Thyroid

Improved diagnostic accuracy of thyroid fine-needle aspiration cytology with artificial intelligence technology

Journal:	*Thyroid*
Manuscript ID	THY-2023-0384.R6
Manuscript Type:	Clinical or Basic Original Study
Date Submitted by the Author:	28-Mar-2024
Complete List of Authors:	Lee, Yujin; The Catholic University of Korea College of Medicine, Department of Hospital Pathology Alam, Mohammad Rizwan ; The Catholic University of Korea College of Medicine, Department of Hospital Pathology Park, Hongsik ; The Catholic University of Korea College of Medicine, Department of Hospital Pathology Yim, Kwangil; The Catholic University of Korea College of Medicine, Department of Hospital Pathology Seo, Kyung Jin; The Catholic University of Korea Uijeongbu St Mary's Hospital, Deptartment of Hospital Pathology Hwang, Gisu ; AI Team, DeepNoid Inc Kim, Dahyeon ; AI Team, DeepNoid Inc Chung, Yeonsoo ; AI Team, DeepNoid Inc Gong, Gyungyub; Asan Medical Center, Department of Pathology Cho, Nam Hoon ; Yonsei University College of Medicine Yoo, Chong Woo ; National Cancer Center Hospital, Department of Pathology Chong, Yosep; The Catholic University of Korea College of Medicine, Department of Hospital Pathology Choi, Hyun Joo; The Catholic University of Korea College of Medicine, Department of Hospital Pathology
Keyword:	Thyroid Nodules, Pathology-Thyroid Cytology, Clinical Research
Manuscript Keywords (Search Terms):	thyroid, cytology, fine-needle aspiration, artificial intelligence, thyroid neoplasms, deep learning
Abstract:	Background: Artificial intelligence (AI) is increasingly being applied in pathology and cytology, showing promising results. We collected a large dataset of whole slide image of thyroid fine needle aspiration cytology (FNA), incorporating z-stacking, from institutions across the nation to develop an AI model. Methods: We conducted a multi-center retrospective diagnostic accuracy study using thyroid FNA dataset from the Open AI Dataset Project that consists of digitalized images samples collected from three university hospitals and 215 Korean institutions through extensive quality check during the case selection, scanning, labeling, and reviewing process. Multiple z-layer images were captured using three different scanners and image patches were extracted from whole slide images and resized after focus-fusion and color normalization. We pre-tested six AI models,

	determining Inception ResNet v2 as the best model using a subset of dataset, and subsequently tested the final model with total datasets. Additionally, we compared the performance of AI and cytopathologists using randomly selected 1,031 image patches and reevaluated the cytopathologists' performance after reference to AI results. Results: A total of 10,332 image patches from 306 thyroid FNAs, comprising 78 malignant (Papillary thyroid carcinoma) and 228 benign from 86 institutions were used for the AI training. Inception ResNet v2 achieved highest accuracy of 99.7%, 97.7%, and 94.9% for training, validation, and test dataset, respectively (Sensitivity 99.9%, 99.6%, and 100% and specificity 99.6%, 96.4%, and 90.4% for training, validation, and test dataset, respectively). In the comparison between AI and human, AI model showed higher accuracy and specificity than the average expert cytopathologists beyond the two-standard deviation (Accuracy 99.71% (95% CI, 99.38-100.00%) vs. 88.91% (95% CI, 86.99-90.83%), sensitivity 99.81% (95% CI, 99.54-100.00%) vs. 87.26% (95% CI, 85.22-89.30%), and specificity 99.61% (95% CI, 99.23-99.99%) vs. 90.58% (95% CI, 88.80-92.36%). Moreover, after referring to the AI results, all the performance of the experts increased (Accuracy 96%, 95%, and 96%, respectively) as well as diagnostic agreement (from 0.64 to 0.84). Conclusions: These results suggest that the application of AI technology to thyroid FNA cytology may improve the diagnostic accuracy as well as intra- and inter-observer variability among pathologists. Further confirmatory research is needed.

Improved diagnostic accuracy of thyroid fine-needle aspiration cytology with artificial intelligence technology

Running title: AI Improves Accuracy of Thyroid FNA Diagnosis

Yujin Lee[1], Mohammad Rizwan Alam[2], Hongsik Park[1], Kwangil Yim[2], Kyung Jin Seo[2], Gisu Hwang[3], Dahyeon Kim[3], Yeonsoo Chung[3], Gyungyub Gong[4], Nam Hoon Cho[5], Chong Woo Yoo[6], Yosep Chong[2,*], Hyun Joo Choi[1,*]

[1]Department of Hospital Pathology, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, Suwon, Republic of Korea; wondoocha@naver.com (Y.L.); griselbrand@gmail.com (H.S.P.); chj0103@catholic.ac.kr (H.J.C)

[2]Department of Hospital Pathology, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Uijeongbu, Republic of Korea.; rizwan@catholic.ac.kr (M.R.A.); kangse_manse@catholic.ac.kr (K.Y.); ywacko@catholic.ac.kr (K.J.S.); ychong@catholic.ac.kr (Y.C.)

[3]AI Team, DeepNoid Inc., Seoul, Korea; kisu031@gmail.com (G.H.); anniy8920@outlook.kr (D.K.); yeonsoo00.chung@gmail.com (Y.S.C)

[4]Department of Pathology, Asan Medical Center, Seoul, Korea; gygong@amc.seoul.kr (G.G.)

[5]Department of Pathology, Yonsei University College of Medicine, Seoul, Korea; CHO1988@yuhs.ac (N.C.)

[6]Department of Pathology, National Cancer Center, Ilsan, Gyeonggi-do, Republic of Korea; cwy@ncc.re.kr (C.W.Y.)

***Correspondence:**
Hyun Joo Choi, MD., PhD.

Address: Department of Hospital Pathology, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, 93, Jungbu-daero, Paldal-gu, Suwon 16247, Gyeonggi-do, Republic of Korea

Tel: (+82)-031-249-7592

E-mail: chj0103@catholic.ac.kr

Fax: (+82)-031-244-6786

Yosep Chong, MD., PhD

Address: Department of Hospital Pathology, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 271, Cheonbo-ro, Uijeongbu 11765, Gyeonggi-do, Republic of Korea.

Tel: (+82)-032-820-3160

E-Mail: ychong@catholic.ac.kr

Fax: (+82)-032-820-3877

Conflict of Interests: The authors declare that they have no competing interests.

Funding: This work was partially supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (2021R1A2C2013630).

Abstract

Background: Artificial intelligence (AI) is increasingly being applied in pathology and cytology, showing promising results. We collected a large dataset of whole slide image of thyroid fine needle aspiration cytology (FNA), incorporating z-stacking, from institutions across the nation to develop an AI model.

Methods: We conducted a multi-center retrospective diagnostic accuracy study using thyroid FNA dataset from the Open AI Dataset Project that consists of digitalized images samples collected from three university hospitals and 215 Korean institutions through extensive quality check during the case selection, scanning, labeling, and reviewing process. Multiple z-layer images were captured using three different scanners and image patches were extracted from whole slide images and resized after focus-fusion and color normalization. We pre-tested six AI models, determining Inception ResNet v2 as the best model using a subset of dataset, and subsequently tested the final model with total datasets. Additionally, we compared the performance of AI and cytopathologists using randomly selected 1,031 image patches and reevaluated the cytopathologists' performance after reference to AI results.

Results: A total of 10,332 image patches from 306 thyroid FNAs, comprising 78 malignant (Papillary thyroid carcinoma) and 228 benign from 86 institutions were used for the AI training. Inception ResNet v2 achieved highest accuracy of 99.7%, 97.7%, and 94.9% for training, validation, and test dataset, respectively (Sensitivity 99.9%, 99.6%, and 100% and specificity 99.6%, 96.4%, and 90.4% for training, validation, and test dataset, respectively). In the comparison between AI and human, AI model showed higher accuracy and specificity than the average expert cytopathologists beyond the two-standard deviation (Accuracy 99.71% (95% CI, 99.38-100.00%) vs. 88.91% (95% CI, 86.99-90.83%), sensitivity 99.81% (95% CI, 99.54-100.00%) vs. 87.26% (95% CI, 85.22-89.30%), and specificity 99.61% (95% CI, 99.23-99.99%) vs. 90.58% (95% CI, 88.80-92.36%). Moreover, after referring to the AI results, all the performance of the experts increased (Accuracy 96%, 95%, and 96%, respectively) as well as diagnostic agreement (from 0.64 to 0.84).

Conclusions: These results suggest that the application of AI technology to thyroid FNA cytology may improve the diagnostic accuracy as well as intra- and inter-observer variability among pathologists. Further confirmatory research is needed.

Keywords: thyroid; cytology; fine-needle aspiration; artificial intelligence; thyroid neoplasms; deep learning

Introduction

Artificial intelligence (AI) is emerging in cytopathologic image analysis with promising results. Advancement in AI have enabled the application of computer models, such as artificial neural networks, deep learning, genetic algorithms, and classification, to support diagnosis and treatment decisions.[1] Computational models have been developed to perform pathological image analyses using machine learning.[2-5] Image analysis can be used not only for surgical or biopsy slide samples but also for cytologic slides.[6] It can be used to objectify the diagnostic process, mitigate human effort, and improve the productivity of cytopathological examinations.

In non-gynecologic cytology, thyroid fine needle aspiration (FNA) is the most actively explored field of AI application.[7] This is because AI diagnosis is the preferred modality for the evaluation of thyroid nodules owing to easy access and a generally high accuracy. FNA cytology for thyroid nodules is ideal for applying AI owing to the relatively specific diagnostic criteria for thyroid nodules. Thyroid nodules are observed in an estimated 10% of the general population, of which 5–7% suffer from malignant thyroid nodules.[8] FNA is currently considered the preferred modality for the evaluation of thyroid nodules.[9,10] FNA diagnosis is less invasive compared to others; it is also rapid and has high accuracy. Cytopathologists follow The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC)[11,12], which comprises six categories to evaluate the risk of malignancy in thyroid FNA samples. The categories are as follows: non-diagnostic, benign, atypia of undetermined significance/follicular lesion of undetermined significance, follicular neoplasm/suspicious for follicular neoplasm, suspicious for malignancy, and malignant. The sensitivity and specificity of the current system range from 68% to 98% and 56% to 100%, respectively.[8,11,13,14]

Studies on the application of AI in thyroid FNA have promising results; however, they had several limitations, such as a small sample size, lack of z-stacking, and the collection of data only from a single institute. Computational methods have been applied to the cytological analysis of thyroid nodules since 1980.[15,16] In 1999, Karakitsos et al.[17] classified the images of benign and malignant thyroid tumors using a learning vector quantizer artificial neural network and achieved 97.8% accuracy. Study by Cochand-Priollet, B. et al. identified malignant and benign nodules using May-Grunwald-Giemsa-stained smears and four independent classifiers. Subsequently, the performances of the classifiers were compared. The most successful classifier, the k-nearest neighbor, achieved an overall accuracy of 88.71%.[18] Convolutional neural networks have been used to predict malignancy from 908 whole slide images (WSIs), and a sensitivity of 92.0% and specificity of 90.5%[19] have been achieved.

In another study, an artificial neural network was used to classify 447 cases into two categories, benign and malignant, using the radial basis function, and an overall sensitivity of 95.0% and specificity of 95.5% were achieved.[20] Guan, Q. et al. conducted a study using a deep convolutional neural network and fragmented image, which were classified using VGG-16; they obtained an accuracy of 97.66%.[21] Previous studies have evaluated the diagnostic performance of AI, but few studies have explored the AI usefulness with enough amount of training and validation datasets.

The second challenge of developing effective AI for thyroid cytology is the lack of consideration for z-stacking in the training and validation datasets. In cytology, distinct from histologic samples, individual cells and three-dimensional cell clusters are suspended in a medium, floating in the space between the glass and cover slides, typically spanning 50-100 microns. This space is referred to as the z-axis, and it is crucial to scan cytologic slides at various z-layers of focus to capture clearly focused images of these clusters.[22] This is especially important for samples like thyroid FNAs, which naturally include large papillary clusters. Previously, there has been no special focus on z-stacking for digital cytology images in AI research related to cytology. [22,23] However, through our recent comparative analysis of image quality across different scanners, we have identified the minimal requirements for appropriate z-stack layers based on sample types.[24] For instance, at least three layers are necessary for liquid-based preparations (LBP) and five layers for conventional smears. Scanning digital cytology images with these specified z-stack layers, tailored to the sample type, is vital for developing robust AI models that can be effectively applied in daily practice. Moreover, as the third challenge, AI models in past studies have been developed and validated using datasets from only a single or two institutions.[23] This approach lacks the diversity and scale needed for good generalizability. There is a pressing need for multi-institutional or national level datasets to enhance the AI models' applicability and reliability in various clinical settings.

To address the aforementioned limitations, we built a dataset that contained a sufficient number of cases and z-stacks from nationwide institutions to develop an AI model for thyroid FNAs.

Material and Methods

The study was approved by the Institutional Review Board of the Catholic University of Korea, College of Medicine (UC21SNSI0064), the Institutional Review Board of the Yonsei University College of Medicine (4-2021-0569), Institutional Review Board of the National Cancer Center (NCC2021-0145), and Institutional Review Board of the St. Vincent's Hospital College of Medicine, Catholic University of Korea (VC22RISI0131). The informed consents from the patients were waived by these IRBs. The use of quality assurance program slides of the Korean Society of Cytopathology was approved by the society executive board. A schematic overview of the

proposed method and workflow is shown in Fig. 1.

Data collection

This is a multi-center retrospective diagnostic study. Training, validation, and testing were performed using images of thyroid FNA specimens from a dataset provided by an AI research project known as "the Open AI Dataset Project" (Fig. 2). The Open AI dataset consists of digitalized images of cytopathology slides collected from three university hospitals and quality control slides from 215 institutions in South Korea. Because of the data imbalance, we only included papillary carcinoma but no other histologic subtypes for malignant cases. The slides were scanned using AT2 (Leica Biosystems, Germany), NanoZoomer S360 (Hammatsu, Japan), and Pannoramic 250 Flash III (3DHistech, Hungary) at a ×40 objective magnification at the focal planes. Each dataset contained multiple z-stacks, with conventional smears having five layers and LBP slides having one to three layers. The selection of cytologic samples was stringently managed by 12 board-certified cytopathologists and 5 cytologists. Each sample was initially confirmed to match its corresponding histologic slides by the institute's board-certified cytopathologists. Following this, another set of board-certified cytopathologists assessed the slide quality and determined their suitability for inclusion in the Open AI dataset. Subsequently, the quality of the scanned images was meticulously reviewed by another group of board-certified cytopathologists.

A total of 5500 WSIs including all the non-gynecologic samples such as respiratory tract, body fluid, urine, thyroid FNAs, and other FNAs were collected. 1100 WSIs of textbook cases were from quality control slides of 215 all registered cytopathology laboratories in Korea achieved for quality assurance program of the Korean Society for Cytopathology, while another 4400 WSIs of daily practice cases were collected from 12 major hospitals in Korea. Among the Open AI Dataset, we exclusively utilized thyroid FNA samples other than other types of samples for this study. We also excluded the samples that are not eligible for AI training such as bad image quality. Finally, thyroid 306 FNAs samples were collected from 86 institutions and used in this study (Supplementary Table 1). We included LBP samples as well as conventional smear, Papanicolaou stained samples as well as hematoxylin and eosin-stained samples.

From the 78 malignant WSIs, 7,994 malignant image patches and 33,245 benign image patches were created. From the 228 benign WSIs, 202,345 benign image patches were created. Due to data imbalance, all the benign patches from malignant and benign WSIs were merged, resulting in a total of 235,590 benign image patches. To address this imbalance, 227,699 benign image patches were randomly excluded. Finally, for this study, 7,994 malignant and 7,891 benign image patches were used for the development of the AI model.

Image pre-processing

To deal with the multiple images with different z-axis, we first generated the extended z-stack images from 3-6 images with different z-axis by enhanced focus fusion technology (Fig. 3). By this process, we obtained WSIs with only 1 evenly focused layer for 3-dimensional cell clusters. Color normalization was performed on the integrated images before patch extraction. Images were cropped and extracted as 1024 × 1024 pixels-sized image patches and resized to 256 × 256 pixels before using them as inputs for model training (Fig. 4). The image patches were evenly divided into grids from the WSIs. Each patch image was reviewed by trained cytopathologists to confirm the benign or malignant and another group of cytopathologists reviewed the labeling and the image patches with discordant opinions were excluded from the dataset as mentioned in the prior section. Image patches with no follicular cells have been removed, and image patches with both benign and malignant follicular cells have been considered as malignant. The RAND function in Excel was utilized to assign a random number to each image patch for the selection Microsoft Excel (Version 2021, Build. 14827.20192).

Image patch labeling

For the categorization of image patches, those extracted from negative samples were collected as negative image patches. In contrast, patches derived from positive samples were further classified into two categories: those containing cancer cells and those without. This classification was performed by board-certified cytopathologists. Finally, to ensure the utmost accuracy and consistency, these categorized image patches underwent a further review by yet another group of board-certified cytopathologists. If there is a discordant opinion between 1st and 2nd reviewers, the image patches were removed from the dataset. This rigorous process underscores the commitment to the highest standards of accuracy and reliability in this dataset for AI training.

From the 78 malignant WSIs, 7,994 malignant image patches and 33,245 benign image patches were created. From the 228 benign WSIs, 202,345 benign image patches were created. Due to data imbalance, all the benign patches from malignant and benign WSIs were merged, resulting in a total of 235,590 benign image patches. To address this imbalance, 227,699 benign image patches were randomly excluded. Finally, for this study, 7,994 malignant and 7,891 benign image patches were used for the development of the AI model.

