterial presence, high cell number, and cell overlap were significantly more common in the under-50 age group. Conversely, in the over-50 age group, a higher prevalence of atrophy was observed. n.s.: Not significant.

lower positions remained lower after sorting. In Bag28 (Fig. 3f), diagnostically important cases, such as HSIL and ADC, were sorted lower. The lower anomaly score in these cases was attributed to the lower number of tile images generated (indicating a lower number of cells in the specimen) and a lower occurrence of atypical cells.

The sorting results for all other bags analyzed are shown in Supplementary Figs. S2–S7 online.

Discussion

In our study, we addressed the limitations of current methods in cervical cytology screening by proposing a new model and workflow that utilizes AI technology. The AI model developed through our research has demonstrated the capability to rapidly and accurately evaluate cytology images, significantly reducing screening time while potentially improving diagnostic quality. Although the double-check process aids in error prevention and quality control, it can also lead to diminished personal accountability (the Ringelmann effect)[17] and reduced effort (the social loafing phenomenon)[18] within the group, consequently risking a decline in check quality. In other words, using the same method for double-checking may repeat initial oversights[19], and in tasks such as cytological examination, which involve considerable physical and mental strain, the risk of missing something is particularly high. Introducing an AI-based system under such circumstances could be an effective measure to reduce potential oversights. In addition, cytology is not always a specialized task in our country and is often performed alongside other laboratory tasks. These additional tasks increase cytologists' burden, leading to delays in reporting cytology results. The current double-check process depends heavily on cytologists' skills and experience.

We previously focused on developing models for screening purposes, particularly for evaluating low-magnification images. The developed models required approximately 30 s per case for determination, highlighting the necessity to improve the determination performance and further reduce processing time[16]. This study addresses these challenges by adopting a visual language model. It quantizes it to develop a model capable of the high-accuracy and high-speed evaluation of many tile images. The time required for screening was approximately

