Deep learning models for thyroid nodules diagnosis of fine-needle aspiration biopsy: a retrospective, prospective, multicentre study in China

Jue Wang, Nafen Zheng*, Huan Wan*, Qinyue Yao*, Shijun Jia*, Xin Zhang*, Sha Fu, Jingliang Ruan, Gui He, Xulin Chen, Suiping Li, Rui Chen, Boan Lai, Jin Wang†, Qingping Jiang†, Nengtai Ouyang†, Yin Zhang†*

Summary

Background Accurately distinguishing between malignant and benign thyroid nodules through fine-needle aspiration cytopathology is crucial for appropriate therapeutic intervention. However, cytopathologic diagnosis is time consuming and hindered by the shortage of experienced cytopathologists. Reliable assistive tools could improve cytopathologic diagnosis efficiency and accuracy. We aimed to develop and test an artificial intelligence (AI)-assistive system for thyroid cytopathologic diagnosis according to the Thyroid Bethesda Reporting System.

Methods 11 254 whole-slide images (WSIs) from 4037 patients were used to train deep learning models. Among the selected WSIs, cell level was manually annotated by cytopathologists according to The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) guidelines of the second edition (2017 version). A retrospective dataset of 5638 WSIs of 2914 patients from four medical centres was used for validation. 469 patients were recruited for the prospective study of the performance of AI models and their 537 thyroid nodule samples were used. Cohorts for training and validation were enrolled between Jan 1, 2016, and Aug 1, 2022, and the prospective dataset was recruited between Aug 1, 2022, and Jan 1, 2023. The performance of our AI models was estimated as the area under the receiver operating characteristic (AUROC), sensitivity, specificity, accuracy, positive predictive value, and negative predictive value. The primary outcomes were the prediction sensitivity and specificity of the model to assist cyto-diagnosis of thyroid nodules.

Findings The AUROC of TBSRTC III+ (which distinguishes benign from TBSRTC classes III, IV, V, and VI) was 0·930 (95% CI 0·921–0·939) for Sun Yat-sen Memorial Hospital of Sun Yat-sen University (SYSMH) internal validation and 0·944 (0·929 – 0·959), 0·939 (0·924–0·955), 0·971 (0·938–1·000) for The First People's Hospital of Foshan (FPHF), Sichuan Cancer Hospital & Institute (SCHI), and The Third Affiliated Hospital of Guangzhou Medical University (TAHGMU) medical centres, respectively. The AUROC of TBSRTC V+ (which distinguishes benign from TBSRTC classes V and VI) was 0·990 (95% CI 0·986–0·995) for SYSMH internal validation and 0·988 (0·980–0·995), 0·965 (0·953–0·977), and 0·991 (0·972–1·000) for FPHF, SCHI, and TAHGMU medical centres, respectively. For the prospective study at SYSMH, the AUROC of TBSRTC III+ and TBSRTC V+ was 0·977 and 0·981, respectively. With the assistance of AI, the specificity of junior cytopathologists was boosted from 0·887 (95% CI 0·8440–0·922) to 0·993 (0·974–0·999) and the accuracy was improved from 0·877 (0·846–0·904) to 0·948 (0·926–0·965). 186 atypia of undetermined significance samples from 186 patients with *BRAF* mutation information were collected; 43 of them harbour the $BRAF^{V600E}$ mutation. 91% (39/43) of $BRAF^{V600E}$-positive atypia of undetermined significance samples were identified as malignant by the AI models.

Interpretation In this study, we developed an AI-assisted model named the Thyroid Patch-Oriented WSI Ensemble Recognition (ThyroPower) system, which facilitates rapid and robust cyto-diagnosis of thyroid nodules, potentially enhancing the diagnostic capabilities of cytopathologists. Moreover, it serves as a potential solution to mitigate the scarcity of cytopathologists.

Funding Guangdong Science and Technology Department.