Pre-test for AI-model selection

Prior to model selection, we tested six models: Inception ResNet v2, MobileNet v2, Efficientnet-b1, Densenet 121, ResNext50, and ResNet50. Subsequently, we compared their performances. The training, validation, and testing were performed using random WSI patches of 78 malignant and 88 benign cases. A total of 2,351 image

patches were created from the malignant and 2,518 from the benign WSIs. The number of benign and malignant patches used for training, validation, and testing was adjusted similarly, and a total of 3,797 image patches were used for training, 565 for validation, and 507 for testing.

AI model training

For model training, 15,975 augmented patch images were created from the dataset. The images were flipped vertically and horizontally and rotated clockwise for data augmentation. The augmented dataset included 7,981 benign and 7,994 malignant image patches. A total of 14,903 image patches were used for training, 565 for validation, and 507 for testing (Table 1). The performances of the AI models trained using augmented, non-augmented, and reduced datasets were also compared along with a 95% confidence interval (95% CI).

Comparison of the performances between experts and AI model

Of the 15,975 image patches, 1,031 image patches were randomly selected to compare the diagnostic accuracies of experts and the AI model. The RAND function in Excel was utilized to assign a random number to each image patch for the selection Microsoft Excel (Version 2021, Build. 14827.20192). The diagnostic accuracy of the AI model was compared with that of three experienced cytopathologists. In addition, the diagnostic accuracy of the cytopathologists after referring to the AI results was investigated. Standard deviations were also analyzed for the comparison results.

Statistical analysis

To analyze the diagnostic agreement among pathologists, the Fleiss' Kappa coefficient was used. The 95% CI of the Fleiss' Kappa coefficient was also analyzed. The analysis was performed using R statistical programming (Version 3.4.1; http//www.r-project.org, accessed on May 21, 2023).

Results

Data collection

A total of 306 WSIs were included in the dataset, 78 of which were malignant (Papillary thyroid carcinoma) and 228 were benign (25.5% vs. 74.5%), 133 cases were the textbook cases with typical cytologic features from the Quality Assurance Program slides from 85 institutions, while the other 173 cases were daily practice-level cases collected from Uijeongbu St. Mary's Hospital and Yonsei University Severance Hospital (43.5% vs. 56.5%). Of

the slides included in the dataset, 121 were hematoxylin and eosin-stained, and the remaining 185 were Papanicolaou-stained (39.5% vs. 60.5%). The dataset contained 239 LBP slides, which is approximately thrice the number of conventional cytology slides (67, 78.1% vs. 21.9%). 149 cases were scanned by Leica AT2, 83 by Hamamatsu, and 74 by 3DHistech scanner (48.7%, 27.1%, and 24.2%, respectively). Each WSI scan contained at least three to six z-stack layers. Supplementary Table 2 summarizes clinical and pathological information about the dataset included.

Pre-test for AI model selection

Inception ResNet v2 exhibited the highest accuracy of 97.04%. MobileNet v2, Efficientnet-b1, Densenet 121, ResNext50, and ResNet50 achieved an accuracy of 95.07%, 94.67%, 95.07%, 94.28% and 95.27%, respectively. Inception ResNet v2 exhibited a high sensitivity and specificity. MobileNet v2 achieved 100% sensitivity; however, the accuracy was low because the specificity was only 90.77%. By contrast, ResNext50 exhibited the highest specificity but a relatively low sensitivity (Fig. 5).

AI model training

Inception ResNet v2 showed the highest accuracy when the augmented datasets were used for training (Table 2). The model obtained 99.72% (95% CI, 99.64-99.80%) accuracy, 99.87% (95% CI, 99.81-99.93%) sensitivity, and 99.58% (95% CI, 99.48-99.68%) specificity for training dataset, 97.70% (95% CI, 96.46-98.94%) accuracy, 99.57% (95% CI, 99.03-100.00%) sensitivity, and 96.37% (95% CI, 94.83-97.91%) specificity for validation dataset, 94.87% (95% CI, 92.95-96.79%) accuracy, 100% (95% CI, 100.00-100.00%) sensitivity, and 90.41% (95% CI, 87.85-92.97%) specificity for test dataset. Examples correctly diagnosed and misdiagnosed samples by AI (Fig. 6). Result of the non-augmented, and the reduced dataset is shown in Supplementary Table 3.

Comparison of the performances between the AI model and the experts

A total of 1,031 image patches were randomly selected from all datasets for comparison. The image patches included 513 negative cells and 518 malignant cells (Papillary thyroid carcinoma). Three pathologists and the AI model reviewed and labeled the patches. The three pathologists showed an average sensitivity of 87.26% (95% CI, 85.22-89.30%), a specificity of 90.58% (95% CI, 88.80-92.36%), and an accuracy of 88.91% (95% CI, 86.99-90.83%). The Fleiss' Kappa coefficient for diagnostic agreement between pathologists A, B and C was 0.66 (95% CI, 0.63-0.68), indicating moderate agreement. The AI model showed an average sensitivity of 99.81% (95% CI,

99.54-100.00%), a specificity of 99.61% (95% CI, 99.23-99.99%), and an accuracy of 99.71% (95% CI, 99.38-100.00%). Examples of images diagnosed differently by the pathologists and the AI model are shown in Fig. 7. After referring to the diagnostic results obtained by the AI model, the diagnostic accuracy of the three pathologists increased to 95.76% (95% CI, 94.53-96.99%). The sensitivity and specificity increased to 95.24% (95% CI, 93.94-96.54%) and 96.30% (95% CI, 95.15-97.42%), respectively (Fig. 8). The Kappa coefficient for diagnostic agreement between pathologists also increased from 0.66 (95% CI, 0.63-0.68) to 0.86 (95% CI, 0.83-0.88), indicating very good agreement.

Discussion

This study developed an AI model, which exhibited a competitive performance with an accuracy of 99.7%, using 15,975 thyroid image patches obtained from 306 WSIs. The image patches were collected from approximately 200 hospitals nationwide with the help of the Korean Society for Cytopathology and three major institutions participating in the Open AI Dataset Project. The accuracy of the AI model was higher than that of the pathologists (99.71% vs. 88.91%). In addition, the diagnostic accuracy of the pathologists was improved upon referring to the AI results.

The slides included in the dataset were collected nationwide. To include a sufficient number of z-stacks in the image and reduce the data-storage space required by the z-stacks, at least three layers were integrated to create an extended z-stack image. Previous studies have reported that utilizing a z-stack was difficult because it required a large storage space.[19] Many studies either avoided using any z-stack images or used only the middle layer of complete images;[19,25,26] however, multi-focus images must be analyzed to increase the accuracy of diagnosis based on cytology slides. Image integration can be applied to obtain clear images of multiple foci without necessitating large capacities. There are a recent technical advances in the scanning technology for cytology such as volumetric scanning by Hologic (Marlborough, MA), multi-camera image capture and focus fusion technology by Vieworks (Anyang, Republic of Korea), and pixel level z-scanning technology by Pramana, Inc. (Morrisville, North Carolina). [27,28] This new technique will revolutionize the practical application of digital cytology and the use of AI.

Although the revised College of American Pathologists guideline regarding the validation of cytology for diagnostic purposes indicates insufficient evidence supporting the use of digital cytology for primary diagnosis, recently many efforts have been made to strengthen the idea of the role of digital cytology and AI.[29-31] Recent

studies using AI technology on thyroid FNAs have shown very promising results in addition to the gynecologic cytology screening AI software that were recently introduced in the market with CE-mark and FDA-approval by several companies, such as Hologic (Genius Digital Diagnostic System, USA), TechCyte (UT, USA), Datexim (CytoProcessor, France), Cell Solutions (BestCyte Cell Sorter Imaging System, USA), Landing Med (China), and KF Bio (China).[27,32] Some of these AI software adopted more advanced scanning technology that enhanced the scanning speed and image quality such as volumetric scanning by Hologic (Marlborough, MA), multi-camera focus-fusion technology by Vieworks (Anyang, Republic of Korea), and pixel level z-stacking by Pramana, Inc. (Morrisville, North Carolina).[28] These changes show the possibility of fundamental changes in digital cytology and the potential application of AI soon.

Our model showed the best diagnostic performance of 99.71% accuracy in the thyroid FNAs . The highest accuracy reported for the cytological classification of thyroid nodules in the previous studies is 97.66%.[21] A study in 2020, developed an artificial neural network model and obtain an accuracy of 95% using a larger number of image patches and 447 WSIs[20] (Supplementary Table 4). Among the papers reported to date, the study that utilized the largest number of samples, 908 WSIs, achieved a sensitivity of 92% and specificity of 90.5% through VGG-11, a type of Convolutional neural network.[19] In a subsequent study conducted by the same authors, the performances of AI models were compared with that of expert pathologists in predicting malignancy from WSIs; the AI models and experts showed similar accuracies.[26] Interestingly, the performances of the AI models improved and the diagnosis became similar to that of the experts when the models underwent a fully supervised training to identify the region of interest.

In this study, the diagnostic accuracy of the AI model was significantly higher than that of the experts, and the diagnostic accuracy of the experts improved after referring to the AI results. Previous studies have compared human and AI performance and assess diagnostic agreement with screeners[26], this is the first study to evaluate how diagnostic accuracy changes when humans refer to AI results. Of the 1,031 image patches used for comparison, only one image patch of a malignant case was accurately diagnosed by the experts but was misdiagnosed by the AI model (Fig. 7d). As is well known, typical papillary thyroid carcinoma has nuclear features such as enlargement, elongation, crowding, irregular contours, grooves, pseudoinclusions and chromatin clearing.[33] The case exhibited the characteristic features of typical papillary thyroid carcinoma, such as overlapping nuclei and intranuclear pseudoinclusions; however, the image patch showed relatively fine chromatin and smooth cell membrane. The AI model misdiagnosed benign nodules as malignant in two image patches. In

both patches, it showed many lymphocytes and histiocytes, which might have been mistaken for malignant cells (Fig. 7c). In lymphocytic thyroiditis, the nuclei of neighboring follicular cells frequently undergo a slight enlargement and exhibit atypical features in response to chronic inflammation. This phenomenon could account for the AI model's erroneous positive predictions. By contrast, the AI model accurately distinguished the characteristics of malignant cells with pseudoinclusions or irregular cell membranes, even in the images of regions with low cellularity (Fig. 7b). After referring to the AI's results, the pathologists were able to identify small or fuzzy pseudoinclusions to base their judgment on. In addition, the AI model accurately diagnosed benign nodules containing clumping follicular cells, which the experts misdiagnosed as malignant nodules (Fig. 7a). In the actual benign cases where AI got the diagnosis right and the pathologist got it wrong, the confusing factors that influenced the pathologist's judgment were features like focal overlapping nuclei, high cell density, and relatively irregular nuclear size. However, the nuclei of these cells retained a fine chromatin pattern and smooth membrane, which lacks evidence of malignancy.

While it is possible to compare images of cases which the AI made a correct or incorrect diagnosis and assume that cells such as histiocytes would have been misdiagnosed as malignant, or a small intranuclear inclusion may helped it diagnose the malignant cell correctly, it is actually difficult to interpret exactly what criteria the AI uses to distinguish between malignant and benign cells. In the case of this study, it's especially hard to know which cases AI is more likely to misdiagnose, since AI got most of the diagnoses right. This makes it difficult to decide which of the AI's judgments to trust and which not to trust. Recently, quantitative scoring methods have been attempted to develop interpretable AI[4], and in the case of typical papillary thyroid carcinoma, we can think of ways to let the AI calculate the nuclear scoring system[33] to make a quantitative assessment or mark areas in the image that are likely to be malignant cells. However, it is important to consider that there is a trade-off between explainability and performance.[34-37]

Our study made three pathologists determine the initial diagnosis of 1,031 image patches, revise their diagnoses based on the AI's diagnosis, and then compare their results. After noting that some of their diagnoses were different from AI readings, the pathologists reviewed and revised at least one of their misdiagnoses. Referring to AI gave pathologists the opportunity to review their diagnosis in more detail if it differs from the AI's, and ultimately contributed to improving agreement level and diagnostic accuracy. Initially, the Fleiss' Kappa coefficient diagnostic agreement between pathologists A, B and C was 0.66, indicating moderate agreement. Pathologist A tended to have high specificity and low sensitivity, while pathologist C had high sensitivity and low specificity

(95.71%, 71.04%, 95.37% and 85.96%, respectively, data not shown). Pathologist B's sensitivity was higher than A's and the same as C's, specificity was higher than C's and lower than A's, and accuracy was the highest. The sensitivity, specificity, and accuracy of all three pathologists increased after referring the AI's diagnosis. Notably, pathologist A's sensitivity increased to 90.73%, and pathologist C's specificity increased to 96.88% after referring the AI's diagnosis (data not shown). The diagnostic agreement between pathologists also increased to 0.86, indicating very good agreement. This means that AI can help calibrate pathologists' diagnoses to increase sensitivity or specificity and reduce inter-observer variation due to individual pathologist tendencies.

However, this study has a few limitations. We classified thyroid cell slides into only two categories, malignant and benign, without segmenting the diagnosis according to TBSRTC. The model produces highly accurate results; however, there are several limitations in categorizing cases as indeterminate lesions or FN using the model. Up to 30% of the FNA specimens were diagnosed as indeterminate [8]. In addition, because TBSRTC is not an ordinal system, the risk of misclassification and intra- or inter-observer variation are inevitable. Because indeterminate diagnosis leads to unnecessary surgery [8], the accuracy of FNA diagnostic screening must be increased. In addition, recent studies have already focused on building AI neural networks that can identify papillary carcinoma and follicular neoplasm[19,26]; therefore, the model employed in the present study, which only distinguishes between papillary carcinoma and benign nodules, is relatively conventional. In other words, malignancy predicted by the classifier should be evaluated for the cases that are correctly identified as atypia with undetermined significance or follicular neoplasm.

Another limitation is that the study was conducted at the image patch level, but the actual diagnosis is performed at the WSI level and is based on clinical information in routine practice, making it difficult to directly apply the results of this study to the clinical field. The performance of this AI cannot be guaranteed in situations where more comprehensive judgment is required, and external validation using external datasets, preferably WSI, unrelated to training, is required to prove its usefulness as a diagnostic tool.

To address the aforementioned limitations, it is necessary to develop an algorithm that guarantees accuracy both at the WSI level and image level. Furthermore, it would be ideal to develop a tool that can classify slide images as per the diagnostic criteria recommended by TBSRTC, such as a model that can categorize cases into atypia with undetermined significance or follicular neoplasm rather than only distinguish between the nuclei of papillary carcinoma and benign nodules. We can consider training with a setup called multiple instance learning as a way to let the AI predict a diagnosis of WSI from multiple image patches. Multiple instance learning groups multiple

separated items into a single bag with a global decision and requires semi-supervised learning, which uses less training effort and is considered to be optimized for methodologies such as ours[38]; however, literature has shown that multiple instance learning has lower accuracy than the methods using supervised learning.[25] Further studies are needed to address these challenges.

Conclusion

We have successfully developed an AI model that distinguishes the malignant papillary thyroid carcinoma from benign lesions using image patches from FNA cytology slides of thyroid nodules. The model helped improve the diagnostic accuracy of expert pathologists. This study is significant in that it used datasets collected from multiple nationwide institutions, utilized images including multiple z-stacks, showed a high accuracy rate of 99.7%, and helped improve the diagnostic accuracy of expert pathologists. The performance in WSI prediction cannot be guaranteed, and the results are based on classification into only two categories, benign and malignant, rather than a diagnosis based on TBSRTC categories, which remains a challenge and should be addressed in future research.

Acknowledgments: We thank Ms. In Park (CUK), Dr. Ah Reum Kim (CUK), Dr. Seona Shin (National Cancer Center (NCC)), Dr. Na Young Han (NCC), Dr. Joon Young Shin, and Dr. Ms. Sook Hee Kang for their data collection, retrieval, management, annotation, and quality checks. We used datasets from The Open AI Dataset Project (AI-Hub, South Korea). All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

Author Contributions: Conceptualization: Gyungyub Gong, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Methodology: Gyungyub Gong, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; software: Yosep Chong.; Validation: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Formal Analysis: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Investigation: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Resources: Yosep Chong.; Data Curation: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Writing and Original Draft Preparation: Yujin Lee, Hyun Joo Choi, and Yosep Chong.; Writing Review and Editing: Yujin Lee,

Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Visualization: Yujin Lee, Hyun Joo Choi, and Yosep Chong.; Supervision: Hyun Joo Choi, and Yosep Chong.; and Project Administration: Yosep Chong. All the authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors, Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong, declare no conflicts of interest.

Funding: This work was partially supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (2021R1A2C2013630).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Figures Legends:

Fig. 1. Schematic workflow of this study.

Fig. 2. Overview of the Open AI Dataset Project.

Fig. 3. Enhanced focus fusion processing to generate evenly perfectly focused ‘extended z-stack images’ for 3 dimensional cell clusters of the cytology samples from 3-6 layers of different z-axis focus.

Fig. 4. Image pre-processing including focus fusion (extended z-stack image generation), color normalization, image patches extraction from WSIs, and resizing.

Fig. 5. Sensitivity, specificity, and accuracy of the convolutional neural network models in the pre-test to select the best model for benign/cancer image patch classification.

Fig. 6. Representative images of malignant (A) and benign (B) samples that were correctly diagnosed by the AI model; (C) malignant cells that were misdiagnosed as benign by the AI model; (D) clusters of benign follicular cells misdiagnosed as malignant.

Fig. 7. Representative images classified differently by the cytopathologists and the AI model. AI correctly diagnosed benign (a) and malignant (b) samples that were misdiagnosed by all the cytopathologists; examples of benign (c) and malignant (d) smears correctly diagnosed by all the cytopathologists but misdiagnosed by the AI

model.

Fig. 8. Comparison of the performances between the human cytopathologists and AI model. In addition to displaying the sensitivity, specificity, and accuracy results for humans and their combination with AI, the figure also presents the standard deviation.

Table 1. Number of whole-slide images (image patches) used for training, validation, and testing.