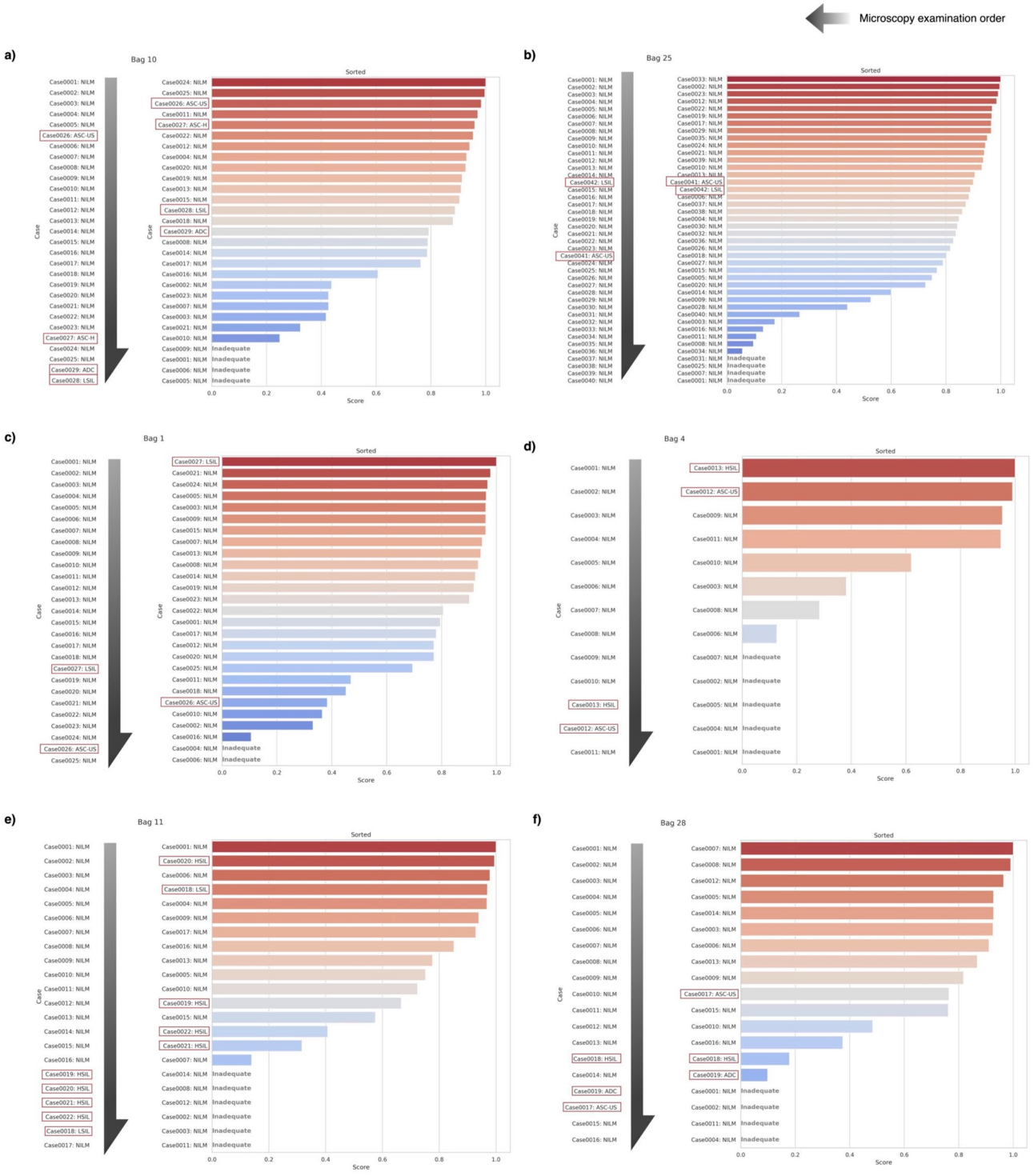

Fig. 3. Description of representative bags. (**a**) A bag containing many daily diagnostic cases, including LSIL, ASC abnormalities, and ADC. When screening in the order of specimen reception, ADC and LSIL were last. However, after sorting, all abnormal cases were sorted into the top 50%. (**b**) A bag with many daily diagnostic cases, including LSIL and ASC abnormalities. Although the ranking of originally higher-positioned LSIL changed minimally, ASC-US was sorted higher. (**c**) A bag where lower-positioned abnormal cases were sorted higher. A lower-positioned LSIL was sorted to the top. (**d**) A bag where lower-positioned abnormal cases were sorted higher. Despite one-third of the cases being deemed inadequate specimens, lower-positioned HSIL and ASC-US were sorted to the top. (**e**) A bag containing LSIL and HSIL, with no change in the sorting results. Although LSIL was sorted higher, three out of four originally lower-positioned HSIL cases remained lower after sorting. (**f**) A bag containing ASC-US, HSIL, and ADC, with no change in the sorting results. Although ASC-US was originally sorted into the top 50%, the lower-positioned HSIL and ADC remained lower.

10 s per case, whereas digitizing a case and generating tile images required approximately 270 s. Furthermore, we established a workflow to sort daily cases using anomaly scores and detect inadequate cases. This model and workflow have the potential to contribute to screening efficiency substantially, reduce the burden on cytologists, and expedite reporting.

In traditional workflows, cytologists must screen each case meticulously without knowing the location of the abnormal cases, leading to considerable physical and mental strain. Our validation results showed that the sorted outcomes within a bag can be categorized into three zones (Fig. 4). The Top5 zone, also known as the "Critical observation zone," requires meticulous scrutiny by cytologists. These include abnormal cases, those with a high cell count and strong overlaps, atrophic changes in older individuals requiring differentiation from HSIL, and cases with a high number of bacteria or severe inflammation in younger individuals. The bottom zone, also known as the "Diagnostic ambiguity zone," consists of potentially inappropriate or ambiguous diagnoses, often inadequate specimens, or cases with a low cell count. However, ASC abnormalities can also be observed in this zone. The most crucial area is the "High-risk sorting zone," between the critical observation zone and diagnostic ambiguity zone, where over 90% of abnormal cases are sorted. Atrophic cases, or those with overlapping cells sorted into the critical observation zone, were considered significant. Atrophic cases may require differentiation from HSIL[20,21], and cases with numerous overlapping cells require careful adjustment of the microscope's focus in the z-direction. Therefore, these cases should be considered important by cytologists. Appropriate feedback on inadequate cases is crucial to prevent overlooking high-grade lesions, reduce the burden on patients for re-examinations, and contribute to improving clinicians' specimen collection techniques[22]. However, feedback on inadequate cases is currently left to the subjective judgment of cytologists. Our study objectively assessed inadequate cases based on cell quantity and the number of generated tile images, which we consider useful for providing appropriate feedback.

The model and workflow developed in this study demonstrated the potential for enhancing the efficiency and accuracy of the cytological screening process. However, some limitations have also become apparent. First, the sorting outcome was significantly influenced by the number of tile images generated and the occurrence of abnormal cells. When a small number of tile images were generated from a single case, the low anomaly score