Lancet Digit Health 2024; 6: e458–69

Published Online June 6, 2024 https://doi.org/10.1016/S2589-7500(24)00085-2

*Contributed equally

†Joint last authors

For the Chinese translation of the abstract see Online for appendix 1

Department of Cellular and Molecular Diagnostics Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China (Ju Wang MS, N Zheng BSc, H Wan BSc, S Fu MD, G He BSc, Prof N Ouyang MD, Y Zhang PhD); **Cells Vision (Guangzhou) Medical Technology, Guangzhou, China** (Q Yao MS, X Chen MS, S Li MS, R Chen BSc, Ji Wang PhD); **Department of Pathology, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu, China** (S Jia MS); **Department of Pathology, The First People's Hospital of Foshan, Foshan, China** (X Zhang MS); **Department of Ultrasound, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China** (J Ruan MD); **Department of Pathology, The Third Affiliated Hospital, Guangzhou Medical University, Guangzhou, China** (B Lai BSc, Prof Q Jiang MD); **Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China** (Prof N Ouyang, Y Zhang)

Introduction

The high prevalence of thyroid nodules is a global health issue.[1] High-resolution ultrasound can detect thyroid nodules in up to 68% of randomly selected individuals.[2] About 10–15% of thyroid nodules are proven to be thyroid cancer,[3] which is one of the most frequently diagnosed endocrine malignancies.[4] Given the high prevalence of thyroid nodules and the low percentage of malignant cases, accurately distinguishing between benign and malignant thyroid nodules is crucial.

Fine-needle aspiration (FNA) cytology is reported to be the most precise single test, which provides a clear

Correspondence to:
Dr Yin Zhang, Department of Cellular and Molecular Diagnostics Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510030, China
zhangy525@mail.sysu.edu.cn

Research in context

Evidence before this study

Fine-needle aspiration (FNA) cytopathology diagnosis is important for thyroid nodule management and is recommended by various clinical practice guidelines. However, the widespread usage of FNA cytopathology for prompt diagnosis of thyroid nodules is hindered by the national shortage of cytopathologists in China. Artificial intelligence (AI) could offer a promising solution to this problem. We searched PubMed and Google Scholar for publications about artificial intelligence in thyroid cytopathology between Jan 1, 2010, and March 18, 2024. We used the keywords "artificial intelligence/deep learning", "thyroid nodules/cancer", "cytopathology/pathology", and "diagnosis". We identified ten studies with sample sizes ranging between 43 and 806, and all of these studies were single-centred. Seven of these studies only performed a simple binary classification (only distinguishing benign and malignant); another study explored multi-classification in 393 samples but did not follow The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) guidelines, a standardised, category-based reporting system that is adopted worldwide. The literature search results showed that high performance and reliable AI models trained and validated on large, multicentric datasets that follow the TBSRTC clinical guidelines are still absent.

Added value of this study

Many studies have shown that larger datasets lead to better classification performance, and multicentric validation is critical to reduce the risks of biased prediction by AI tools. In this study, we developed and validated an AI system for assisted diagnostics of thyroid nodules with a total of 17 966 whole-slide images of 8426 smears from 7420 patients with thyroid nodules from four centres in China. Additional, a prospective validation set was included to assess the robustness of the model. To the best of our knowledge, this is the largest thyroid FNA dataset with experienced cytopathologist supervision that followed the TBSRTC guidelines.

Implications of all the available evidence

With innovative deep learning models and a large dataset, our system had high performance on both internal and external validation datasets. We offer an AI-assisted system that has potential to be applied in clinical practice for the accurate and efficient diagnosis of thyroid nodules.

diagnosis of benign or malignant thyroid disease in most cases.[5] FNA cytology diagnosis of thyroid nodules with suspected sonographic patterns is recommended by various clinical practice guidelines.[6–8] Thyroid nodule FNA cytology should be reported using the diagnostic groups outlined in The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC), a standardised, category-based reporting system that is adopted worldwide.[9] According to the second version of TBSRTC, each thyroid FNA report should begin with one of six diagnostic categories.[9] Each category has an implied cancer risk that ranges from 0–3% to 97–99% for the benign to malignant categories. However, cytological diagnosis requires experience but is also labour intensive and time consuming. There is a shortage of cytopathologists and pathologists in China, which not only increases diagnostic waiting times but also hinders the widespread usage of FNA cytopathology for thyroid nodule diagnosis.[10,11] Therefore, robust assistive tools are urgently needed to improve FNA diagnosis efficiency to accommodate growing patient needs.