	Training	Validation	Test	Total
Benign	207(7,379)	7(331)	14(271)	228(7,981)
Malignant	63(7,524)	6(234)	9(236)	78(7,994)
Total	270(14,903)	13(565)	23(507)	306(15,975)

Table 2. Performance of AI model (Inception ResNet v2) for training, validation, and test dataset (Augmented dataset).

	Training	Validation	Test
Accuracy	99.72% (99.64-99.80%)	97.70% (96.46-98.94%)	94.87% (92.95-96.79%)
Sensitivity	99.87% (99.81-99.93%)	99.57% (99.03-100.00%)	100.00% (100.00-100.00%)
Specificity	99.58% (99.48-99.68%)	96.37% (94.83-97.91%)	90.41% (87.85-92.97%)

The data in parentheses presents a 95% confidence interval.

418 Supplementary Table 1. Institutions Participating in the Open AI dataset.

Institutions	Number of WSIs
Ace Pathology Clinic	4
Asan Chungmu Hospital	1
BHS Hanseo Hospital	1
BML Clinic	1
Bundang Jesaeng Hospital	1
Busan Medical Center	2
Busan National University Yangsan Hospital	1
Busan St. Mary's Hospital	1
C&Y Pathology Clinic	1
Catholic University of Korea Bucheon St. Mary's Hospital	2
Catholic University of Korea Incheon St. Mary's Hospital	1
Catholic University of Korea Seoul St. Mary's Hospital	2
Catholic University of Korea St. Vincent's Hospital	1
Catholic University of Korea Uijeongbu St. Mary's Hospital	**74**
Cha University Bundang Cha Hospital	1
Cha University Gangnam Cha Hospital	1
Cha University Gumi Cha Hospital	1
Cheongju Medical Center	1
Cheongju St. Mary's Hospital	1
Chonnam National University Bitgoeul Hospital	2
Daedong Hospital	1
Daegu Catholic University Hospital	2
Daejeon Seon Hospital	1
Dankook University Hospital	1
Dongguk University Ilsan Buddhism Hospital	1
Dongmasan Hospital	1
Ehwa Clinic	1
Eulji University Hospital	2
Eulji University Hospital	1
Ewon Pathology Clinic	7
For You Pathology Clinic	3
Gangneung Asan Hospital	1
Gil Hospital	2
Gimpo Woori Hospital	2
Guro Sacred Heart Hospital	3
Gwangju Christian Hospital	3
Gyeongsang National University Hospital	1
H Plus Yangji Hospital	1
Hallym Hospital	2
Hallym University Dongtan Sacred Heart Hospital	1
Hallym University Gangnam Sacred Heart Hospital	1
Hanil Hospital	1

Hanmaeum Hospital	1
Hanmi Clinic	3
Hanyang University Guri Hospital	1
Hwasun Chonnam National University Hospital	2
Inha University Hospital	1
Inje University Busan Paik Hospital	1
Inje University Haeundae Paik Hospital	1
Inje University Ilsan Paik Hospital	1
Jeil Hospital	1
Jeju Halla Hospital	2
Jesus Hospital	1
Jisam Hospital	2
Keimyung University Dongsan Hospital	1
Kim Minkyung Pathology Clinic	1
Konkuk University Hospital	1
Korea CFC Pathology Clinic	2
Korea Medical Research Institute Central Branch	1
Korea University College of Medicine, Anam Hospital	2
Korea University Guro Hospital	1
Kosin University Gospel Hospital	1
Kyungpook National University Hospital	1
Mokpo Korea Hospital	2
MS Pathology Clinic	2
Ms. Medi Hospital	1
National Medical Center	1
Open Doctors Pathology Clinic	2
Raphael Hospital	3
Samsung Changwon Hospital	3
Samyuk Seoul Hospital	1
Sejong Hospital	3
Seongnam Central Hospital	2
Seoul Medical Center	4
Seoul Metropolitan Boramae Hospital	1
Seoul National University Bundang Hospital	1
Soonchunhyang University Cheonan Hospital	1
Soonchunhyang University Gumi Hospital	3
St. Carollo Hospital	3
St. Mary's Pathology Clinic	2
Sungae Hospital	1
Yeocheon Jeonnam Hospital	1
Yeongnam University Hospital	1
Yonsei University Severance Hospital	99
Yonsei University Wonju Severance Christian Hospital	1
Yujin Pathology Clinic	1
Total 86 institutions	306

Supplementary Table 2. Whole-slide image data distribution.

Characteristics		Number of WSIs
Sex	Male	110
	Female	196
Age	< 55	132
	≥ 55	174
Classification	Malignant	78
	Benign	228
Method	Conventional	67
	LBP	239
Staining	H&E	121
	Pap	185
Number of z-stack layers	3	96
	5	209
	6	1

LBP, Liquid-based preparation; Pap, Papanicolaou; WSIs, whole slide images.

Supplementary Table 3. Performance of AI model (Inception ResNet v2) for training, validation, and test dataset with augmented dataset, non-augmented (all) dataset, and reduced dataset.

	Training			validation			Test		
	Augmented dataset	Non-augmented (All) dataset	Reduced dataset	Augmented dataset	Non-augmented (All) dataset	Reduced dataset	Augmented dataset	Non-augmented (All) dataset	Reduced dataset
Accuracy	99.72% (99.64-99.80%)	99.82% (99.73-99.91%)	99.89% (99.78-100.00%)	99.70% (96.46-98.94%)	99.47% (99.87-100.00%)	100.00% (100.00-100.00%)	94.87% (92.95-96.79%)	95.66% (93.89-97.43%)	97.04% (95.56-98.52%)
Sensitivity	99.87% (99.81-99.93%)	99.79% (99.70-99.88%)	99.84% (99.71-99.97%)	99.57% (99.03-100.00%)	99.15% (98.39-99.91%)	100.00% (100.00-100.00%)	100.00% (100.00-100.00%)	99.58% (99.02-100.00%)	99.58% (99.02-100.00%)
Specificity	99.58% (99.48-99.68%)	99.82% (99.73-99.91%)	99.95% (99.88-100.00%)	96.37% (94.83-97.91%)	99.70% (99.25-100.00%)	100.00% (100.00-100.00%)	90.41% (87.85-92.97%)	92.25% (89.92-94.58%)	94.83% (92.90-96.76%)

The data in parentheses presents a 95% confidence interval.

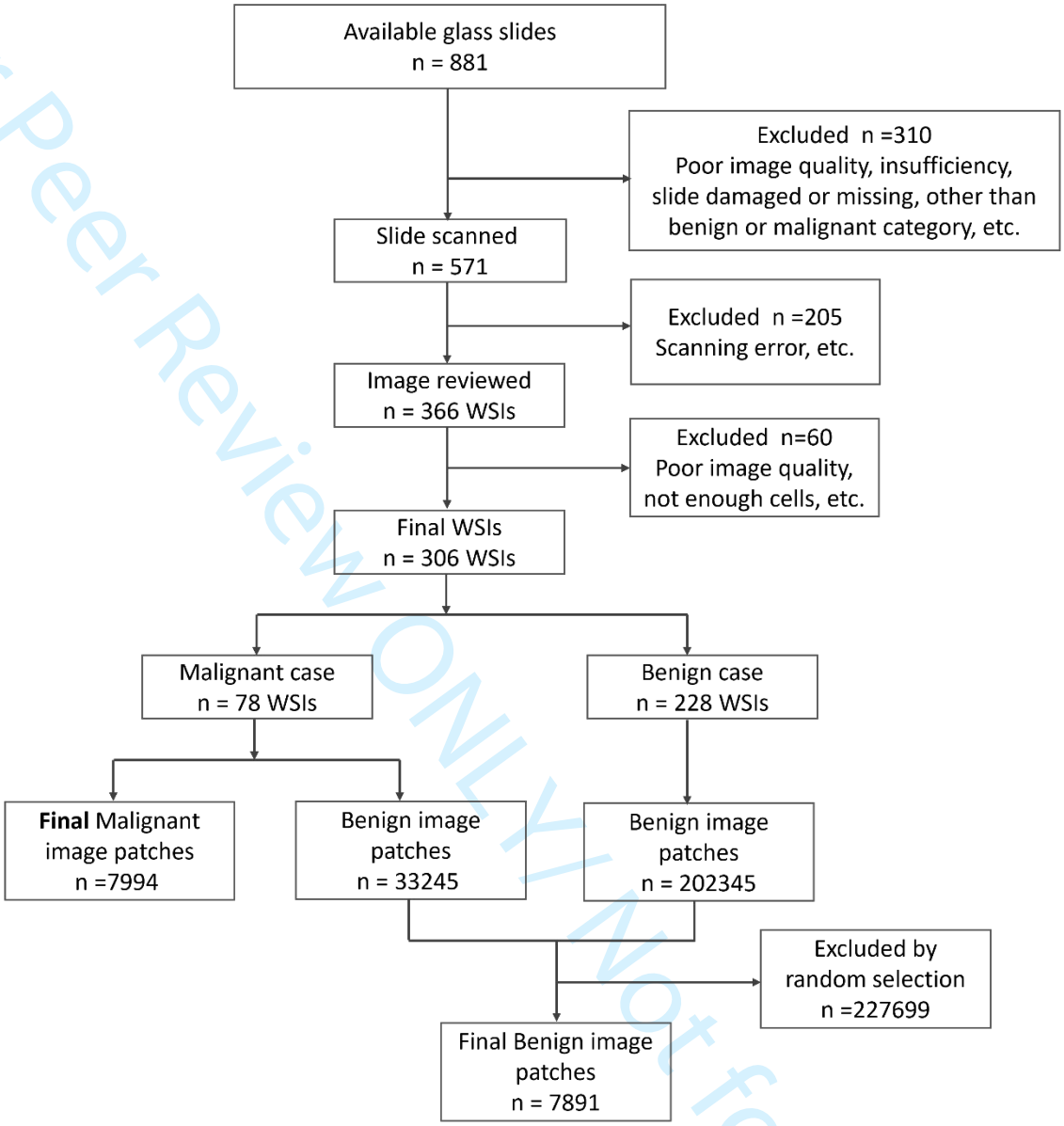
Supplementary Table 4. Performance of AI models in the classification of thyroid nodules in literature.

No.	Author	Year	Country	Task	Staining and preparation Method	Dataset	Pixel-level	Sampling	Z stacking images	External Cross-validation	Base Model	Performance	Pathologist number
1	Varlatzidou [39]	2011	Greece	**Classification** Benign/Malignant	Pap	335 patients (32 887 nuclei)	1024 × 768	FNAC	ND	ND	ANN (LVQ)	Sens: 93.80% Spec: 94.11% Acc: 94.05%	NA
2	Gopinath (1) [40]	2013	India	**Nuclear segmentation/ Classification** Benign/Malignant	Pap	110 patches	256 × 256	FNAC	ND	ND	SVM/ k-NN,	Sens: 95% Spec: 100% Acc: 96.7%	ATLAS committee
3	Gopinath (2) [41]	2013	India	**Nuclear segmentation/ Classification** Benign/Malignant	Pap	110 patches	256 × 256	FNAC	ND	ND	SVM/ ENN/ k-NN	Sens: 90% Spec: 100% Acc: 93.3%	ATLAS committee
4	Gopinath (3) [42]	2015	India	**Nuclear segmentation/ Classification** Benign/Malignant	Pap	110 patches	256 × 256	FNAC	ND	ND	SVM/ ENN/ k-NN/ DT	Sens: 100% Spec: 90% Acc: 96.6%	ATLAS committee
5	Savala [43]	2017	India	**Classification** FA/FC	May Grunwald–Giemsa/H&E	57 cases (57patches)	NA	FNAC	ND	ND	ANN	Acc: 100% AUC: 1.00%	2

6	Gopinath (4) [44]	2018	India	**Classification** Benign/Malignant	Pap	110 patches	256 × 256	FNAC	ND	ND	ANN/ ENN	Sens: 95% Spec: 100% Acc: 96.7%	ATLAS committee
7	Sanyal [45]	2018	India	**Classification** PTC/non PTC	Pap	370 patches	512 × 512	FNAC	ND	ND	CNN	Sens: 90.48% Spec: 83.33% Acc: 85.1%	NA
8	Dov [25]	2019	USA	**Classification** Benign/Malignant	Pap	908 WSIs (5461 patches)	150000 × 100000	FNAC	ND	ND	CNN (VGG-11)	Sens: 92% Spec: 90.5%	3
9	Guan [21]	2019	China	**Classification** Benign/PTC	H&E	279 WSI (887 patch images)	224 × 224	FNAC	ND	ND	VGG-16/ Inception-V3	Sens 100% Spec 94.91% Acc: 97.6%	1
10	Range [19]	2020	USA	**Classification** Benign/Malignant	Pap	659 patients (908 WSIs) (4494 patches)	NA	FNAC	Yes	ND	Machine learning & CNNs	Sens: 92.0% Spec: 90.5% AUC: 0.93%	1
11	Frago-poulos [20]	2020	Greece	**Classification** Benign/Malignant	Pap	447 WSI (41,324 nuclei)	1024 × 768	FNAC	ND	ND	ANN (RBF)	Sens: 95.0%, Spec: 95.5%	NA

12	Dov [26]	2022	USA	**Classification** Benign/Malignant	Diff-Quik Pap	908 WSIs (100 ROIs per each cases)	NA	FNAC	Yes	ND	CNN (VGG-11)	AUC: 93.1%	1
13	Current study	2023	Korea, Republic of	**Classification** Benign/Malignant	Pap/H&E	306 WSIs (15,975 patches)	1024 × 1024	FNAC	Yes	ND	CNN (Inception ResNet v2)	Sens: 99.81% Spec: 99.61% Acc: 99.71%	3

Pap, Papanicolaou; FNAC, fine-needle aspiration cytology; ND, not done; ANN, artificial neural network; NA, not applicable; LVQ, learning vector quantization; Sens, sensitivity; Spec, specificity; Acc, accuracy; SVM, state variable model; k-NN, k-nearest neighbor algorithm; ENN, environmental neural network; DT, digital transformation; FA, follicular adenoma; FC, follicular carcinoma; PTC, papillary thyroid carcinoma; CNN, convolutional neural network; VGG, visual geometry group; RBF, radial basis function.



Supplementary Figure 1. participant flow diagram

Table: The STARD Checklist Table.

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	P2
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	P2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	P3-P5
	4	Study objectives and hypotheses	P5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	P5
Participants	**6**	Eligibility criteria	P5-P6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	P6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	N/A
	9	Whether participants formed a consecutive, random or convenience series	N/A
Test methods	**10a**	Index test, in sufficient detail to allow replication	N/A
	10b	Reference standard, in sufficient detail to allow replication	4
	11	Rationale for choosing the reference standard (if alternatives exist)	4
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	N/A
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	N/A
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	N/A
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	N/A
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	P6-P7
	15	How indeterminate index test or reference standard results were handled	N/A
	16	How missing data on the index test and reference standard were handled	N/A
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	N/A
	18	Intended sample size and how it was determined	P7-P8
RESULTS			
Participants	**19**	Flow of participants, using a diagram	P22
	20	Baseline demographic and clinical characteristics of participants	P18
	21a	Distribution of severity of disease in those with the target condition	N/A
	21b	Distribution of alternative diagnoses in those without the target condition	N/A
	22	Time interval and any clinical interventions between index test and reference standard	N/A
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	N/A
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	N/A
	25	Any adverse events from performing the index test or the reference standard	N/A
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	P11-P12
	27	Implications for practice, including the intended use and clinical role of the index test	N/A
OTHER INFORMATION			
	28	Registration number and name of registry	N/A
	29	Where the full study protocol can be accessed	P3-P7
	30	Sources of funding and other support; role of funders	P13

References

1. Pouliakis A, Karakitsou E, Margari N, et al. Artificial Neural Networks as Decision Support Tools in Cytopathology: Past, Present, and Future. Biomed Eng Comput Biol 2016;7(1-18, doi:10.4137/BECB.S31601

2. Thakur N, Yoon H, Chong Y. Current Trends of Artificial Intelligence for Colorectal Cancer Pathology Image Analysis: A Systematic Review. Cancers (Basel) 2020;12(7):1884, doi:10.3390/cancers12071884

3. Ailia MJ, Thakur N, Abdul-Ghafar J, et al. Current Trend of Artificial Intelligence Patents in Digital Pathology: A Systematic Evaluation of the Patent Landscape. Cancers (Basel) 2022;14(10):2400, doi:10.3390/cancers14102400

4. Li Z, Jiang Y, Li B, et al. Development and Validation of a Machine Learning Model for Detection and Classification of Tertiary Lymphoid Structures in Gastrointestinal Cancers. JAMA Netw Open 2023;6(1):e2252553, doi:10.1001/jamanetworkopen.2022.52553

5. Park HS, Chong Y, Lee Y, et al. Deep Learning-Based Computational Cytopathologic Diagnosis of Metastatic Breast Carcinoma in Pleural Fluid. Cells 2023;12(14):1847

6. Dey P. Artificial neural network in diagnostic cytology. Cytojournal 2022;19(27, doi:10.25259/Cytojournal_33_2021

7. Lollie TK, Krane JF. Applications of Computational Pathology in Head and Neck Cytopathology. Acta Cytol 2021;65(4):330-334, doi:10.1159/000513286

8. Ho AS, Sarti EE, Jain KS, et al. Malignancy rate in thyroid nodules classified as Bethesda category III (AUS/FLUS). Thyroid 2014;24(5):832-9, doi:10.1089/thy.2013.0317

9. Gharib H, Goellner JR. Fine-needle aspiration biopsy of the thyroid: an appraisal. Ann Intern Med 1993;118(4):282-9, doi:10.7326/0003-4819-118-4-199302150-00007

10. Bongiovanni M, Krane JF, Cibas ES, et al. The atypical thyroid fine-needle aspiration: past, present, and future. Cancer Cytopathol 2012;120(2):73-86, doi:10.1002/cncy.20178

11. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid 2017;27(11):1341-1346, doi:10.1089/thy.2017.0500