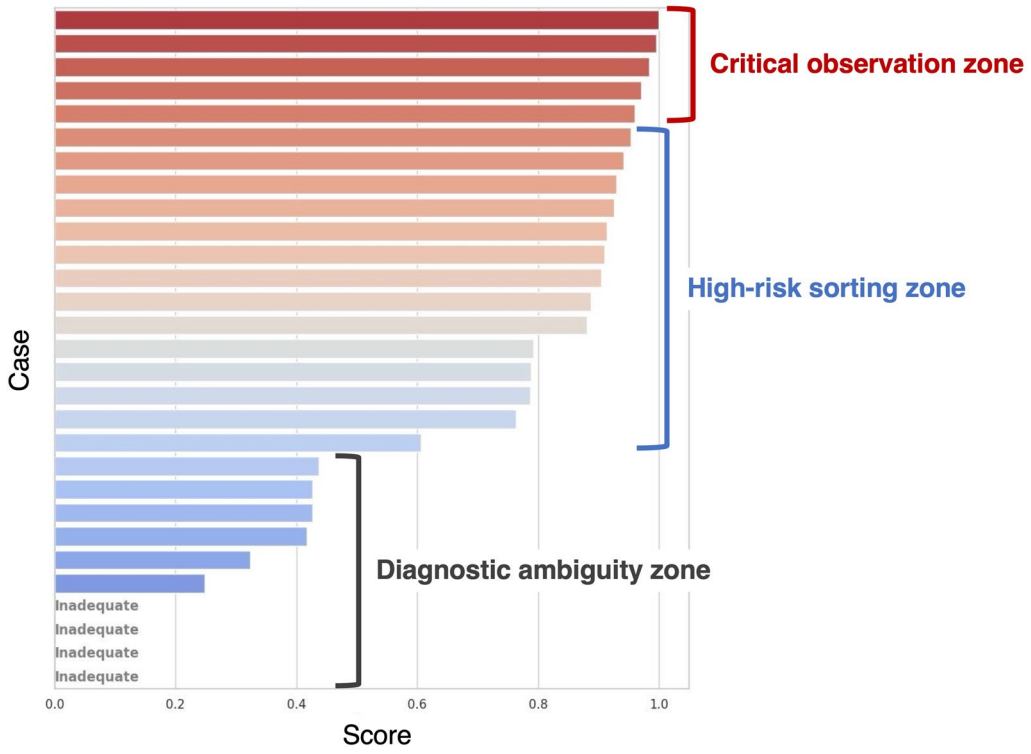

Fig. 4. Sorting results were categorized into three zones. The critical observation zone corresponds to the top five sorted cases. This zone is the "area where cytologists must carefully observe." It includes abnormal cases and cases with a high quantity of cells and strong overlaps, atrophic changes in older adults that require differentiation from HSIL, and cases in younger individuals with a high number of bacteria or strong inflammation. The diagnostic ambiguity zone corresponds to a lower sorted area. This zone is characterized as "potentially inappropriate or ambiguous for diagnosis." This characterization included cases with inadequate specimens or low cell count. However, ASC abnormalities may also be observed in this zone. The high-risk sorting zone lies between critical observation and diagnostic ambiguity zones. It is where "more than 90% of abnormal cases are sorted." This zone is crucial for identifying cases requiring further attention and careful evaluation.

demonstrated that the screening accuracy was significantly affected by the number of cells in the specimen (see Supplementary Fig. S8 online). This observation is important considering the widespread use of LBC in cervical cytology. LBC facilitates efficient cell recovery compared with the direct smear method and enables stable specimen preparation. However, the number of cells in a specimen depends on the specimen collection technique and LBC preparation method. When the cell count in a specimen is low, the specimen is considered inadequate, and re-examination or re-preparation of the specimen is recommended[2].

Conversely, an excessively high cell quantity may result in challenges during WSI creation, such as poor focus or difficulty in observation because of overlapping cells, rendering the specimen unsuitable for diagnosis. Another challenge is the low diagnostic accuracy for ASC cases. ASC determination often occurs when diagnostic findings are lacking, particularly in cases of insufficient nuclear atypia or a low frequency of occurrences[2]. This influence may be attributed to only including definitively diagnosed abnormally labeled tile images in the training data. In particular, cases diagnosed as ASC-US, known to include normal cases, pose a challenge in differentiating them from normally labeled tile images. However, as a subjective judgment by cytologists plays a crucial role in ASC determination, further discussion is needed regarding its addition to future datasets.

Furthermore, in this workflow, we opted not to scale the anomaly scores based on the number of images to avoid losing important information regarding the number of cells in the specimen. Instead, our workflow detects inadequacies using the number of generated tile images as an objective indicator. Cytology screening aims to evaluate the entire specimen, and the number of cells in the specimen is crucial for clinical feedback. However, the number of cells distributed on a specimen varies, and its assessment has traditionally been subjective[22]. In particular, feedback on inadequate specimens requires careful consideration; however, there are variations among facilities. Some diagnoses are made without feedback, even when inadequacy is suspected. The primary cause of inadequacy, especially in the prevalence of LBC, is the method of specimen collection. Cases with a low cell count in the specimen were designed to be sorted lower, despite all tile images being deemed abnormal, to ensure accurate clinical feedback. The anomaly score may be low for cases with numerous tile images if abnormal cells are exceptionally rare. In this study, we identified five cases with few abnormal cells. Variations in specimen collection site, instrument, and clinician's technique level led to substantial differences between facilities. Training for proper specimen collection from the appropriate sites will likely improve this issue.

Our approach emulates the actual workflow of cytological examinations and utilizes AI to assess cytology slides, aiming to reduce the workload of cytotechnologists. As depicted in Fig. 5, the cytology process involves multiple complex stages, particularly the time-consuming and labor-intensive microscopy observations. Unlike typical AI-based diagnostic studies that only handle images with already detected abnormalities, our method targets all slides, including those with undetected anomalies, thereby enhancing the pre-diagnosis screening process. The sorting algorithm used in this study was not integrated into the AI model algorithm and had a simple structure, allowing customization to suit individual facilities. For example, sorting can be based on conditions such as HPV test results or the presence of follow-up information. As these data vary by facility, this study used only the anomaly score and age, which are generally available for sorting. Sorting using various information enables customization tailored to each facility's characteristics. For example, sorting solely based on anomaly scores and age may be effective for screening facilities with numerous new patients. In contrast, larger facilities focusing on referrals may require sorting that considers follow-up information, medical history, and surgical