About 15–30% of thyroid nodules are classified as indeterminate nodules, between benign and malignant, including 7–10% of AUS samples.[12] Uncertain diagnosis can cause excessive, uncontrollable worry in patients.[13,14] Many patients with indeterminate cytological features will have diagnostic thyroid surgery, which might lead to severe surgical complications, such as damage to the laryngeal nerve.[15] Increasing diagnostic certainty is an essential strategy for reducing overtreatment, a significant problem for thyroid nodule management.[16–18] Some molecular markers have high diagnostic precision, serving as a powerful method to assist thyroid cancer diagnosis, particularly for 'rule-in' tests of indeterminate cytopathologic results.[19] For example, *BRAF*^{V600E} mutations are significant indicators of malignancy in thyroid nodules and can crucially influence diagnostic and therapeutic decisions. If a *BRAF*^{V600E} mutation is found in a thyroid nodule, the risk of papillary thyroid carcinoma is nearly 100%.[20] However, gene detection is not feasible for all patients for many reasons, such as cell insufficiency of the sample, unaffordability, and unavailability of the test.[21]

Over the past 5 years, advances in deep-learning technology have provided emerging opportunities for artificial intelligence (AI) applications in cytology diagnosis, particularly for thyroid nodules. Computer-aided diagnosis systems effectively reduce diagnostic time and improve accuracy by offering an exhaustive and detailed scan of every cell, which would ease the workload for experienced cytopathologists and increase diagnosis efficiency. Many models have been developed to help screen and diagnose different histological and cytological samples.[22–24] Some studies have applied deep learning technology to cytological digital images for thyroid cancer diagnosis.[25–28] Although the results are encouraging, the two major limitations of current studies are: (1) only performing a simple binary classification (benign *vs* malignant); and (2) only being based on a small number of cytological FNA images from a few hundred patients.[26] In real-world clinical practice, closely following the TBSRTC category guidelines and making an AI model