12. Kim M, Park HJ, Min HS, et al. The Use of the Bethesda System for Reporting Thyroid Cytopathology in Korea: A Nationwide Multicenter Survey by the Korean Society of Endocrine Pathologists. J Pathol Transl Med 2017;51(4):410-417, doi:10.4132/jptm.2017.04.05

13. Straccia P, Rossi ED, Bizzarro T, et al. A meta-analytic review of the Bethesda System for Reporting Thyroid Cytopathology: Has the rate of malignancy in indeterminate lesions been underestimated? Cancer Cytopathol 2015;123(12):713-22, doi:10.1002/cncy.21605

14. Ko YS, Hwang TS, Kim JY, et al. Diagnostic Limitation of Fine-Needle Aspiration (FNA) on Indeterminate Thyroid Nodules Can Be Partially Overcome by Preoperative Molecular Analysis: Assessment of RET/PTC1 Rearrangement in BRAF and RAS Wild-Type Routine Air-Dried FNA Specimens. Int J Mol Sci 2017;18(4):806, doi:10.3390/ijms18040806

15. Boon ME, Lowhagen T, Willems JS. Planimetric studies on fine needle aspirates from follicular adenoma and follicular carcinoma of the thyroid. Acta Cytol 1980;24(2):145-8

16. Chain K, Legesse T, Heath JE, et al. Digital image-assisted quantitative nuclear analysis

improves diagnostic accuracy of thyroid fine-needle aspiration cytology. Cancer Cytopathol 2019;127(8):501-513, doi:10.1002/cncy.22120

17. Karakitsos P, Cochand-Priollet B, Pouliakis A, et al. Learning vector quantizer in the investigation of thyroid lesions. Anal Quant Cytol Histol 1999;21(3):201-8

18. Cochand-Priollet B, Koutroumbas K, Megalopoulou TM, et al. Discriminating benign from malignant thyroid lesions using artificial intelligence and statistical selection of morphometric features. Oncol Rep 2006;15 Spec no.(1023-6, doi:10.3892/or.15.4.1023

19. Elliott Range DD, Dov D, Kovalsky SZ, et al. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. Cancer Cytopathol 2020;128(4):287-295, doi:10.1002/cncy.22238

20. Fragopoulos C, Pouliakis A, Meristoudis C, et al. Radial Basis Function Artificial Neural Network for the Investigation of Thyroid Cytological Lesions. J Thyroid Res 2020;2020(5464787, doi:10.1155/2020/5464787

21. Guan Q, Wang Y, Ping B, et al. Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. J Cancer 2019;10(20):4876-4882, doi:10.7150/jca.28769

22. Donnelly AD, Mukherjee MS, Lyden ER, et al. Optimal z-axis scanning parameters for gynecologic cytology specimens. Journal of pathology informatics 2013;4(1):38

23. Thakur N, Alam MR, Abdul-Ghafar J, et al. Recent application of artificial intelligence in non-gynecological cancer cytopathology: a systematic review. Cancers 2022;14(14):3529

24. Chong Y, Hong SA, Oh HK, et al. Diagnostic proficiency test using digital cytopathology and comparative assessment of whole slide images of cytologic samples for quality assurance program in Korea. Journal of Pathology and Translational Medicine 2023;57(5):251-264

25. Dov D, Kovalsky SZ, Cohen J, et al. Thyroid cancer malignancy prediction from whole slide cytopathology images. PMLR: 2019.

26. Dov D, Kovalsky SZ, Feng Q, et al. Use of Machine Learning-Based Software for the Screening of Thyroid Cytopathology Whole Slide Images. Arch Pathol Lab Med 2022;146(7):872-878, doi:10.5858/arpa.2020-0712-OA

27. Genius Digital Diagnostic System, USA. Avaliable at: https://www.hologic.com/hologic-products/cytology/genius-digital-diagnostics-system/ Acessed December 25 2023.

28. Pramana, Inc. (Morrisville, North Carolina). Avaliable at: https://pramana.ai/ Acessed December 25 2023.

29. Evans AJ, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology: guideline update from the College of American Pathologists in collaboration with the American Society for Clinical Pathology and the Association for Pathology Informatics. Archives of pathology & laboratory medicine 2022;146(4):440-450

30. Marletta S, Salatiello M, Pantanowitz L, et al. Delphi expert consensus for whole slide imaging in thyroid cytopathology. Cytopathology 2023;

31. Girolami I, Marletta S, Pantanowitz L, et al. Impact of image analysis and artificial

intelligence in thyroid pathology, with particular reference to cytological aspects. Cytopathology 2020;31(5):432-444

32. Chantziantoniou N. BestCyte® Cell Sorter Imaging System: Primary and adjudicative whole slide image rescreening review times of 500 ThinPrep Pap test thin-layers-An intra-observer, time-surrogate analysis of diagnostic confidence potentialities. Journal of Pathology Informatics 2022;13(100095

33. Nikiforov YE, Seethala RR, Tallini G, et al. Nomenclature Revision for Encapsulated Follicular Variant of Papillary Thyroid Carcinoma: A Paradigm Shift to Reduce Overtreatment of Indolent Tumors. JAMA Oncol 2016;2(8):1023-9, doi:10.1001/jamaoncol.2016.0386

34. Abeyrathna KD, Granmo O-C, Goodwin M. Extending the Tsetlin Machine With Integer-Weighted Clauses for Increased Interpretability. IEEE Access 2021;9(8233-8248, doi:10.1109/access.2021.3049569

35. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020;58(82-115, doi:10.1016/j.inffus.2019.12.012

36. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018;6(52138-52160, doi:10.1109/access.2018.2870052

37. Nazir S, Dickson DM, Akram MU. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Comput Biol Med 2023;156(106668, doi:10.1016/j.compbiomed.2023.106668

38. Quellec G, Cazuguel G, Cochener B, et al. Multiple-instance learning for medical image and video analysis. IEEE reviews in biomedical engineering 2017;10(213-234

39. Varlatzidou A, Pouliakis A, Stamataki M, et al. Cascaded learning vector quantizer neural networks for the discrimination of thyroid lesions. Anal Quant Cytol Histol 2011;33(6):323-34

40. Gopinath B, Shanthi N. Support Vector Machine based diagnostic system for thyroid cancer using statistical texture features. Asian Pac J Cancer Prev 2013;14(1):97-102, doi:10.7314/apjcp.2013.14.1.97

41. Gopinath B, Shanthi N. Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images. Australas Phys Eng Sci Med 2013;36(2):219-30, doi:10.1007/s13246-013-0199-8

42. Gopinath B, Shanthi N. Development of an Automated Medical Diagnosis System for Classifying Thyroid Tumor Cells using Multiple Classifier Fusion. Technol Cancer Res Treat 2015;14(5):653-62, doi:10.7785/tcrt.2012.500430

43. Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. Diagn Cytopathol 2018;46(3):244-249, doi:10.1002/dc.23880

44. Gopinath B. A benign and malignant pattern identification in cytopathological images of thyroid nodules using gabor filter and neural networks. Asian Journal of Convergence In Technology 2018;4(1):

45. Sanyal P, Mukherjee T, Barui S, et al. Artificial Intelligence in Cytopathology: A Neural Network to Identify Papillary Carcinoma on Thyroid Fine-Needle Aspiration Cytology Smears. J Pathol Inform 2018;9(43, doi:10.4103/jpi.jpi_43_18

Improved diagnostic accuracy of thyroid fine-needle aspiration cytology with artificial intelligence technology

Running title: AI Improves Accuracy of Thyroid FNA Diagnosis

Yujin Lee[1], Mohammad Rizwan Alam[2], Hongsik Park[1], Kwangil Yim[2], Kyung Jin Seo[2], Gisu Hwang[3], Dahyeon Kim[3], Yeonsoo Chung[3], Gyungyub Gong[4], Nam Hoon Cho[5], Chong Woo Yoo[6], Yosep Chong[2,*], Hyun Joo Choi[1,*]

[1]Department of Hospital Pathology, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, Suwon, Republic of Korea; wondoocha@naver.com (Y.L.); griselbrand@gmail.com (H.S.P.); chj0103@catholic.ac.kr (H.J.C)

[2]Department of Hospital Pathology, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Uijeongbu, Republic of Korea.; rizwan@catholic.ac.kr (M.R.A.); kangse_manse@catholic.ac.kr (K.Y.); ywacko@catholic.ac.kr (K.J.S.); ychong@catholic.ac.kr (Y.C.)

[3]AI Team, DeepNoid Inc., Seoul, Korea; kisu031@gmail.com (G.H.); anniy8920@outlook.kr (D.K.); yeonsoo00.chung@gmail.com (Y.S.C)

[4]Department of Pathology, Asan Medical Center, Seoul, Korea; gygong@amc.seoul.kr (G.G.)

[5]Department of Pathology, Yonsei University College of Medicine, Seoul, Korea; CHO1988@yuhs.ac (N.C.)

[6]Department of Pathology, National Cancer Center, Ilsan, Gyeonggi-do, Republic of Korea; cwy@ncc.re.kr (C.W.Y.)

***Correspondence:**
Hyun Joo Choi, MD., PhD.

Address: Department of Hospital Pathology, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, 93, Jungbu-daero, Paldal-gu, Suwon 16247, Gyeonggi-do, Republic of Korea

Tel: (+82)-031-249-7592

E-mail: chj0103@catholic.ac.kr

Fax: (+82)-031-244-6786

Yosep Chong, MD., PhD

Address: Department of Hospital Pathology, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 271, Cheonbo-ro, Uijeongbu 11765, Gyeonggi-do, Republic of Korea.

Tel: (+82)-032-820-3160

E-Mail: ychong@catholic.ac.kr

Fax: (+82)-032-820-3877

Conflict of Interests: The authors declare that they have no competing interests.

Funding: This work was partially supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (2021R1A2C2013630).

Abstract

Background: Artificial intelligence (AI) is increasingly being applied in pathology and cytology, showing promising results. We collected a large dataset of whole slide image of thyroid fine needle aspiration cytology (FNA), incorporating z-stacking, from institutions across the nation to develop an AI model.

Methods: We conducted a multi-center retrospective diagnostic accuracy study using thyroid FNA dataset from the Open AI Dataset Project that consists of digitalized images samples collected from three university hospitals and 215 Korean institutions through extensive quality check during the case selection, scanning, labeling, and reviewing process. Multiple z-layer images were captured using three different scanners and image patches were extracted from whole slide images and resized after focus-fusion and color normalization. We pre-tested six AI models, determining Inception ResNet v2 as the best model using a subset of dataset, and subsequently tested the final model with total datasets. Additionally, we compared the performance of AI and cytopathologists using randomly selected 1,031 image patches and reevaluated the cytopathologists' performance after reference to AI results.

Results: A total of 10,332 image patches from 306 thyroid FNAs, comprising 78 malignant (Papillary thyroid carcinoma) and 228 benign from 86 institutions were used for the AI training. Inception ResNet v2 achieved highest accuracy of 99.7%, 97.7%, and 94.9% for training, validation, and test dataset, respectively (Sensitivity 99.9%, 99.6%, and 100% and specificity 99.6%, 96.4%, and 90.4% for training, validation, and test dataset, respectively). In the comparison between AI and human, AI model showed higher accuracy and specificity than the average expert cytopathologists beyond the two-standard deviation (Accuracy 99.71% (95% CI, 99.38-100.00%~~0.99-1.00~~) vs. 88.91% (95% CI, 86.99-90.83%~~0.86-0.90~~), sensitivity 99.81% (95% CI,99.54-100.00% ~~0.99-1.00~~) vs. 87.26% (95% CI, 85.22-89.30%~~0.85-0.89~~), and specificity 99.61% (95% CI, 99.23-99.99%~~0.99-0.99~~) vs. 90.58% (95% CI, 88.80-92.36%~~0.88-0.92~~). Moreover, after referring to the AI results, all the performance of the experts increased (Accuracy 96%, 95%, and 96%, respectively) as well as diagnostic agreement (from 0.64 to 0.84).

Conclusions: These results suggest that the application of AI technology to thyroid FNA cytology may improve the diagnostic accuracy as well as intra- and inter-observer variability among pathologists. Further confirmatory research is needed.

Keywords: thyroid; cytology; fine-needle aspiration; artificial intelligence; thyroid neoplasms; deep learning

Introduction

Artificial intelligence (AI) is emerging in cytopathologic image analysis with promising results. Advancement in AI have enabled the application of computer models, such as artificial neural networks, deep learning, genetic algorithms, and classification, to support diagnosis and treatment decisions.[1] Computational models have been developed to perform pathological image analyses using machine learning.[2-5] Image analysis can be used not only for surgical or biopsy slide samples but also for cytologic slides.[6] It can be used to objectify the diagnostic process, mitigate human effort, and improve the productivity of cytopathological examinations.

In non-gynecologic cytology, thyroid fine needle aspiration (FNA) is the most actively explored field of AI application.[7] This is because AI diagnosis is the preferred modality for the evaluation of thyroid nodules owing to easy access and a generally high accuracy. FNA cytology for thyroid nodules is ideal for applying AI owing to the relatively specific diagnostic criteria for thyroid nodules. Thyroid nodules are observed in an estimated 10% of the general population, of which 5–7% suffer from malignant thyroid nodules.[8] FNA is currently considered the preferred modality for the evaluation of thyroid nodules.[9,10] FNA diagnosis is less invasive compared to others; it is also rapid and has high accuracy. Cytopathologists follow The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC)[11,12], which comprises six categories to evaluate the risk of malignancy in thyroid FNA samples. The categories are as follows: non-diagnostic, benign, atypia of undetermined significance/follicular lesion of undetermined significance, follicular neoplasm/suspicious for follicular neoplasm, suspicious for malignancy, and malignant. The sensitivity and specificity of the current system range from 68% to 98% and 56% to 100%, respectively.[8,11,13,14]

Studies on the application of AI in thyroid FNA have promising results; however, they had several limitations, such as a small sample size, lack of z-stacking, and the collection of data only from a single institute. Computational methods have been applied to the cytological analysis of thyroid nodules since 1980.[15,16] In 1999, Karakitsos et al.[17] classified the images of benign and malignant thyroid tumors using a learning vector quantizer artificial neural network and achieved 97.8% accuracy. Study by Cochand-Priollet, B. et al. identified malignant and benign nodules using May-Grunwald-Giemsa-stained smears and four independent classifiers. Subsequently, the performances of the classifiers were compared. The most successful classifier, the k-nearest neighbor, achieved an overall accuracy of 88.71%.[18] Convolutional neural networks have been used to predict malignancy

from 908 whole slide images (WSIs), and a sensitivity of 92.0% and specificity of 90.5%[19] have been achieved. In another study, an artificial neural network was used to classify 447 cases into two categories, benign and malignant, using the radial basis function, and an overall sensitivity of 95.0% and specificity of 95.5% were achieved.[20] Guan, Q. et al. conducted a study using a deep convolutional neural network and fragmented image, which were classified using VGG-16; they obtained an accuracy of 97.66%.[21] Previous studies have evaluated the diagnostic performance of AI, but few studies have explored the AI usefulness with enough amount of training and validation datasets.

The second challenge of developing effective AI for thyroid cytology is the lack of consideration for z-stacking in the training and validation datasets. In cytology, distinct from histologic samples, individual cells and three-dimensional cell clusters are suspended in a medium, floating in the space between the glass and cover slides, typically spanning 50-100 microns. This space is referred to as the z-axis, and it is crucial to scan cytologic slides at various z-layers of focus to capture clearly focused images of these clusters.[22] This is especially important for samples like thyroid FNAs, which naturally include large papillary clusters. Previously, there has been no special focus on z-stacking for digital cytology images in AI research related to cytology. [22,23] However, through our recent comparative analysis of image quality across different scanners, we have identified the minimal requirements for appropriate z-stack layers based on sample types.[24] For instance, at least three layers are necessary for liquid-based preparations (LBP) and five layers for conventional smears. Scanning digital cytology images with these specified z-stack layers, tailored to the sample type, is vital for developing robust AI models that can be effectively applied in daily practice. Moreover, as the third challenge, AI models in past studies have been developed and validated using datasets from only a single or two institutions.[23] This approach lacks the diversity and scale needed for good generalizability. There is a pressing need for multi-institutional or national level datasets to enhance the AI models' applicability and reliability in various clinical settings.

To address the aforementioned limitations, we built a dataset that contained a sufficient number of cases and z-stacks from nationwide institutions to develop an AI model for thyroid FNAs.

Material and Methods

The study was approved by the Institutional Review Board of the Catholic University of Korea, College of Medicine (UC21SNSI0064), the Institutional Review Board of the Yonsei University College of Medicine (4-2021-0569), Institutional Review Board of the National Cancer Center (NCC2021-0145), and Institutional Review Board of the St. Vincent's Hospital College of Medicine, Catholic University of Korea (VC22RISI0131). The informed consents from the patients were waived by these IRBs. The use of quality assurance program slides of

the Korean Society of Cytopathology was approved by the society executive board. A schematic overview of the proposed method and workflow is shown in Fig. 1.

Data collection

This is a multi-center retrospective diagnostic study. Training, validation, and testing were performed using images of thyroid FNA specimens from a dataset provided by an AI research project known as "the Open AI Dataset Project" (Fig. 2). The Open AI dataset consists of digitalized images of cytopathology slides collected from three university hospitals and quality control slides from 215 institutions in South Korea. Because of the data imbalance, we only included papillary carcinoma but no other histologic subtypes for malignant cases. The slides were scanned using AT2 (Leica Biosystems, Germany), NanoZoomer S360 (Hammatsu, Japan), and Pannoramic 250 Flash III (3DHistech, Hungary) at a ×40 objective magnification at the focal planes. Each dataset contained multiple z-stacks, with conventional smears having five layers and LBP slides having one to three layers. The selection of cytologic samples was stringently managed by 12 board-certified cytopathologists and 5 cytologists. Each sample was initially confirmed to match its corresponding histologic slides by the institute's board-certified cytopathologists. Following this, another set of board-certified cytopathologists assessed the slide quality and determined their suitability for inclusion in the Open AI dataset. Subsequently, the quality of the scanned images was meticulously reviewed by another group of board-certified cytopathologists.