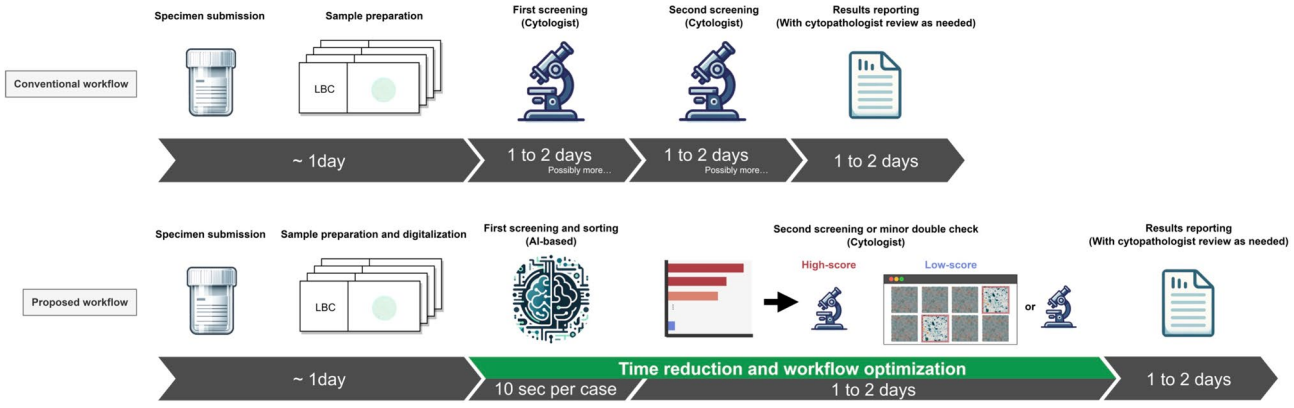

Fig. 5. Overview diagram of conventional workflow and proposed workflow. In traditional workflows, the process from specimen submission to result typically takes 4–7 d. This process is considered intense for cytologists and cytopathologists, leading to considerable mental and physical strain. The burden associated with the first and second screenings is a critical issue. The proposed workflow introduces AI into the first screening, substantially streamlining the process. Using AI, the first screening can be completed in approximately 10 s per case, efficiently passing the sorted group of cases to the second screener. As an operational example in this optimized process, cases with high anomaly scores undergo detailed microscopic observation, whereas those with low scores primarily review only the identified abnormal tiles. Even for cases with low Anomaly scores, microscopic confirmation can be considered if abnormalities are observed in the tile images. This AI-integrated approach reduces the workload and enhances the precision and speed of the screening process. It allows for a more focused review where it is most needed, potentially improving the accuracy of diagnoses and the well-being of the medical professionals involved.

history. Implementing this new workflow for the first screening step can streamline the screening process and reduce the burden on cytologists (Fig. 5). However, this study was limited to validation at a single facility; therefore, multi-facility validation is essential in the future.

Additionally, data containing actual diagnostic statements, not just text data, are needed to improve the model performance further. Diagnostic statements include subtle nuances perceived by cytologists, which are important features of model training. Developing large-scale visual language models with a comprehensive global open dataset, considering text data, image data, and the multiscale nature of WSI, is crucial for automating cytology screening and enhancing its accuracy. Furthermore, significantly enhancing the integration of our AI model into medical information systems can be achieved by interfacing with onsite microscopic cameras for real-time assistance and incorporating decision processes during WSI creation. Leveraging on-premise servers or mobile devices can also improve accessibility and convenience in low-resource environments. However, realizing these technological advances requires overcoming several technical challenges. Specifically, further development in model quantization and optimization, strengthening data-centric learning infrastructures, and accelerating inference through hardware acceleration and edge computing technologies are necessary. These advancements are essential for expanding the adoption of AI models in clinical environments and ensuring their practical utility.

In this study, we proposed a new cytology-based screening workflow that incorporates AI to improve the early detection and treatment of cervical cancer. We achieved more efficient and accurate screening via AI technology to address the physical and mental burden on cytologists and cytopathologists, reduce delays in the screening, and overcome oversights in traditional methods. AI-enhanced screening accelerates and enhances the precision of test results, ultimately expanding early detection and treatment possibilities for cervical cancer. Additionally, the objectivity and reproducibility introduced by AI in cytological examinations provide a reliable diagnostic support tool for healthcare providers, ensuring faster and more accurate treatments for patients. This study represents a new paradigm for screening cervical cytology specimens with potential future applications in other cancer types and medical fields. However, further validation and improvements are needed for clinical implementation, and addressing these challenges in future studies is crucial. Our approach focuses on outputting the degree of abnormality and sorting results of specimens, rendering traditional definitions of false negatives and false positives inapplicable, which complicates statistical comparisons. This is because our system is not designed to provide diagnoses but to assist cytotechnologists in the screening process. Therefore, direct comparisons with similar studies are challenging. The contributions of this study lie in its unique approach and practicality. To further illustrate the application and effectiveness of our workflow, we have included additional validation results in the supplemental materials (see Supplementary Fig. S9 online). These figures demonstrate how the expanded dataset contributes to refining our workflow, offering insights into potential optimizations and adjustments that could enhance the utility and accuracy of the model in clinical settings. Future research should explore further optimization of this new workflow and the potential benefits of integrating it with other diagnostic tools.

Methods

Description of the dataset

From October 2020 to August 2023, cervical specimens were collected from patients at the JA Shizuoka Koseiren Enshu Hospital (400 beds, annual cytology cases: 6766, cervical cytology cases: 3491). Each specimen was subjected to LBC using BD SurePath (Becton Dickinson, Inc., Franklin Lakes, NJ, USA) and standard Papanicolaou staining. Two cytopathologists with over 20 and 10 years of experience and three cytologists, each with over 10 years of experience, diagnosed all cases according to the Bethesda System. This study was approved by the Ethics Review Committees of Hamamatsu University School of Medicine and JA Shizuoka Koseiren Enshu Hospital (Approval No. 21-131). All methods were carried out in accordance with relevant institutional guidelines and regulations. The study was specifically designed and conducted in accordance with the Declaration of Helsinki. The explanation to participants was made using an opt-out process, which the ethics above review committees approved. Informed consent was waived or declared not required by the Ethics Review Committees of Hamamatsu University School of Medicine and JA Shizuoka Koseiren Enshu Hospital as our study did not directly involve obtaining informed consent from participants, as it utilized anonymized tissue samples that were previously collected as part of routine clinical care. These samples were provided by the Hamamatsu University School of Medicine and JA Shizuoka Koseiren Enshu Hospital, with all patient identifiers removed to ensure anonymity and privacy.