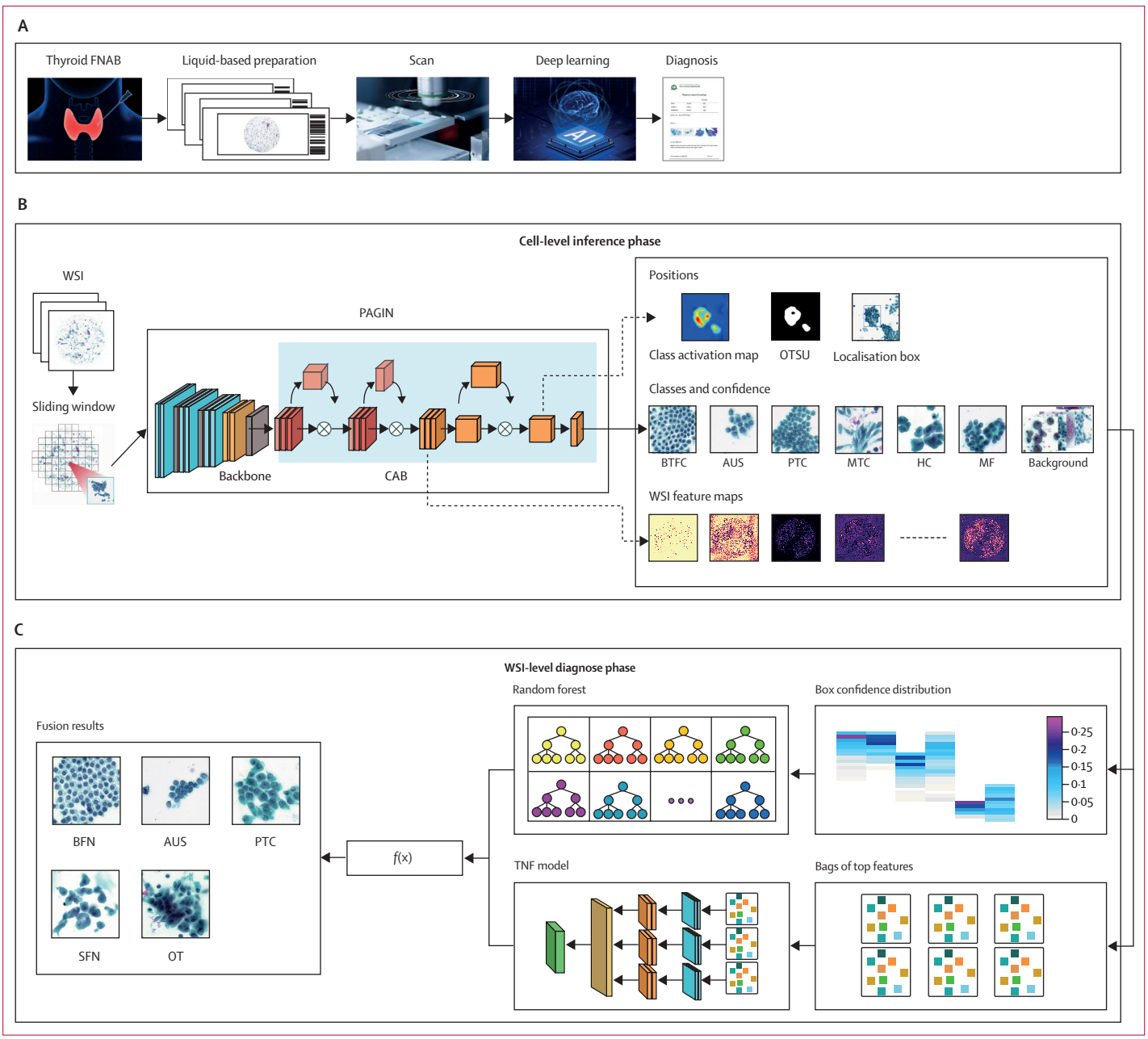


Figure 1: **Workflow of the ThyroPower system**

(A) Overview of the whole processes of AI-assisted diagnosis of thyroid FNA. (B) The ThyroPower automates the thyroid FNA screening process using deep learning algorithms. The system provides a WSI-level diagnostic decision, with the abnormal areas identified. After the slide preparation and scanning process, a patch-level model called PAGIN was applied to extract cell-level features and locate abnormalities by sliding a fixed-size window of size 768×768 pixels throughout the WSI. (C) The WSI-level classification models of the ThyroPower system are trained on a large and well-annotated dataset to distinguish different categories. Bounding boxes of abnormal cells were generated using OTSU thresholding on the class activation maps. Cell-level features include cell classes, confidences, and feature maps. After the feature extraction, two sub-models were employed for WSI-level diagnosis, utilising different perspectives of the extracted features. One is a random forest model based on customised handcrafted features. The other is a multi-instance-learning model proposed by this study using deep learning feature maps. Bags of top features refers to selecting the top patches with the highest confidence for each class and using their features for model classification. The final WSI-level diagnostic decision was produced by the fusion of the two sub-models, which imitates when the diagnosis is made by multiple experts together. The descriptions of different categories are listed in appendix 2 (pp 9–10). AI=artificial intelligence. AUS=atypia of undetermined significance. BFN=benign follicular nodule. BTFC=benign thyroid follicular cell. CAB= convolutional attention block. CC=cyst-lining cell. FC=fuzzy cell. FNA=fine-needle aspiration. FNAB=fine-needle aspiration biopsy. HC=Hürthle cell. LC=lymphocyte. MF=microfollicle formation. MGC=multinucleated giant cell. MP=macrophages. MTC=medullary thyroid carcinoma. OT=other tumours. OTSU=Otsu's method. PAGIN=Patch Attention Global Importance Pooling Network. PTC=papillary thyroid carcinoma. RBC=red blood cell. SFN=suspicious for a follicular neoplasm. SPTC=suspicious papillary thyroid carcinoma. TBSRTC=The Bethesda System for Reporting Thyroid Cytopathology. TNF=Top N Feature. UTC=undifferentiated carcinoma. WSI=whole-slide image.