A total of 5500 WSIs including all the non-gynecologic samples such as respiratory tract, body fluid, urine, thyroid FNAs, and other FNAs were collected. 1100 WSIs of textbook cases were from quality control slides of 215 all registered cytopathology laboratories in Korea achieved for quality assurance program of the Korean Society for Cytopathology, while another 4400 WSIs of daily practice cases were collected from 12 major hospitals in Korea. Among the Open AI Dataset, we exclusively utilized thyroid FNA samples other than other types of samples for this study. We also excluded the samples that are not eligible for AI training such as bad image quality. Finally, thyroid 306 FNAs samples were collected from 86 institutions and used in this study (Supplementary Table 1). We included LBP samples as well as conventional smear, Papanicolaou stained samples as well as hematoxylin and eosin-stained samples.

From the 78 malignant WSIs, 7,994 malignant image patches and 33,245 benign image patches were created. From the 228 benign WSIs, 202,345 benign image patches were created. Due to data imbalance, all the benign patches from malignant and benign WSIs were merged, resulting in a total of 235,590 benign image patches. To address this imbalance, 227,699 benign image patches were randomly excluded. Finally, for this study, 7,994 malignant and 7,891 benign image patches were used for the development of the AI model.

Image pre-processing

To deal with the multiple images with different z-axis, we first generated the extended z-stack images from 3-6 images with different z-axis by enhanced focus fusion technology (Fig. 3). By this process, we obtained WSIs with only 1 evenly focused layer for 3-dimensional cell clusters. Color normalization was performed on the integrated images before patch extraction. Images were cropped and extracted as 1024 × 1024 pixels-sized image patches and resized to 256 × 256 pixels before using them as inputs for model training (Fig. 4). The image patches were evenly divided into grids from the WSIs. Each patch image was reviewed by trained cytopathologists to confirm the benign or malignant and another group of cytopathologists reviewed the labeling and the image patches with discordant opinions were excluded from the dataset as mentioned in the prior section. Image patches with no follicular cells have been removed, and image patches with both benign and malignant follicular cells have been considered as malignant. The RAND function in Excel was utilized to assign a random number to each image patch for the selection Microsoft Excel (Version 2021, Build. 14827.20192).

Image patch labeling

For the categorization of image patches, those extracted from negative samples were collected as negative image patches. In contrast, patches derived from positive samples were further classified into two categories: those containing cancer cells and those without. This classification was performed by board-certified cytopathologists. Finally, to ensure the utmost accuracy and consistency, these categorized image patches underwent a further review by yet another group of board-certified cytopathologists. If there is a discordant opinion between 1st and 2nd reviewers, the image patches were removed from the dataset. This rigorous process underscores the commitment to the highest standards of accuracy and reliability in this dataset for AI training.

From the 78 malignant WSIs, 7,994 malignant image patches and 33,245 benign image patches were created. From the 228 benign WSIs, 202,345 benign image patches were created. Due to data imbalance, all the benign patches from malignant and benign WSIs were merged, resulting in a total of 235,590 benign image patches. To address this imbalance, 227,699 benign image patches were randomly excluded. Finally, for this study, 7,994 malignant and 7,891 benign image patches were used for the development of the AI model.

Pre-test for AI-model selection

Prior to model selection, we tested six models: Inception ResNet v2, MobileNet v2, Efficientnet-b1, Densenet 121, ResNext50, and ResNet50. Subsequently, we compared their performances. The training, validation, and testing were performed using random WSI patches of 78 malignant and 88 benign cases. A total of 2,351 image

patches were created from the malignant and 2,518 from the benign WSIs. The number of benign and malignant patches used for training, validation, and testing was adjusted similarly, and a total of 3,797 image patches were used for training, 565 for validation, and 507 for testing.

AI model training

For model training, 15,975 augmented patch images were created from the dataset. The images were flipped vertically and horizontally and rotated clockwise for data augmentation. The augmented dataset included 7,981 benign and 7,994 malignant image patches. A total of 14,903 image patches were used for training, 565 for validation, and 507 for testing (Table 1). The performances of the AI models trained using augmented, non-augmented, and reduced datasets were also compared along with a 95% confidence interval (95% CI).

Comparison of the performances between experts and AI model

Of the 15,975 image patches, 1,031 image patches were randomly selected to compare the diagnostic accuracies of experts and the AI model. The RAND function in Excel was utilized to assign a random number to each image patch for the selection Microsoft Excel (Version 2021, Build. 14827.20192). The diagnostic accuracy of the AI model was compared with that of three experienced cytopathologists. In addition, the diagnostic accuracy of the cytopathologists after referring to the AI results was investigated. Standard deviations were also analyzed for the comparison results.

Statistical analysis

To analyze the diagnostic agreement among pathologists, the Fleiss' Kappa coefficient was used. The 95% CI of the Fleiss' Kappa coefficient was also analyzed. The analysis was performed using R statistical programming (Version 3.4.1; http//www.r-project.org, accessed on May 21, 2023).

Results

Data collection

A total of 306 WSIs were included in the dataset, 78 of which were malignant (Papillary thyroid carcinoma) and 228 were benign (25.5% vs. 74.5%), 133 cases were the textbook cases with typical cytologic features from the Quality Assurance Program slides from 85 institutions, while the other 173 cases were daily practice-level cases collected from Uijeongbu St. Mary's Hospital and Yonsei University Severance Hospital (43.5% vs. 56.5%). Of

the slides included in the dataset, 121 were hematoxylin and eosin-stained, and the remaining 185 were Papanicolaou-stained (39.5% vs. 60.5%). The dataset contained 239 LBP slides, which is approximately thrice the number of conventional cytology slides (67, 78.1% vs. 21.9%). 149 cases were scanned by Leica AT2, 83 by Hamamatsu, and 74 by 3DHistech scanner (48.7%, 27.1%, and 24.2%, respectively). Each WSI scan contained at least three to six z-stack layers. Supplementary Table 2 summarizes clinical and pathological information about the dataset included.

Pre-test for AI model selection

Inception ResNet v2 exhibited the highest accuracy of 97.04%. MobileNet v2, Efficientnet-b1, Densenet 121, ResNext50, and ResNet50 achieved an accuracy of 95.07%, 94.67%, 95.07%, 94.28% and 95.27%, respectively. Inception ResNet v2 exhibited a high sensitivity and specificity. MobileNet v2 achieved 100% sensitivity; however, the accuracy was low because the specificity was only 90.77%. By contrast, ResNext50 exhibited the highest specificity but a relatively low sensitivity (Fig. 5).

AI model training

Inception ResNet v2 showed the highest accuracy when the augmented datasets were used for training (Table 2). The model obtained 99.72% (95% CI, 99.64-99.80%~~0.99-0.99~~) accuracy, 99.87% (95% CI, 99.81-99.93%~~0.99-0.99~~) sensitivity, and 99.58% (95% CI, 99.48-99.68%~~0.99-0.99~~) specificity for training dataset, 97.70% (95% CI, 96.46-98.94%~~0.96-0.98~~) accuracy, 99.57% (95% CI, 99.03-100.00%~~0.99-1.00~~) sensitivity, and 96.37% (95% CI, 94.83-97.91%~~0.94-0.97~~) specificity for validation dataset, 94.87% (95% CI, 92.95-96.79%~~0.92-0.96~~) accuracy, 100% (95% CI, 100.00-100.00%~~1.00-1.00~~) sensitivity, and 90.41% (95% CI, 87.85-92.97%~~0.87-0.92~~) specificity for test dataset. Examples correctly diagnosed and misdiagnosed samples by AI (Fig. 6). Result of the non-augmented, and the reduced dataset is shown in Supplementary Table 3.

Comparison of the performances between the AI model and the experts

A total of 1,031 image patches were randomly selected from all datasets for comparison. The image patches included 513 negative cells and 518 malignant cells (Papillary thyroid carcinoma). Three pathologists and the AI model reviewed and labeled the patches. The three pathologists showed an average sensitivity of 87.26% (95% CI, 85.22-89.30%~~0.85-0.89~~), a specificity of 90.58% (95% CI, 88.80-92.36%~~0.88-0.92~~), and an accuracy of 88.91% (95% CI, 86.99-90.83%~~0.86-0.90~~). The Fleiss' Kappa coefficient for diagnostic agreement between pathologists

A, B and C was 0.66 (95% CI, 0.63-0.68), indicating moderate agreement. The AI model showed an average sensitivity of 99.81% (95% CI, 99.54-100.00%~~0.99-1.00~~), a specificity of 99.61% (95% CI, 99.23-99.99%~~0.99-0.99~~), and an accuracy of 99.71% (95% CI, 99.38-100.00%~~0.99-1.00~~). Examples of images diagnosed differently by the pathologists and the AI model are shown in Fig. 7. After referring to the diagnostic results obtained by the AI model, the diagnostic accuracy of the three pathologists increased to 95.76% (95% CI, 94.53-96.99%~~0.94-0.96~~). The sensitivity and specificity increased to 95.24% (95% CI, 93.94-96.54%~~0.93-0.96~~) and 96.30% (95% CI, 95.15-97.42%~~0.95-0.97~~), respectively (Fig. 8). The Kappa coefficient for diagnostic agreement between pathologists also increased from 0.66 (95% CI, 0.63-0.68) to 0.86 (95% CI, 0.83-0.88), indicating very good agreement.

Discussion

This study developed an AI model, which exhibited a competitive performance with an accuracy of 99.7%, using 15,975 thyroid image patches obtained from 306 WSIs. The image patches were collected from approximately 200 hospitals nationwide with the help of the Korean Society for Cytopathology and three major institutions participating in the Open AI Dataset Project. The accuracy of the AI model was higher than that of the pathologists (99.71% vs. 88.91%). In addition, the diagnostic accuracy of the pathologists was improved upon referring to the AI results.

The slides included in the dataset were collected nationwide. To include a sufficient number of z-stacks in the image and reduce the data-storage space required by the z-stacks, at least three layers were integrated to create an extended z-stack image. Previous studies have reported that utilizing a z-stack was difficult because it required a large storage space.[19] Many studies either avoided using any z-stack images or used only the middle layer of complete images;[19,25,26] however, multi-focus images must be analyzed to increase the accuracy of diagnosis based on cytology slides. Image integration can be applied to obtain clear images of multiple foci without necessitating large capacities. There are a recent technical advances in the scanning technology for cytology such as volumetric scanning by Hologic (Marlborough, MA), multi-camera image capture and focus fusion technology by Vieworks (Anyang, Republic of Korea), and pixel level z-scanning technology by Pramana, Inc. (Morrisville, North Carolina). [27,28] This new technique will revolutionize the practical application of digital cytology and the use of AI.

Although the revised College of American Pathologists guideline regarding the validation of cytology for

diagnostic purposes indicates insufficient evidence supporting the use of digital cytology for primary diagnosis, recently many efforts have been made to strengthen the idea of the role of digital cytology and AI.[29-31] Recent studies using AI technology on thyroid FNAs have shown very promising results in addition to the gynecologic cytology screening AI software that were recently introduced in the market with CE-mark and FDA-approval by several companies, such as Hologic (Genius Digital Diagnostic System, USA), TechCyte (UT, USA), Datexim (CytoProcessor, France), Cell Solutions (BestCyte Cell Sorter Imaging System, USA), Landing Med (China), and KF Bio (China).[27,32] Some of these AI software adopted more advanced scanning technology that enhanced the scanning speed and image quality such as volumetric scanning by Hologic (Marlborough, MA), multi-camera focus-fusion technology by Vieworks (Anyang, Republic of Korea), and pixel level z-stacking by Pramana, Inc. (Morrisville, North Carolina).[28] These changes show the possibility of fundamental changes in digital cytology and the potential application of AI soon.

Our model showed the best diagnostic performance of 99.71% accuracy in the thyroid FNAs . The highest accuracy reported for the cytological classification of thyroid nodules in the previous studies is 97.66%.[21] A study in 2020, developed an artificial neural network model and obtain an accuracy of 95% using a larger number of image patches and 447 WSIs[20] (Supplementary Table 4). Among the papers reported to date, the study that utilized the largest number of samples, 908 WSIs, achieved a sensitivity of 92% and specificity of 90.5% through VGG-11, a type of Convolutional neural network.[19] In a subsequent study conducted by the same authors, the performances of AI models were compared with that of expert pathologists in predicting malignancy from WSIs; the AI models and experts showed similar accuracies.[26] Interestingly, the performances of the AI models improved and the diagnosis became similar to that of the experts when the models underwent a fully supervised training to identify the region of interest.

In this study, the diagnostic accuracy of the AI model was significantly higher than that of the experts, and the diagnostic accuracy of the experts improved after referring to the AI results. Previous studies have compared human and AI performance and assess diagnostic agreement with screeners[26], this is the first study to evaluate how diagnostic accuracy changes when humans refer to AI results. Of the 1,031 image patches used for comparison, only one image patch of a malignant case was accurately diagnosed by the experts but was misdiagnosed by the AI model (Fig. 7d). As is well known, typical papillary thyroid carcinoma has nuclear features such as enlargement, elongation, crowding, irregular contours, grooves, pseudoinclusions and chromatin clearing.[33] The case exhibited the characteristic features of typical papillary thyroid carcinoma, such as

overlapping nuclei and intranuclear pseudoinclusions; however, the image patch showed relatively fine chromatin and smooth cell membrane. The AI model misdiagnosed benign nodules as malignant in two image patches. In both patches, it showed many lymphocytes and histiocytes, which might have been mistaken for malignant cells (Fig. 7c). In lymphocytic thyroiditis, the nuclei of neighboring follicular cells frequently undergo a slight enlargement and exhibit atypical features in response to chronic inflammation. This phenomenon could account for the AI model's erroneous positive predictions. By contrast, the AI model accurately distinguished the characteristics of malignant cells with pseudoinclusions or irregular cell membranes, even in the images of regions with low cellularity (Fig. 7b). After referring to the AI's results, the pathologists were able to identify small or fuzzy pseudoinclusions to base their judgment on. In addition, the AI model accurately diagnosed benign nodules containing clumping follicular cells, which the experts misdiagnosed as malignant nodules (Fig. 7a). In the actual benign cases where AI got the diagnosis right and the pathologist got it wrong, the confusing factors that influenced the pathologist's judgment were features like focal overlapping nuclei, high cell density, and relatively irregular nuclear size. However, the nuclei of these cells retained a fine chromatin pattern and smooth membrane, which lacks evidence of malignancy.

While it is possible to compare images of cases which the AI made a correct or incorrect diagnosis and assume that cells such as histiocytes would have been misdiagnosed as malignant, or a small intranuclear inclusion may helped it diagnose the malignant cell correctly, it is actually difficult to interpret exactly what criteria the AI uses to distinguish between malignant and benign cells. In the case of this study, it's especially hard to know which cases AI is more likely to misdiagnose, since AI got most of the diagnoses right. This makes it difficult to decide which of the AI's judgments to trust and which not to trust. Recently, quantitative scoring methods have been attempted to develop interpretable AI[4], and in the case of typical papillary thyroid carcinoma, we can think of ways to let the AI calculate the nuclear scoring system[33] to make a quantitative assessment or mark areas in the image that are likely to be malignant cells. However, it is important to consider that there is a trade-off between explainability and performance.[34-37]

Our study made three pathologists determine the initial diagnosis of 1,031 image patches, revise their diagnoses based on the AI's diagnosis, and then compare their results. After noting that some of their diagnoses were different from AI readings, the pathologists reviewed and revised at least one of their misdiagnoses. Referring to AI gave pathologists the opportunity to review their diagnosis in more detail if it differs from the AI's, and ultimately contributed to improving agreement level and diagnostic accuracy. Initially, the Fleiss' Kappa coefficient

diagnostic agreement between pathologists A, B and C was 0.66, indicating moderate agreement. Pathologist A tended to have high specificity and low sensitivity, while pathologist C had high sensitivity and low specificity (95.71%, 71.04%, 95.37% and 85.96%, respectively, data not shown). Pathologist B's sensitivity was higher than A's and the same as C's, specificity was higher than C's and lower than A's, and accuracy was the highest. The sensitivity, specificity, and accuracy of all three pathologists increased after referring the AI's diagnosis. Notably, pathologist A's sensitivity increased to 90.73%, and pathologist C's specificity increased to 96.88% after referring the AI's diagnosis (data not shown). The diagnostic agreement between pathologists also increased to 0.86, indicating very good agreement. This means that AI can help calibrate pathologists' diagnoses to increase sensitivity or specificity and reduce inter-observer variation due to individual pathologist tendencies.

However, this study has a few limitations. We classified thyroid cell slides into only two categories, malignant and benign, without segmenting the diagnosis according to TBSRTC. The model produces highly accurate results; however, there are several limitations in categorizing cases as indeterminate lesions or FN using the model. Up to 30% of the FNA specimens were diagnosed as indeterminate [8]. In addition, because TBSRTC is not an ordinal system, the risk of misclassification and intra- or inter-observer variation are inevitable. Because indeterminate diagnosis leads to unnecessary surgery [8], the accuracy of FNA diagnostic screening must be increased. In addition, recent studies have already focused on building AI neural networks that can identify papillary carcinoma and follicular neoplasm[19,26]; therefore, the model employed in the present study, which only distinguishes between papillary carcinoma and benign nodules, is relatively conventional. In other words, malignancy predicted by the classifier should be evaluated for the cases that are correctly identified as atypia with undetermined significance or follicular neoplasm.

Another limitation is that the study was conducted at the image patch level, but the actual diagnosis is performed at the WSI level and is based on clinical information in routine practice, making it difficult to directly apply the results of this study to the clinical field. The performance of this AI cannot be guaranteed in situations where more comprehensive judgment is required, and external validation using external datasets, preferably WSI, unrelated to training, is required to prove its usefulness as a diagnostic tool.