Tile image data acquisition

LBC specimens were scanned at 40× magnification using a whole-slide scanner (NanoZoomer 2.0-HT; Hamamatsu Photonics, Hamamatsu, Japan) and converted to WSIs. They were categorized into small patches called tile images, each measuring 1024 × 1024 pixels (0.92 microns/pixel), equivalent to a 10 × objective lens on an optical microscope. The cell quantity in each tile image was calculated based on the number of pixels, excluding the background. Images with a 30% or more cell quantity were filtered and retained. The tile images were generated using a custom algorithm (available at https://github.com/kuri54/Preprocessing-WSI), implemented using only a CPU. The CPU used was Ryzen Threadripper™ PRO 5965WX (AMD, Santa Clara, CA, USA).

Training datasets

From the dataset, we specifically selected cervical specimens from patients who had not undergone a hysterectomy or cervical conization between October 7, 2020, and May 17, 2023. In total, 215 patients were randomly selected. Only the first specimen was used in patients with multiple samples collected during the study period. The breakdown was as follows: NILM, 150 cases; Low-grade LSIL, 10 cases; HSIL, 10 cases; SCC, four cases; and

ADC, four cases (Fig. 6a). Additionally, the recent introduction of LBC at the facility has limited the availability of a broader historical range of cases. These factors have inevitably influenced the diversity and volume of the data collected, contributing to the limitations of our dataset.

All cases were categorized into tiles. For NILM, one image without cell overlap and one with cell overlap among these tile images were used for a similar image search using image hash (available at https://github.com/JohannesBuchner/imagehash). Color hash was used for the hash value search, and only those with a hash value of three or more were sampled. From the sampled image pool, 500 images were randomly selected for each, and all these images were labeled as "normal" (Fig. 6b).

The non-NILM cases were also categorized into tiles and hand-labeled. When atypical cells appeared in tile images that could be determined, they were labeled as "abnormal," and 200 LSIL, 200 HSIL, 100 SCC, and 100 ADC images, totaling 600 images, were sampled (Fig. 6c). A training dataset of 1600 images was created by combining all "normal" (1000) and "abnormal" (600) images. The average age of the cases included in this dataset is 47.7 years (Max: 89, Min: 21).

Test datasets

From the dataset, 938 cervical specimens submitted for cervical cancer screening between May 18, 2023, and July 14, 2023, were used. These specimens were collected consecutively and may have included multiple samples from the same patient. Details are presented in Table 2. All cases were categorized into tiles, and those not labeled were used as the test dataset (Fig. 6d). Among these, 42 cases could not be scanned because of poor encapsulation or a low cell count in the specimen.