To address the aforementioned limitations, it is necessary to develop an algorithm that guarantees accuracy both at the WSI level and image level. Furthermore, it would be ideal to develop a tool that can classify slide images as per the diagnostic criteria recommended by TBSRTC, such as a model that can categorize cases into atypia with undetermined significance or follicular neoplasm rather than only distinguish between the nuclei of papillary

carcinoma and benign nodules. We can consider training with a setup called multiple instance learning as a way to let the AI predict a diagnosis of WSI from multiple image patches. Multiple instance learning groups multiple separated items into a single bag with a global decision and requires semi-supervised learning, which uses less training effort and is considered to be optimized for methodologies such as ours[38]; however, literature has shown that multiple instance learning has lower accuracy than the methods using supervised learning.[25] Further studies are needed to address these challenges.

Conclusion

We have successfully developed an AI model that distinguishes the malignant papillary thyroid carcinoma from benign lesions using image patches from FNA cytology slides of thyroid nodules. The model helped improve the diagnostic accuracy of expert pathologists. This study is significant in that it used datasets collected from multiple nationwide institutions, utilized images including multiple z-stacks, showed a high accuracy rate of 99.7%, and helped improve the diagnostic accuracy of expert pathologists. The performance in WSI prediction cannot be guaranteed, and the results are based on classification into only two categories, benign and malignant, rather than a diagnosis based on TBSRTC categories, which remains a challenge and should be addressed in future research.

Acknowledgments: We thank Ms. In Park (CUK), Dr. Ah Reum Kim (CUK), Dr. Seona Shin (National Cancer Center (NCC)), Dr. Na Young Han (NCC), Dr. Joon Young Shin, and Dr. Ms. Sook Hee Kang for their data collection, retrieval, management, annotation, and quality checks. We used datasets from The Open AI Dataset Project (AI-Hub, South Korea). All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

Author Contributions: Conceptualization: Gyungyub Gong, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Methodology: Gyungyub Gong, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; software: Yosep Chong.; Validation: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Formal Analysis: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Investigation: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Resources: Yosep Chong.; Data Curation: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung,

Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Writing and Original Draft Preparation: Yujin Lee, Hyun Joo Choi, and Yosep Chong.; Writing Review and Editing: Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong.; Visualization: Yujin Lee, Hyun Joo Choi, and Yosep Chong.; Supervision: Hyun Joo Choi, and Yosep Chong.; and Project Administration: Yosep Chong. All the authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors, Yujin Lee, Mohammad Rizwan Alam, Hongsik Park, Kwangil Yim, Kyung Jin Seo, Gisu Hwang, Dahyeon Kim, Yeonsoo Chung, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Hyun Joo Choi, and Yosep Chong, declare no conflicts of interest.

Funding: This work was partially supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (2021R1A2C2013630).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Figures Legends:

Fig. 1. Schematic workflow of this study.

Fig. 2. Overview of the Open AI Dataset Project.

Fig. 3. Enhanced focus fusion processing to generate evenly perfectly focused ‘extended z-stack images’ for 3 dimensional cell clusters of the cytology samples from 3-6 layers of different z-axis focus.

Fig. 4. Image pre-processing including focus fusion (extended z-stack image generation), color normalization, image patches extraction from WSIs, and resizing.

Fig. 5. Sensitivity, specificity, and accuracy of the convolutional neural network models in the pre-test to select the best model for benign/cancer image patch classification.

Fig. 6. Representative images of malignant (A) and benign (B) samples that were correctly diagnosed by the AI model; (C) malignant cells that were misdiagnosed as benign by the AI model; (D) clusters of benign follicular cells misdiagnosed as malignant.

Fig. 7. Representative images classified differently by the cytopathologists and the AI model. AI correctly diagnosed benign (a) and malignant (b) samples that were misdiagnosed by all the cytopathologists; examples of benign (c) and malignant (d) smears correctly diagnosed by all the cytopathologists but misdiagnosed by the AI model.

Fig. 8. Comparison of the performances between the human cytopathologists and AI model. In addition to displaying the sensitivity, specificity, and accuracy results for humans and their combination with AI, the figure also presents the standard deviation.

Table 1. Number of whole-slide images (image patches) used for training, validation, and testing.

	Training	Validation	Test	Total
Benign	207(7,379)	7(331)	14(271)	228(7,981)
Malignant	63(7,524)	6(234)	9(236)	78(7,994)
Total	270(14,903)	13(565)	23(507)	306(15,975)

Table 2. Performance of AI model (Inception ResNet v2) for training, validation, and test dataset (Augmented dataset).

	Training	Validation	Test
Accuracy	99.72%(99.64-99.80%~~0.99-0.99~~)	97.70% (96.46-98.94%~~0.96-0.98~~)	94.87% (92.95-96.79%~~0.92-0.96~~)
Sensitivity	99.87%(99.81-99.93%~~0.99-0.99~~)	99.57% (99.03-100.00%~~0.99-1.00~~)	100.00%(100.00-100.00%~~1.00-1.00~~)
Specificity	99.58%(99.48-99.68%~~0.99-0.99~~)	96.37% (94.83-97.91%~~0.94-0.97~~)	90.41% (87.85-92.97%~~0.87-0.92~~)

The data in parentheses presents a 95% confidence interval.

421 Supplementary Table 1. Institutions Participating in the Open AI dataset.

Institutions	Number of WSIs
Ace Pathology Clinic	4
Asan Chungmu Hospital	1
BHS Hanseo Hospital	1
BML Clinic	1
Bundang Jesaeng Hospital	1
Busan Medical Center	2
Busan National University Yangsan Hospital	1
Busan St. Mary's Hospital	1
C&Y Pathology Clinic	1
Catholic University of Korea Bucheon St. Mary's Hospital	2
Catholic University of Korea Incheon St. Mary's Hospital	1
Catholic University of Korea Seoul St. Mary's Hospital	2
Catholic University of Korea St. Vincent's Hospital	1
Catholic University of Korea Uijeongbu St. Mary's Hospital	**74**
Cha University Bundang Cha Hospital	1
Cha University Gangnam Cha Hospital	1
Cha University Gumi Cha Hospital	1
Cheongju Medical Center	1
Cheongju St. Mary's Hospital	1
Chonnam National University Bitgoeul Hospital	2
Daedong Hospital	1
Daegu Catholic University Hospital	2
Daejeon Seon Hospital	1
Dankook University Hospital	1
Dongguk University Ilsan Buddhism Hospital	1
Dongmasan Hospital	1
Ehwa Clinic	1
Eulji University Hospital	2
Eulji University Hospital	1
Ewon Pathology Clinic	7
For You Pathology Clinic	3
Gangneung Asan Hospital	1
Gil Hospital	2
Gimpo Woori Hospital	2
Guro Sacred Heart Hospital	3
Gwangju Christian Hospital	3
Gyeongsang National University Hospital	1
H Plus Yangji Hospital	1
Hallym Hospital	2
Hallym University Dongtan Sacred Heart Hospital	1
Hallym University Gangnam Sacred Heart Hospital	1
Hanil Hospital	1

Hanmaeum Hospital	1
Hanmi Clinic	3
Hanyang University Guri Hospital	1
Hwasun Chonnam National University Hospital	2
Inha University Hospital	1
Inje University Busan Paik Hospital	1
Inje University Haeundae Paik Hospital	1
Inje University Ilsan Paik Hospital	1
Jeil Hospital	1
Jeju Halla Hospital	2
Jesus Hospital	1
Jisam Hospital	2
Keimyung University Dongsan Hospital	1
Kim Minkyung Pathology Clinic	1
Konkuk University Hospital	1
Korea CFC Pathology Clinic	2
Korea Medical Research Institute Central Branch	1
Korea University College of Medicine, Anam Hospital	2
Korea University Guro Hospital	1
Kosin University Gospel Hospital	1
Kyungpook National University Hospital	1
Mokpo Korea Hospital	2
MS Pathology Clinic	2
Ms. Medi Hospital	1
National Medical Center	1
Open Doctors Pathology Clinic	2
Raphael Hospital	3
Samsung Changwon Hospital	3
Samyuk Seoul Hospital	1
Sejong Hospital	3
Seongnam Central Hospital	2
Seoul Medical Center	4
Seoul Metropolitan Boramae Hospital	1
Seoul National University Bundang Hospital	1
Soonchunhyang University Cheonan Hospital	1
Soonchunhyang University Gumi Hospital	3
St. Carollo Hospital	3
St. Mary's Pathology Clinic	2
Sungae Hospital	1
Yeocheon Jeonnam Hospital	1
Yeongnam University Hospital	1
Yonsei University Severance Hospital	99
Yonsei University Wonju Severance Christian Hospital	1
Yujin Pathology Clinic	1
Total 86 institutions	306

Supplementary Table 2. Whole-slide image data distribution.

Characteristics		Number of WSIs
Sex	Male	110
	Female	196
Age	< 55	132
	≥ 55	174
Classification	Malignant	78
	Benign	228
Method	Conventional	67
	LBP	239
Staining	H&E	121
	Pap	185
Number of z-stack layers	3	96
	5	209
	6	1

LBP, Liquid-based preparation; Pap, Papanicolaou; WSIs, whole slide images.

Supplementary Table 3. Performance of AI model (Inception ResNet v2) for training, validation, and test dataset with augmented dataset, non-augmented (all) dataset, and reduced dataset.

	Training			validation			Test		
	Augmented dataset	Non-augmented (All) dataset	Reduced dataset	Augmented dataset	Non-augmented (All) dataset	Reduced dataset	Augmented dataset	Non-augmented (All) dataset	Reduced dataset
Accuracy	99.72% (99.64-99.80%0.99-0.99)	99.82% (99.73-99.91%0.99-0.99)	99.89% (99.78-100.00%0.99-1.00)	99.70% (96.46-98.94%0.96-0.98)	99.47% (99.87-100.00%0.99-1.00)	100.00% (100.00-100.00%1.00-1.00)	94.87% (92.95-96.79%0.92-0.96)	95.66% (93.89-97.43%0.93-0.97)	97.04% (95.56-98.52%0.95-0.98)
Sensitivity	99.87% (99.81-99.93%0.99-0.99)	99.79% (99.70-99.88%0.99-0.99)	99.84% (99.71-99.97%0.99-0.99)	99.57% (99.03-100.00%0.99-1.00)	99.15% (98.39-99.91%0.98-0.99)	100.00% (100.00-100.00%1.00-1.00)	100.00% (100.00-100.00%1.00-1.00)	99.58% (99.02-100.00%0.99-1.00)	99.58% (99.02-100.00%0.99-1.00)
Specificity	99.58% (99.48-99.68%0.99-0.99)	99.82% (99.73-99.91%0.99-0.99)	99.95% (99.88-100.00%0.99-	96.37% (94.83-97.91%0.94-0.97)	99.70% (99.25-100.00%0.99-	100.00% (100.00-100.00%1.00-	90.41% (87.85-92.97%0.87-	92.25% (89.92-94.58%0.89-0.94)	94.83% (92.90-96.76%0.92-

~~1.00~~)	~~1.00~~)	~~1.00~~)	~~0.92~~)	~~0.96~~)	

The data in parentheses presents a 95% confidence interval.

Supplementary Table 4. Performance of AI models in the classification of thyroid nodules in literature.

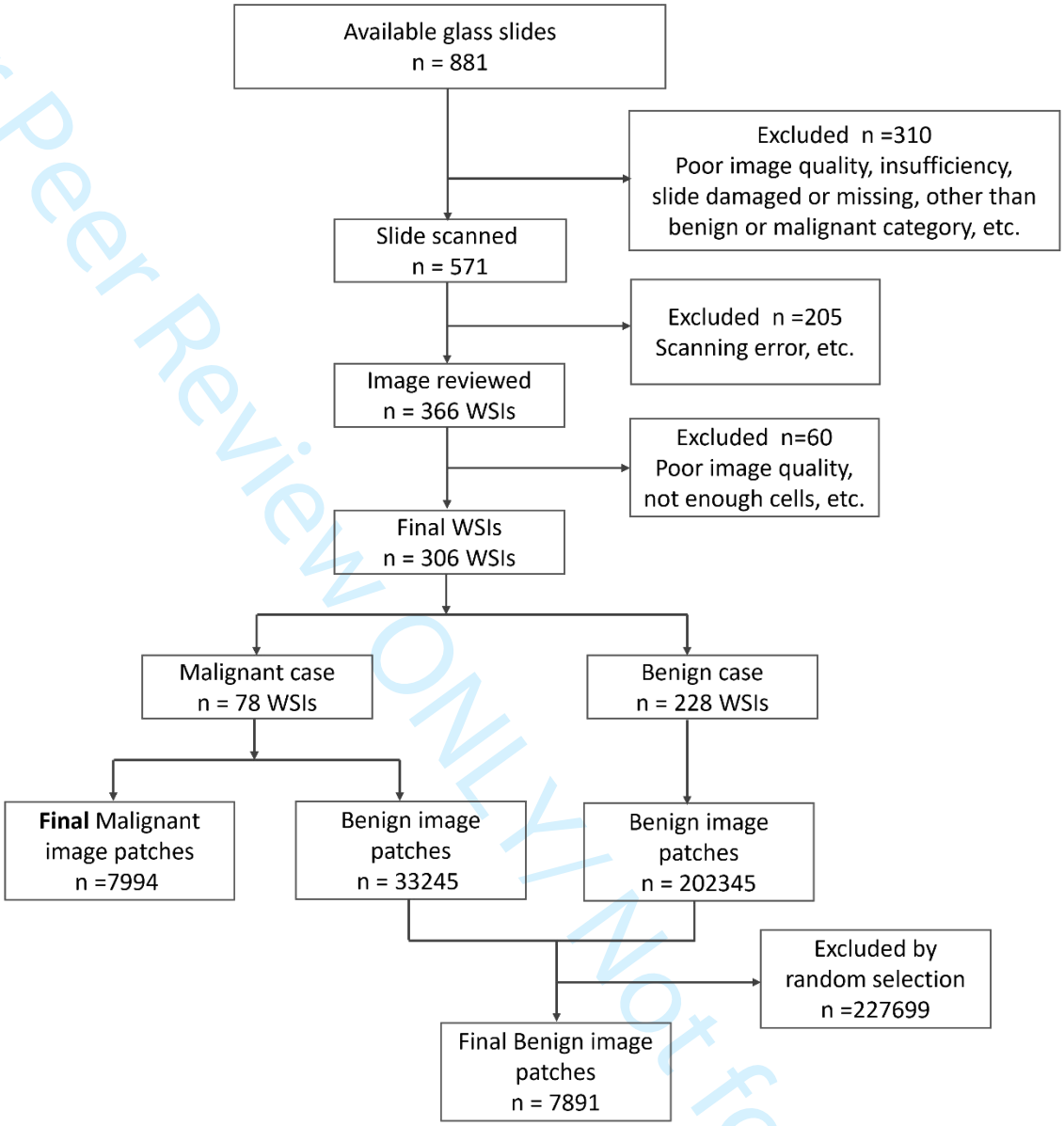
No.	Author	Year	Country	Task	Staining and preparation Method	Dataset	Pixel-level	Sampling	Z stacking images	External Cross-validation	Base Model	Performance	Pathologist number
1	Varlatzidou [39]	2011	Greece	**Classification** Benign/Malignant	Pap	335 patients (32 887 nuclei)	1024 × 768	FNAC	ND	ND	ANN (LVQ)	Sens: 93.80% Spec: 94.11% Acc: 94.05%	NA
2	Gopinath (1) [40]	2013	India	**Nuclear segmentation/ Classification** Benign/Malignant	Pap	110 patches	256 × 256	FNAC	ND	ND	SVM/ k-NN,	Sens: 95% Spec: 100% Acc: 96.7%	ATLAS committee
3	Gopinath (2) [41]	2013	India	**Nuclear segmentation/ Classification** Benign/Maligna	Pap	110 patches	256 × 256	FNAC	ND	ND	SVM/ ENN/ k-NN	Sens: 90% Spec: 100% Acc: 93.3%	ATLAS committee

nt

4	Gopinath (3) [42]	2015	India	**Nuclear segmentation/ Classification** Benign/Malignant	Pap	110 patches	256 × 256	FNAC	ND	ND	SVM/ ENN/ k-NN/ DT	Sens: 100% Spec: 90% Acc: 96.6%	ATLAS committee
5	Savala [43]	2017	India	**Classification** FA/FC	May Grunwald–Giemsa/H&E	57 cases (57patches)	NA	FNAC	ND	ND	ANN	Acc: 100% AUC: 1.00%	2
6	Gopinath (4) [44]	2018	India	**Classification** Benign/Malignant	Pap	110 patches	256 × 256	FNAC	ND	ND	ANN/ ENN	Sens: 95% Spec: 100% Acc: 96.7%	ATLAS committee
7	Sanyal [45]	2018	India	**Classification** PTC/non PTC	Pap	370 patches	512 × 512	FNAC	ND	ND	CNN	Sens: 90.48% Spec: 83.33% Acc: 85.1%	NA
8	Dov [25]	2019	USA	**Classification** Benign/Malignant	Pap	908 WSIs (5461 patches)	150000 × 100000	FNAC	ND	ND	CNN (VGG-11)	Sens: 92% Spec: 90.5%	3
9	Guan [21]	2019	China	**Classification** Benign/PTC	H&E	279 WSI (887 patch images)	224 × 224	FNAC	ND	ND	VGG-16/ Inception-V3	Sens 100% Spec 94.91%	1

												Acc: 97.6%	
10	Range [19]	2020	USA	**Classification** Benign/Malignant	Pap	659 patients (908 WSIs) (4494 patches)	NA	FNAC	Yes	ND	Machine learning & CNNs	Sens: 92.0% Spec: 90.5% AUC: 0.93%	1
11	Frago-poulos [20]	2020	Greece	**Classification** Benign/Malignant	Pap	447 WSI (41,324 nuclei)	1024 × 768	FNAC	ND	ND	ANN (RBF)	Sens: 95.0%, Spec: 95.5%	NA
12	Dov [26]	2022	USA	**Classification** Benign/Malignant	Diff-Quik Pap	908 WSIs (100 ROIs per each cases)	NA	FNAC	Yes	ND	CNN (VGG-11)	AUC: 93.1%	1
13	Current study	2023	Korea, Republic of	**Classification** Benign/Malignant	Pap/H&E	306 WSIs (15,975 patches)	1024 × 1024	FNAC	Yes	ND	CNN (Inception ResNet v2)	Sens: 99.81% Spec: 99.61% Acc: 99.71%	3

Pap, Papanicolaou; FNAC, fine-needle aspiration cytology; ND, not done; ANN, artificial neural network; NA, not applicable; LVQ, learning vector quantization; Sens, sensitivity; Spec, specificity; Acc, accuracy; SVM, state variable model; k-NN, k-nearest neighbor algorithm; ENN, environmental neural network; DT, digital transformation; FA, follicular adenoma; FC, follicular carcinoma; PTC, papillary thyroid carcinoma; CNN, convolutional neural network; VGG, visual geometry group; RBF, radial basis function.