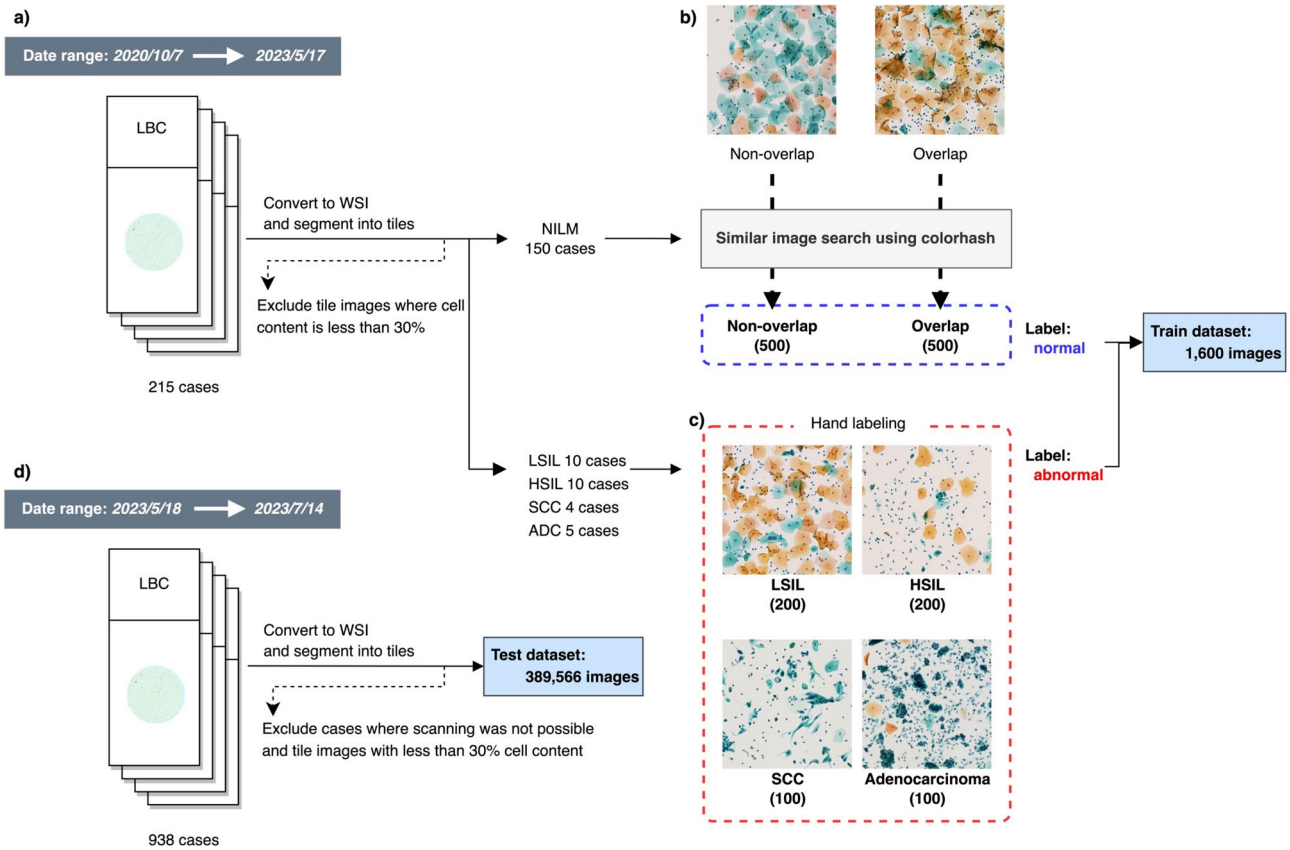

Fig. 6. Overview of the dataset creation process. (**a**) Flowchart for acquiring the training dataset. After converting the LBC specimens to WSIs, they were categorized into tiles, and images with a cell quantity of 30% or more were filtered and selected. (**b**) The process for acquiring data with a "normal" label using a similar image search. One image without cell overlap and one with overlap were used, and a similar image search was conducted using a color hash. From the resulting image pool, 500 images were each randomly selected. They were labeled as "normal." (**c**) The process for acquiring data with an "abnormal" label. Hand labeling was performed for each tile image. If atypical cells were identifiable in the tile image, the label "abnormal" was assigned. (**d**) Flowchart for acquiring the test dataset. After converting LBC specimens to WSIs, cases that could not be scanned were excluded, and the remaining specimens were categorized into tiles. Images with a 30% or more cell quantity were filtered and selected.

Average age	50.9	
Min	19	
Max	92	
Bethesda classification	**Number of cases**	**%**
NILM	845	90.1
LSIL	22	2.3
HSIL	27	2.9
ASC-US	32	3.4
ASC-H	9	1.0
SCC	1	0.1
ADC	2	0.2
	93	9.9

Table 2. Details of the test dataset.

Model training and tuning

All experiments were conducted using Python version 3.8.10. The library versions used in the experiments were as follows: torch v2.0.1, CUDA v11.7.1, CUDNN v8.5.0, torchvision v0.15.2, pillow v9.5.0, scikit-learn v1.3.0, scikit-image v0.21.0, pandas v2.0.3, numpy v1.24.4, seaborn v0.12.2, accelerate v0.23.0, transformers v4.33.1, albumentations v1.3.1, and openslide-python v1.1.1.

Project page (https://github.com/kuri54/GynAIe).

Model training

Our model was constructed by fine-tuning the architecture described by Radford et al.[23], composed of an image encoder, vision transformer (ViT-L/14@336px; available at https://huggingface.co/openai/clip-vit-large-patch14-336, with an input size of 336 × 336 pixels), and text encoder based on a text transformer with a maximum sequence length of 77 tokens. The images were resized to 336 × 336 pixels before inputting into the image encoder. The image and text encoders output 768-dimensional vectorized features and were optimized by minimizing the contrastive loss within a batch. Contrastive learning imparts the model of the correlation between images and text by calculating the cosine similarity between image and text features within a batch (Fig. 7a).

Input prompts were prepared using templates such as ['A photo of {label}.', 'An image of {label}.', 'A picture of {label}.', 'This is a photo of {label}.', 'Here is an image of {label}.', 'Take a look at this photo of {label}.', 'Please see the picture of {label}.', 'You can see the image of {label}.'] randomly selected from each input image. The label, 'normal' or 'abnormal,' was filled in the template (Fig. 7b). For example, with the template 'A photo of {}.' and the label 'normal,' the training prompt became 'A photo of normal.'

We explored combinations of batch sizes and learning rates to identify those minimizing loss. The optimal batch size was 16, and the best learning rate was 1e-8. We set the number of epochs to 400 and conducted mixed-precision training (FP16) using two RTX A6000 GPUs (NVIDIA, Santa Clara, CA, USA) with 48 GB of memory. The trained model was saved at the epoch with the lowest validation loss (399th epoch).

Test case evaluation

The test dataset was grouped (bags) by submission date, creating 42 bags. This grouping strategy reflects the daily variability in specimen submissions, with each bag corresponding to all the cervical cytology specimens received on a particular day. As a result, the number of cases per bag varies, replicating the natural fluctuations in specimen volume typically seen in clinical settings. The cervical cytology specimens included in the test set cover the past two months, ensuring each bag contains a comprehensive snapshot of daily case numbers. This approach simulates the variability observed in actual clinical settings, allowing us to assess the model's real-world applicability and robustness across different volumes of cases. All tile images from the test dataset were evaluated using a quantized (8-bit) trained Contrastive Language-Image Pre-Training (CLIP) model. Cases with fewer than 50 tile images and those that could not be scanned were considered inadequate specimens. A total of 120 cases had fewer than 50 tile images, and when combined with cases that could not be scanned, 162 cases were deemed inadequate.