Supplementary Figure 1. participant flow diagram

Table: The STARD Checklist Table.

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	P2
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	P2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	P3-P5
	4	Study objectives and hypotheses	P5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	P5
Participants	**6**	Eligibility criteria	P5-P6
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	P6
	8	Where and when potentially eligible participants were identified (setting, location and dates)	N/A
	9	Whether participants formed a consecutive, random or convenience series	N/A
Test methods	**10a**	Index test, in sufficient detail to allow replication	N/A
	10b	Reference standard, in sufficient detail to allow replication	4
	11	Rationale for choosing the reference standard (if alternatives exist)	4
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	N/A
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	N/A
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	N/A
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	N/A
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	P6-P7
	15	How indeterminate index test or reference standard results were handled	N/A
	16	How missing data on the index test and reference standard were handled	N/A
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	N/A
	18	Intended sample size and how it was determined	P7-P8
RESULTS			
Participants	**19**	Flow of participants, using a diagram	P22
	20	Baseline demographic and clinical characteristics of participants	P18
	21a	Distribution of severity of disease in those with the target condition	N/A
	21b	Distribution of alternative diagnoses in those without the target condition	N/A
	22	Time interval and any clinical interventions between index test and reference standard	N/A
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	N/A
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	N/A
	25	Any adverse events from performing the index test or the reference standard	N/A
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	P11-P12
	27	Implications for practice, including the intended use and clinical role of the index test	N/A
OTHER INFORMATION			
	28	Registration number and name of registry	N/A
	29	Where the full study protocol can be accessed	P3-P7
	30	Sources of funding and other support; role of funders	P13

References

1. Pouliakis A, Karakitsou E, Margari N, et al. Artificial Neural Networks as Decision Support Tools in Cytopathology: Past, Present, and Future. Biomed Eng Comput Biol 2016;7(1-18, doi:10.4137/BECB.S31601

2. Thakur N, Yoon H, Chong Y. Current Trends of Artificial Intelligence for Colorectal Cancer Pathology Image Analysis: A Systematic Review. Cancers (Basel) 2020;12(7):1884, doi:10.3390/cancers12071884

3. Ailia MJ, Thakur N, Abdul-Ghafar J, et al. Current Trend of Artificial Intelligence Patents in Digital Pathology: A Systematic Evaluation of the Patent Landscape. Cancers (Basel) 2022;14(10):2400, doi:10.3390/cancers14102400

4. Li Z, Jiang Y, Li B, et al. Development and Validation of a Machine Learning Model for Detection and Classification of Tertiary Lymphoid Structures in Gastrointestinal Cancers. JAMA Netw Open 2023;6(1):e2252553, doi:10.1001/jamanetworkopen.2022.52553

5. Park HS, Chong Y, Lee Y, et al. Deep Learning-Based Computational Cytopathologic Diagnosis of Metastatic Breast Carcinoma in Pleural Fluid. Cells 2023;12(14):1847

6. Dey P. Artificial neural network in diagnostic cytology. Cytojournal 2022;19(27, doi:10.25259/Cytojournal_33_2021

7. Lollie TK, Krane JF. Applications of Computational Pathology in Head and Neck Cytopathology. Acta Cytol 2021;65(4):330-334, doi:10.1159/000513286

8. Ho AS, Sarti EE, Jain KS, et al. Malignancy rate in thyroid nodules classified as Bethesda category III (AUS/FLUS). Thyroid 2014;24(5):832-9, doi:10.1089/thy.2013.0317

9. Gharib H, Goellner JR. Fine-needle aspiration biopsy of the thyroid: an appraisal. Ann Intern Med 1993;118(4):282-9, doi:10.7326/0003-4819-118-4-199302150-00007

10. Bongiovanni M, Krane JF, Cibas ES, et al. The atypical thyroid fine-needle aspiration: past, present, and future. Cancer Cytopathol 2012;120(2):73-86, doi:10.1002/cncy.20178

11. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid 2017;27(11):1341-1346, doi:10.1089/thy.2017.0500

12. Kim M, Park HJ, Min HS, et al. The Use of the Bethesda System for Reporting Thyroid Cytopathology in Korea: A Nationwide Multicenter Survey by the Korean Society of Endocrine Pathologists. J Pathol Transl Med 2017;51(4):410-417, doi:10.4132/jptm.2017.04.05

13. Straccia P, Rossi ED, Bizzarro T, et al. A meta-analytic review of the Bethesda System for Reporting Thyroid Cytopathology: Has the rate of malignancy in indeterminate lesions been underestimated? Cancer Cytopathol 2015;123(12):713-22, doi:10.1002/cncy.21605

14. Ko YS, Hwang TS, Kim JY, et al. Diagnostic Limitation of Fine-Needle Aspiration (FNA) on Indeterminate Thyroid Nodules Can Be Partially Overcome by Preoperative Molecular Analysis: Assessment of RET/PTC1 Rearrangement in BRAF and RAS Wild-Type Routine Air-Dried FNA Specimens. Int J Mol Sci 2017;18(4):806, doi:10.3390/ijms18040806

15. Boon ME, Lowhagen T, Willems JS. Planimetric studies on fine needle aspirates from follicular adenoma and follicular carcinoma of the thyroid. Acta Cytol 1980;24(2):145-8

16. Chain K, Legesse T, Heath JE, et al. Digital image-assisted quantitative nuclear analysis

improves diagnostic accuracy of thyroid fine-needle aspiration cytology. Cancer Cytopathol 2019;127(8):501-513, doi:10.1002/cncy.22120

17. Karakitsos P, Cochand-Priollet B, Pouliakis A, et al. Learning vector quantizer in the investigation of thyroid lesions. Anal Quant Cytol Histol 1999;21(3):201-8

18. Cochand-Priollet B, Koutroumbas K, Megalopoulou TM, et al. Discriminating benign from malignant thyroid lesions using artificial intelligence and statistical selection of morphometric features. Oncol Rep 2006;15 Spec no.(1023-6, doi:10.3892/or.15.4.1023

19. Elliott Range DD, Dov D, Kovalsky SZ, et al. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. Cancer Cytopathol 2020;128(4):287-295, doi:10.1002/cncy.22238

20. Fragopoulos C, Pouliakis A, Meristoudis C, et al. Radial Basis Function Artificial Neural Network for the Investigation of Thyroid Cytological Lesions. J Thyroid Res 2020;2020(5464787, doi:10.1155/2020/5464787

21. Guan Q, Wang Y, Ping B, et al. Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. J Cancer 2019;10(20):4876-4882, doi:10.7150/jca.28769

22. Donnelly AD, Mukherjee MS, Lyden ER, et al. Optimal z-axis scanning parameters for gynecologic cytology specimens. Journal of pathology informatics 2013;4(1):38

23. Thakur N, Alam MR, Abdul-Ghafar J, et al. Recent application of artificial intelligence in non-gynecological cancer cytopathology: a systematic review. Cancers 2022;14(14):3529

24. Chong Y, Hong SA, Oh HK, et al. Diagnostic proficiency test using digital cytopathology and comparative assessment of whole slide images of cytologic samples for quality assurance program in Korea. Journal of Pathology and Translational Medicine 2023;57(5):251-264

25. Dov D, Kovalsky SZ, Cohen J, et al. Thyroid cancer malignancy prediction from whole slide cytopathology images. PMLR: 2019.

26. Dov D, Kovalsky SZ, Feng Q, et al. Use of Machine Learning-Based Software for the Screening of Thyroid Cytopathology Whole Slide Images. Arch Pathol Lab Med 2022;146(7):872-878, doi:10.5858/arpa.2020-0712-OA

27. Genius Digital Diagnostic System, USA. Avaliable at: https://www.hologic.com/hologic-products/cytology/genius-digital-diagnostics-system/ Acessed December 25 2023.

28. Pramana, Inc. (Morrisville, North Carolina). Avaliable at: https://pramana.ai/ Acessed December 25 2023.

29. Evans AJ, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology: guideline update from the College of American Pathologists in collaboration with the American Society for Clinical Pathology and the Association for Pathology Informatics. Archives of pathology & laboratory medicine 2022;146(4):440-450

30. Marletta S, Salatiello M, Pantanowitz L, et al. Delphi expert consensus for whole slide imaging in thyroid cytopathology. Cytopathology 2023;

31. Girolami I, Marletta S, Pantanowitz L, et al. Impact of image analysis and artificial

intelligence in thyroid pathology, with particular reference to cytological aspects. Cytopathology 2020;31(5):432-444

32. Chantziantoniou N. BestCyte® Cell Sorter Imaging System: Primary and adjudicative whole slide image rescreening review times of 500 ThinPrep Pap test thin-layers-An intra-observer, time-surrogate analysis of diagnostic confidence potentialities. Journal of Pathology Informatics 2022;13(100095

33. Nikiforov YE, Seethala RR, Tallini G, et al. Nomenclature Revision for Encapsulated Follicular Variant of Papillary Thyroid Carcinoma: A Paradigm Shift to Reduce Overtreatment of Indolent Tumors. JAMA Oncol 2016;2(8):1023-9, doi:10.1001/jamaoncol.2016.0386

34. Abeyrathna KD, Granmo O-C, Goodwin M. Extending the Tsetlin Machine With Integer-Weighted Clauses for Increased Interpretability. IEEE Access 2021;9(8233-8248, doi:10.1109/access.2021.3049569

35. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020;58(82-115, doi:10.1016/j.inffus.2019.12.012

36. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018;6(52138-52160, doi:10.1109/access.2018.2870052

37. Nazir S, Dickson DM, Akram MU. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Comput Biol Med 2023;156(106668, doi:10.1016/j.compbiomed.2023.106668

38. Quellec G, Cazuguel G, Cochener B, et al. Multiple-instance learning for medical image and video analysis. IEEE reviews in biomedical engineering 2017;10(213-234

39. Varlatzidou A, Pouliakis A, Stamataki M, et al. Cascaded learning vector quantizer neural networks for the discrimination of thyroid lesions. Anal Quant Cytol Histol 2011;33(6):323-34

40. Gopinath B, Shanthi N. Support Vector Machine based diagnostic system for thyroid cancer using statistical texture features. Asian Pac J Cancer Prev 2013;14(1):97-102, doi:10.7314/apjcp.2013.14.1.97

41. Gopinath B, Shanthi N. Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images. Australas Phys Eng Sci Med 2013;36(2):219-30, doi:10.1007/s13246-013-0199-8

42. Gopinath B, Shanthi N. Development of an Automated Medical Diagnosis System for Classifying Thyroid Tumor Cells using Multiple Classifier Fusion. Technol Cancer Res Treat 2015;14(5):653-62, doi:10.7785/tcrt.2012.500430

43. Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. Diagn Cytopathol 2018;46(3):244-249, doi:10.1002/dc.23880

44. Gopinath B. A benign and malignant pattern identification in cytopathological images of thyroid nodules using gabor filter and neural networks. Asian Journal of Convergence In Technology 2018;4(1):

45. Sanyal P, Mukherjee T, Barui S, et al. Artificial Intelligence in Cytopathology: A Neural Network to Identify Papillary Carcinoma on Thyroid Fine-Needle Aspiration Cytology Smears. J Pathol Inform 2018;9(43, doi:10.4103/jpi.jpi_43_18

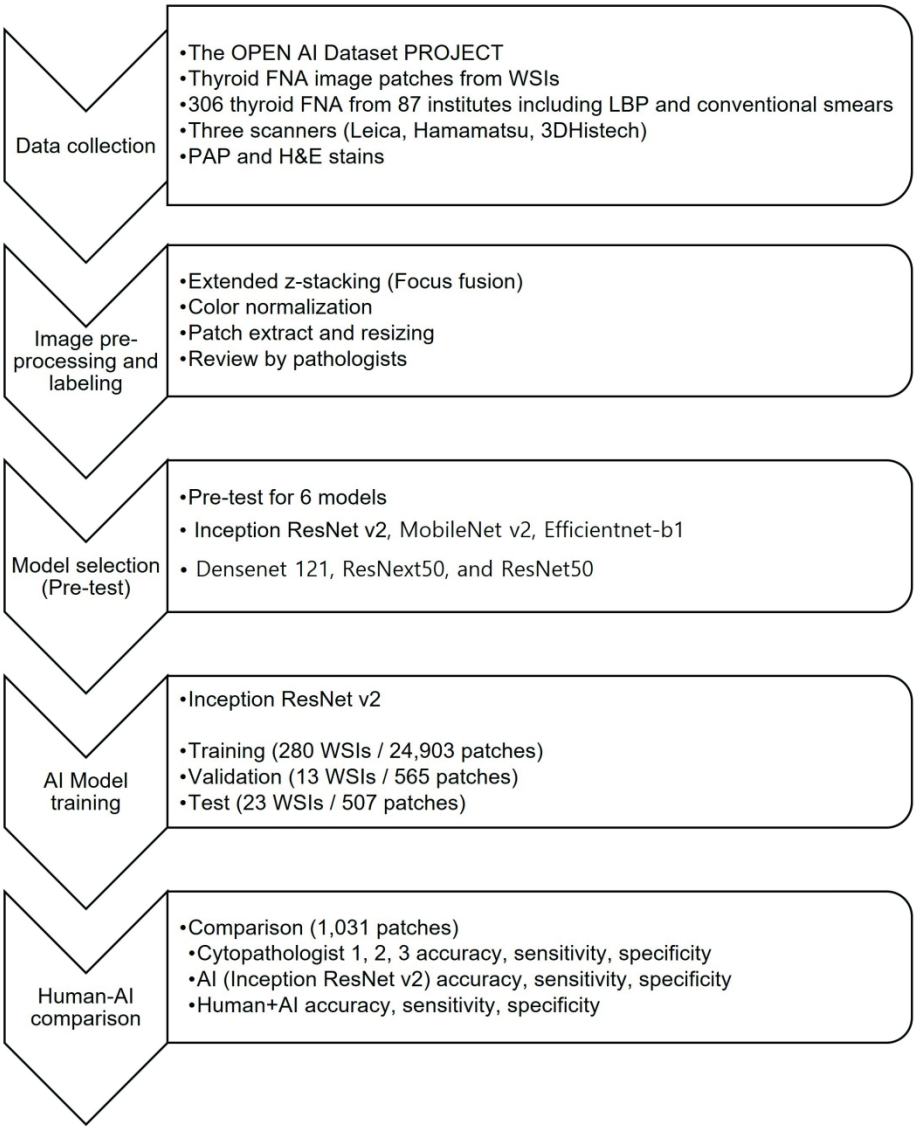


Fig. 1. Schematic workflow of this study.

177x216mm (330 x 330 DPI)

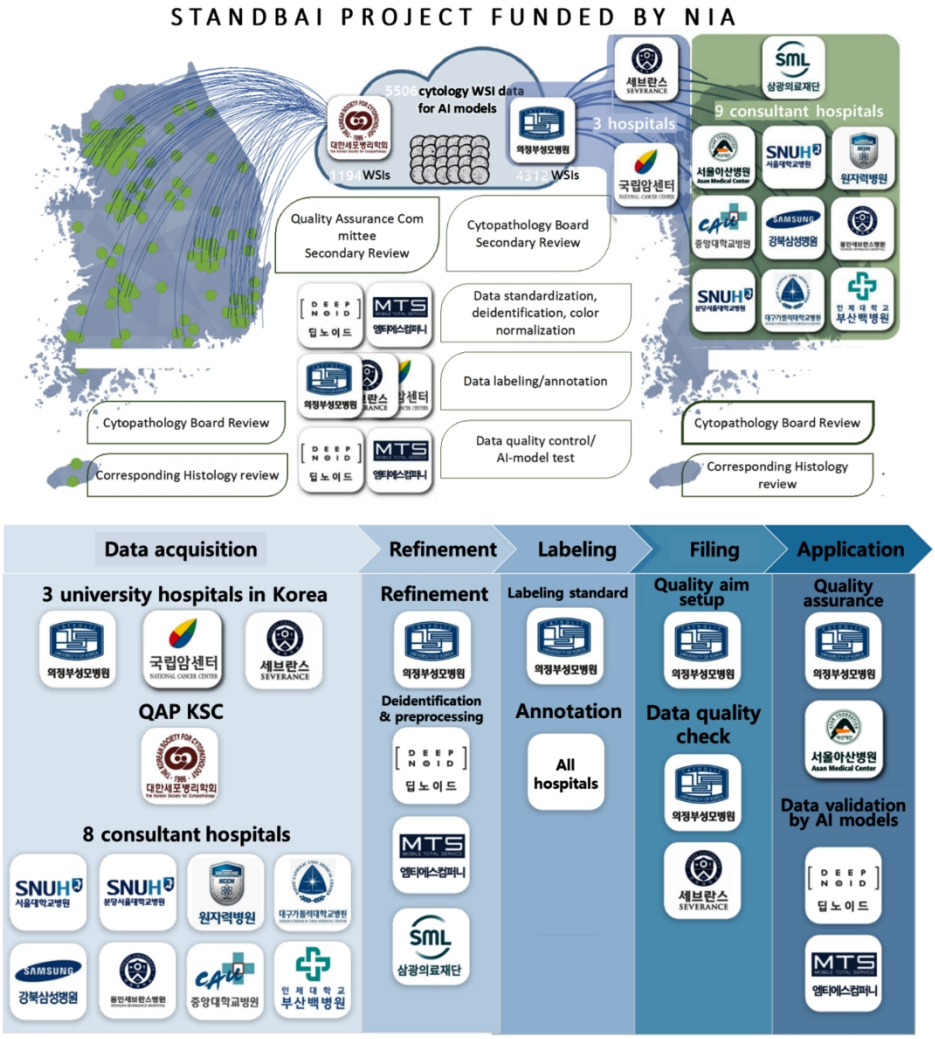


Fig. 2. Overview of the Open AI Dataset Project.