For evaluation, the prompts "a image of normal" and "a image of abnormal" were used, and the CLIP model inferred which of these the input images had a higher representational similarity (Fig. 7c). For each case, the number of tile images determined as "a image of abnormal" was calculated, and this value was normalized to a range of 0–1 within each bag, defining the "anomaly score." Cases within a bag were sorted in (1) descending order of anomaly score and (2) ascending order of age (Fig. 7d). That is, cases with higher anomaly scores were sorted higher within the bag, and among cases with similar anomaly scores, those of younger ages were sorted higher.

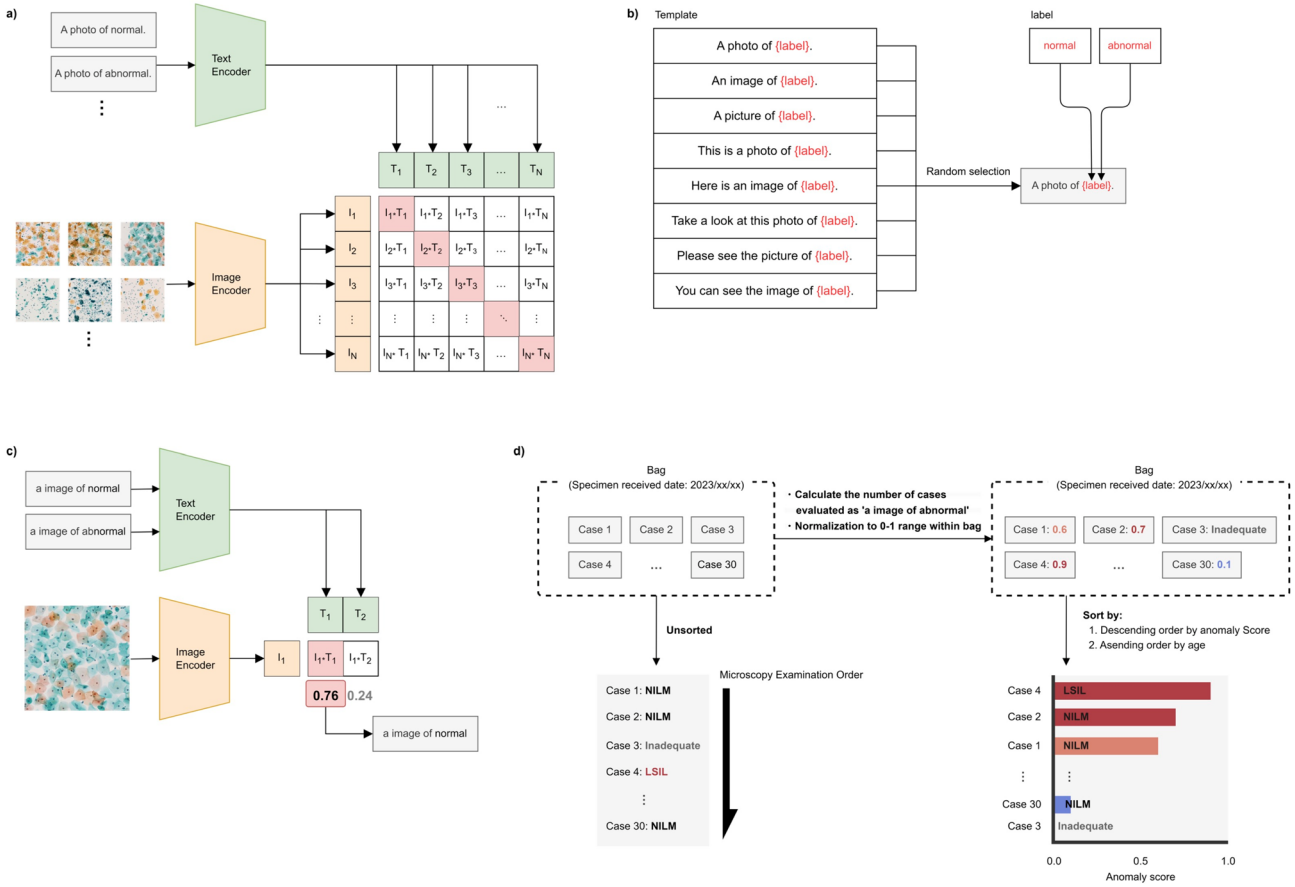

Fig. 7. Overview of the training and evaluation process. (**a**) Training process of the model using contrastive learning. A paired image-text dataset was used, with text inputs for the text encoder derived from the text shown in (**b**). The image encoder received images categorized into tiles. (b) Creation of texts for input into the text encoder. Texts were randomly selected from templates, with {label} filled in with the image label. (**c**) Illustration of the classification process. The classification output was determined by selecting candidate text showing the highest cosine similarity with the input image. (**d**) Overview of evaluation and sorting of test cases. The test dataset was grouped (bag) by submission date, and each case's number of tile images determined as "abnormal" was calculated. These numbers were normalized within the range of 0–1 within each bag, defining the "anomaly score." Cases within a bag were sorted in (1) descending order of anomaly score and (2) ascending order of age. Cases with fewer than 50 tile images and those that could not be scanned were considered inadequate specimens.

Data availability

Ethical restrictions exist on the public sharing of data. There was no provision in the opt-out phase of this study to share data publicly. Therefore, the Ethics Committee of Hamamatsu University School of Medicine restricted these data. If you wish to obtain the datasets, permission must be obtained from the Research Ethics Review Committee of your institution and the Ethics Committee of the Hamamatsu University School of Medicine. For data access requests, please contact the corresponding author, Yuki Kurita.

Received: 7 February 2024; Accepted: 20 August 2024
Published online: 22 August 2024

References

1. Cervical Cancer Statistics (2022) World Health Organization. *Cervical Cancer* https://www.who.int/news-room/fact-sheets/detail/cervical-cancer.
2. Nayar, R. & Wilbur, D. C. The Bethesda system for reporting cervical cytology: A historical perspective. *Acta Cytol.* **61**, 359–372 (2017).
3. Strander, B., Andersson-Ellström, A., Milsom, I., Rådberg, T. & Ryd, W. Liquid-based cytology versus conventional Papanicolaou smear in an organized screening program. *Cancer* **111**, 285–291 (2007).
4. Beerman, H., van Dorst, E. B. L., Kuenen-Boumeester, V. & Hogendoorn, P. C. W. Superior performance of liquid-based versus conventional cytology in a population-based cervical cancer screening program. *Gynecol. Oncol.* **112**, 572–576 (2009).
5. Ito, K. *et al.* A comparison of liquid-based and conventional cytology using data for cervical cancer screening from the Japan Cancer Society. *Jpn. J. Clin. Oncol.* **50**, 138–144 (2020).

6. Yeh, M. W., Demircan, O., Ituarte, P. & Clark, O. H. False-negative fine-needle aspiration cytology results delay treatment and adversely affect outcome in patients with thyroid carcinoma. *Thyroid* **14**, 207–215 (2004).
7. Raab, S. S. *et al.* Double slide viewing as a cytology quality improvement initiative. *Am. J. Clin. Pathol.* **125**, 526–533 (2006).
8. McAlpine, E. D., Pantanowitz, L. & Michelow, P. M. Challenges developing deep learning algorithms in cytology. *Acta Cytol.* **65**, 301–309 (2021).
9. Victória Matias, A. *et al.* What is the state of the art of computer vision-assisted cytology? A systematic literature review. *Comput. Med. Imaging Graph.* **91**, 101934 (2021).
10. Xue, P. *et al.* Deep learning in image-based breast and cervical cancer detection: A systematic review and meta-analysis. *NPJ Digit. Med.* **5**, 19 (2022).
11. Allahqoli, L. *et al.* Diagnosis of cervical cancer and pre-cancerous lesions by artificial intelligence: a systematic review. *Diagnostics* **12**, 2771 (2022).
12. Youneszade, N., Marjani, M. & Pei, C. P. Deep learning in cervical cancer diagnosis: Architecture, opportunities, and open research challenges. *IEEE Access* **11**, 6133–6149 (2023).
13. Jiang, H. *et al.* Deep learning for computational cytology: A survey. *Med. Image Anal.* **84**, 102691 (2023).
14. Sarhangi, H. A., Beigifard, D., Farmani, E. & Bolhasani, H. Deep learning techniques for cervical cancer diagnosis based on pathology and colposcopy images. arXiv [eess.IV] (2023).
15. Jiang, P. *et al.* A systematic review of deep learning-based cervical cytology screening: From cell identification to whole slide image analysis. *Artif. Intell. Rev.* **56**, 2687–2758 (2023).
16. Kurita, Y. *et al.* Accurate deep learning model using semi-supervised learning and noisy student for cervical cancer screening in low magnification images. *PLoS ONE* **18**, e0285996 (2023).
17. Kravitz, D. A. & Martin, B. Ringelmann rediscovered: The original article. *J. Pers. Soc. Psychol.* **50**, 936–941 (1986).
18. Latané, B., Williams, K. & Harkins, S. Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.* **37**, 822–832 (1979).
19. Pfeiffer, Y., Zimmermann, C. & Schwappach, D. L. B. What are we doing when we double check?. *BMJ Qual. Saf.* **29**, 536–540 (2020).
20. Gupta, R., Sodhani, P., Mehrotra, R. & Gupta, S. Cervical high-grade squamous intraepithelial lesion on conventional cytology: Cytological patterns, pitfalls, and diagnostic clues. *Diagn. Cytopathol.* **47**, 1267–1276 (2019).
21. Li, Y., Shoyele, O. & Shidham, V. B. Pattern of cervical biopsy results in cases with cervical cytology interpreted as higher than low grade in the background with atrophic cellular changes. *Cytojournal* **17**, 12 (2020).
22. Yutaka, M. *et al.* Adequacy of cervical cytology by The Bethesda System cell number in specimens in cervical cancer mass screening. *J. Jpn. Soc. Clin. Cytol.* **51**, 110–115 (2012).
23. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the 38th international conference on machine learning* Vol. 139 (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).

Acknowledgements

We thank Yukimi Kouda, Oi Yoshihiro, and Katsuhide Kume for preparing and diagnosing cytology specimens. We thank Mitsue Kawashima, Nao Muranaka, Yuka Homma, and Chisato Nishizawa for their assistance. Part of this study was performed at the Advanced Research Facilities and Services (ARFS) of the Hamamatsu University School of Medicine.

Author contributions

Y.K. was responsible for the conceptualization, data curation, formal analysis, investigation, methodology, resources, validation, visualization, and writing of the original draft. S.M. contributed to conceptualizing and writing the original draft, review, and editing. T.I. managed project administration. I.K., Y.E., H.K., T.K., T.S., and K.S. wrote, reviewed, and edited the manuscript. All the authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70670-6.

Correspondence and requests for materials should be addressed to Y.K. or S.M.