170x183mm (330 x 330 DPI)

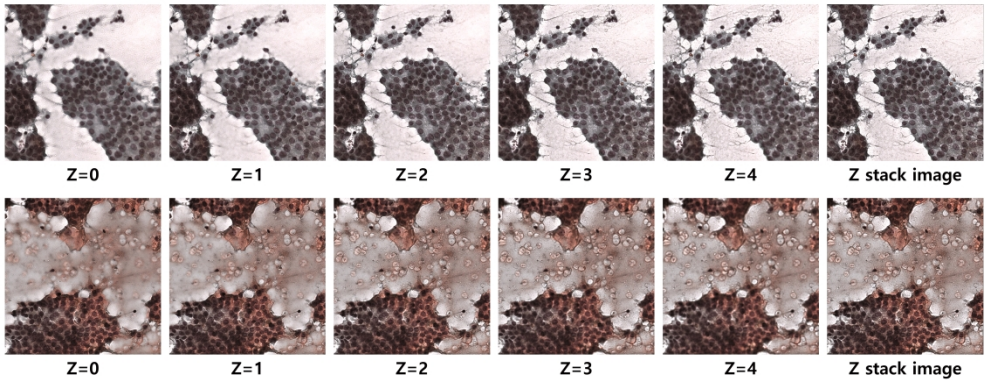


Fig. 3. Enhanced focus fusion processing to generate evenly perfectly focused 'extended z-stack images' for 3 dimensional cell clusters of the cytology samples from 3-6 layers of different z-axis focus.

415x162mm (330 x 330 DPI)

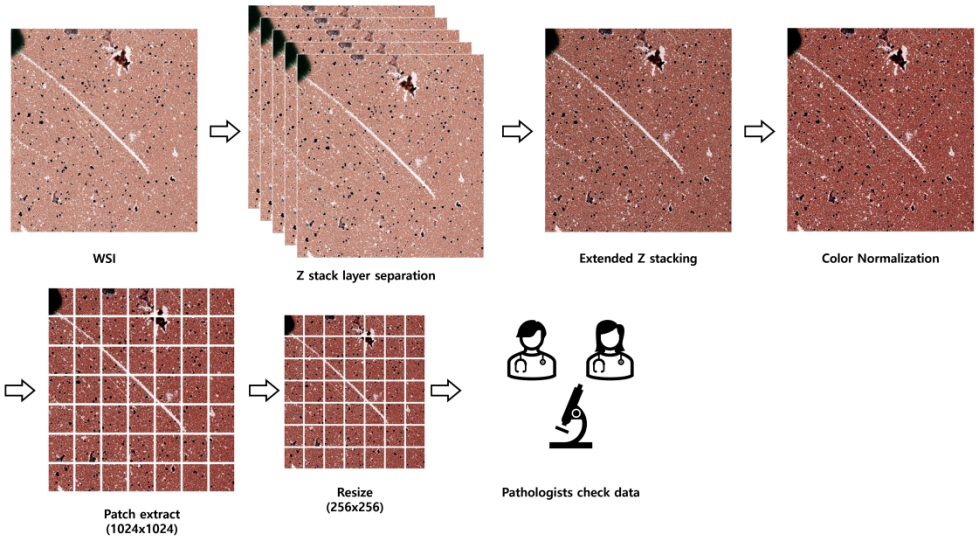


Fig. 4. Image pre-processing including focus fusion (extended z-stack image generation), color normalization, image patches extraction from WSIs, and resizing.

346x199mm (330 x 330 DPI)

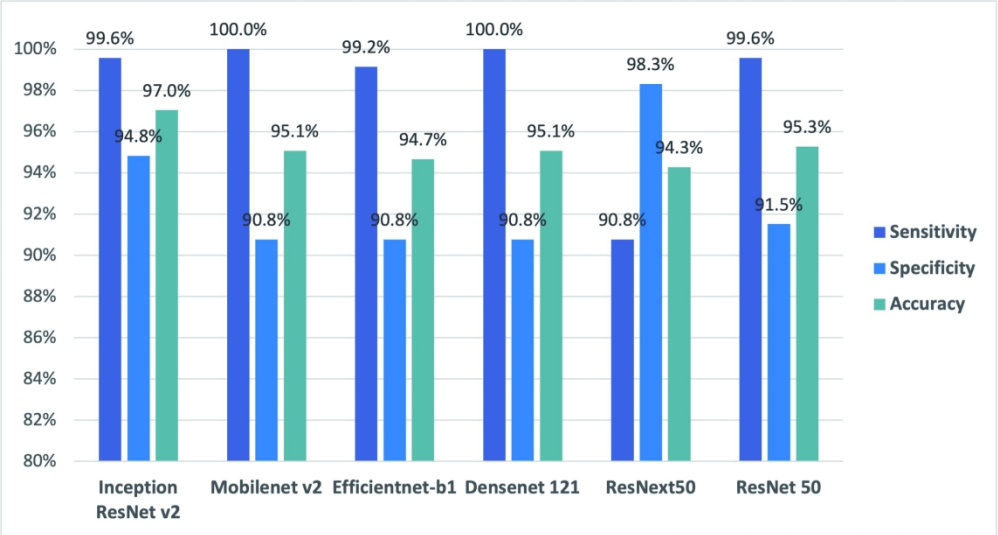


Fig. 5. Sensitivity, specificity, and accuracy of the convolutional neural network models in the pre-test to select the best model for benign/cancer image patch classification.

183x98mm (330 x 330 DPI)

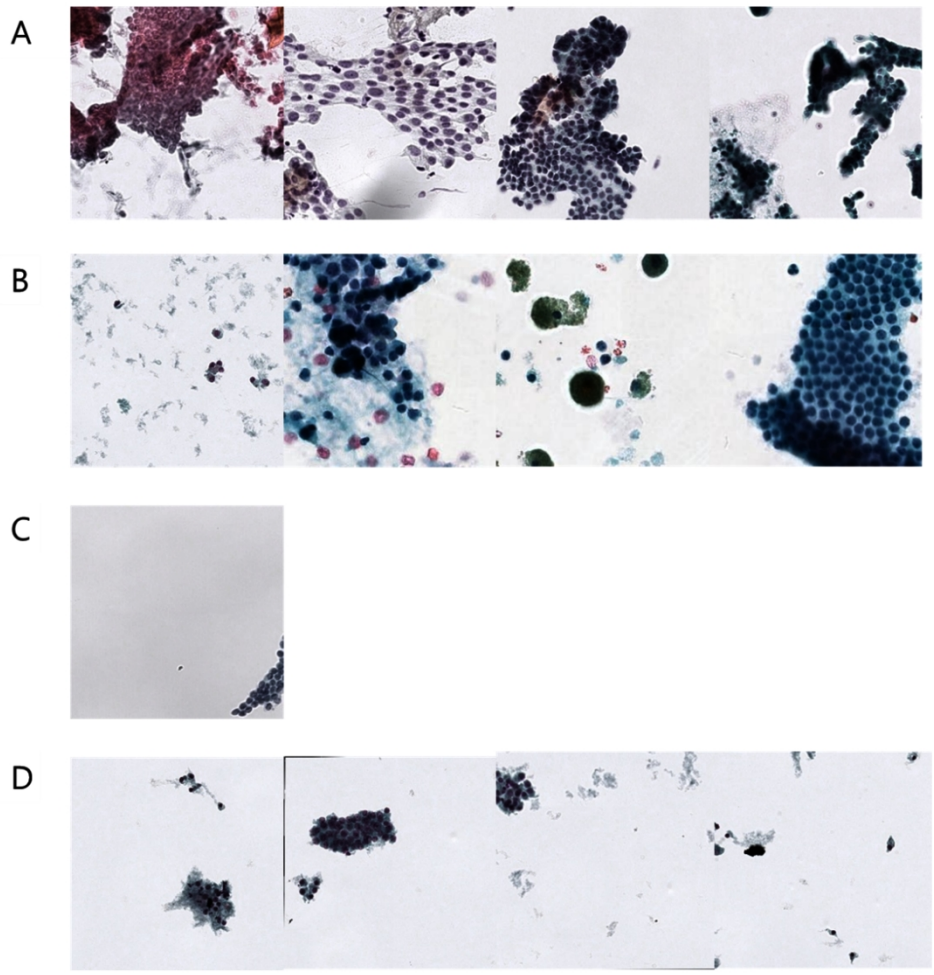


Fig. 6. Representative images of malignant (A) and benign (B) samples that were correctly diagnosed by the AI model; (C) malignant cells that were misdiagnosed as benign by the AI model; (D) clusters of benign follicular cells misdiagnosed as malignant.

177x178mm (330 x 330 DPI)

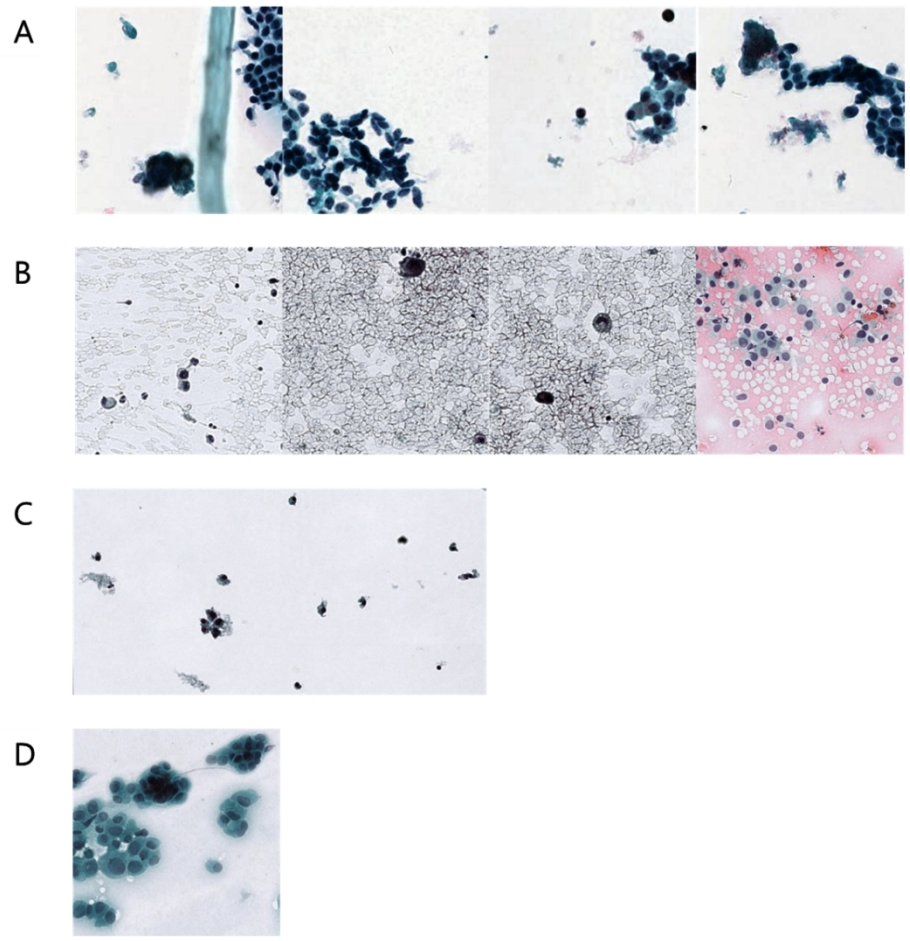


Fig. 7. Representative images classified differently by the cytopathologists and the AI model. AI correctly diagnosed benign (a) and malignant (b) samples that were misdiagnosed by all the cytopathologists; examples of benign (c) and malignant (d) smears correctly diagnosed by all the cytopathologists but misdiagnosed by the AI model.

189x186mm (330 x 330 DPI)

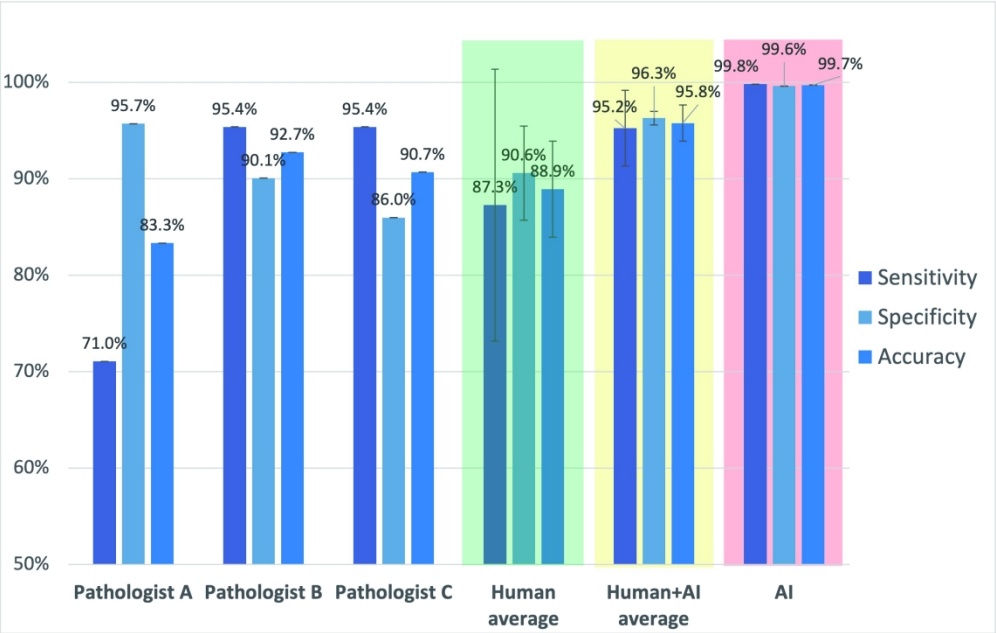


Fig. 8. Comparison of the performances between the human cytopathologists and AI model. In addition to displaying the sensitivity, specificity, and accuracy results for humans and their combination with AI, the figure also presents the standard deviation.

204x129mm (330 x 330 DPI)

editage

Editing Certificate

This document certifies that the manuscript listed below has been edited to ensure language and grammar accuracy and is error free in these aspects. The edit was performed by professional editors at Editage, a brand of Cactus Communications. The author's core research ideas were not altered in any way during the editing process. The quality of the edit has been guaranteed, with the assumption that our suggested changes have been accepted and the text has not been further altered without the knowledge of our editors.

MANUSCRIPT TITLE

Improved diagnostic accuracy of thyroid fine-needle aspiration cytology with artificial intelligence technology

AUTHORS

Yosep Chong

ISSUED ON

February 20, 2023

JOB CODE

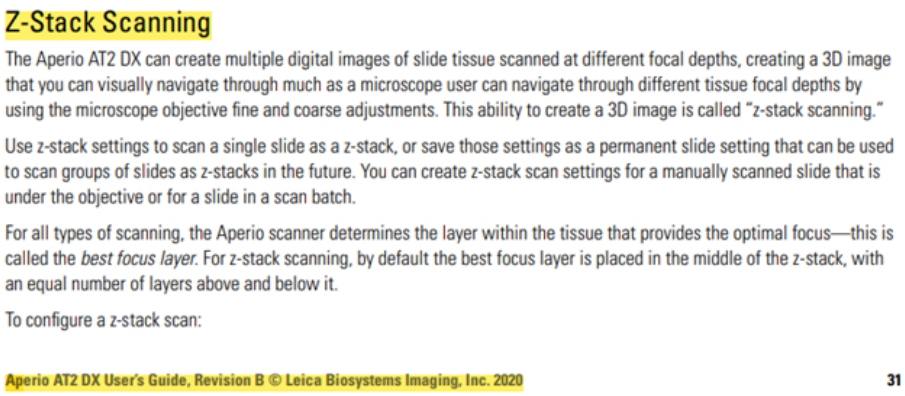
EPONG_22

THE EDITAGE Promise QUALITY GUARANTEE

Prabh Grewal
Senior Vice President - Editage

English Editing Certificate

167x178mm (330 x 330 DPI)



Z-Stack Scanning

The Aperio AT2 DX can create multiple digital images of slide tissue scanned at different focal depths, creating a 3D image that you can visually navigate through much as a microscope user can navigate through different tissue focal depths by using the microscope objective fine and coarse adjustments. This ability to create a 3D image is called "z-stack scanning."

Use z-stack settings to scan a single slide as a z-stack, or save those settings as a permanent slide setting that can be used to scan groups of slides as z-stacks in the future. You can create z-stack scan settings for a manually scanned slide that is under the objective or for a slide in a scan batch.

For all types of scanning, the Aperio scanner determines the layer within the tissue that provides the optimal focus—this is called the *best focus layer*. For z-stack scanning, by default the best focus layer is placed in the middle of the z-stack, with an equal number of layers above and below it.

To configure a z-stack scan:

Screenshot of the Leica AT2 scanner manual

146x62mm (120 x 120 DPI)