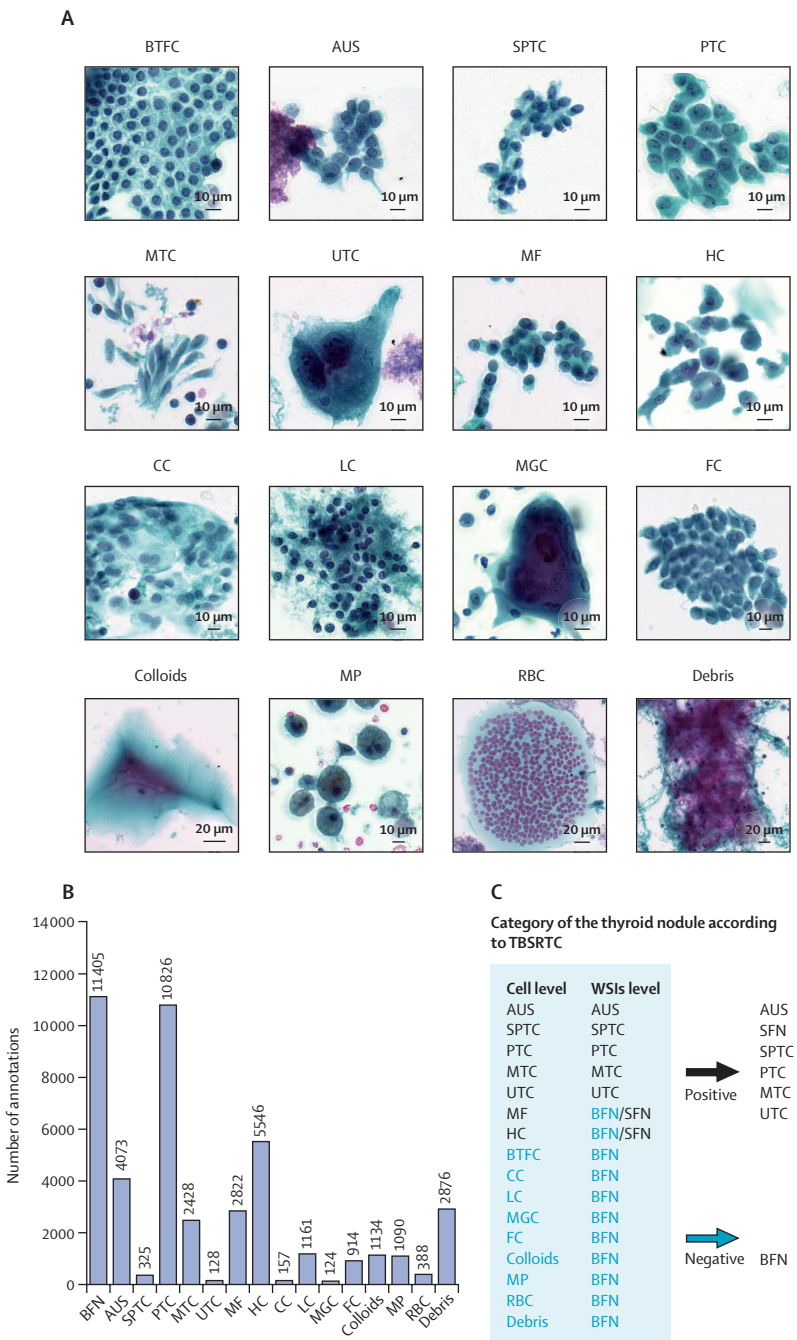


Figure 2: **Preparation of high-quality annotations of different categories of thyroid cytology for AI models**
(A) Representative images of 16 categories of thyroid cytology according to the TBSRTC guidelines. The cells or components are stained using the sedimentation-based Papanicolaou staining method. (B) Annotation number of each category used for AI model development. (C) Category of the thyroid nodule at the cell level and WSI level according to the TBSRTC guidelines. The descriptions of different categories are listed in appendix 2 (pp 9–10). AUS=atypia of undetermined significance. BFN=benign follicular nodule. BTFC=benign thyroid follicular cell. CC=cyst-lining cell. FC=fuzzy cell. HC=Hürthle cell. LC=lymphocyte. MF=microfollicle formation. MGC=multinucleated giant cell. MP=macrophages. MTC=medullary thyroid carcinoma. PTC=papillary thyroid carcinoma. RBC=red blood cell. SFN=suspicious for a follicular neoplasm. SPTC=suspicious papillary thyroid carcinoma. TBSRTC=The Bethesda System for Reporting Thyroid Cytopathology. UTC=undifferentiated carcinoma. WSI=whole-slide image.

with more reliable data are the most meaningful and challenging compared with conventional diagnostic methods.

In this study, we developed an AI-assisted model named the Thyroid Patch-Oriented WSI Ensemble Recognition (ThyroPower) system, to facilitate rapid and robust cyto-diagnosis of thyroid nodules. The model was trained on a large number of samples with high-quality annotations according to TBSRTC guidelines[8] for thyroid cytopathology. The performance was evaluated using retrospective multicentric datasets and a real-world prospective dataset. Moreover, we explored the potential of the ThyroPower system to aid the diagnosis of atypia of undetermined significance (AUS) samples by examining the consistency between *BRAF*^{V600E} mutation status and the AI-predicted confidence level.

Methods

Data acquisition and dataset description

The development workflow of the ThyroPower system is shown in figure 1. First, we collected digital whole-slide images (WSIs) of liquid-based, thin-layer smears of human samples obtained via thyroid FNA biopsy from a total of 7420 patients, which comprised 8426 smears. 17 966 WSIs were collected. All of the patients were enrolled between Jan 1, 2016, and Jan 1, 2023, in the participating hospitals in this study. These samples were diagnosed on the basis of TBSRTC guidelines by senior cytopathologists. The six diagnostic categories of TBSRTC include: non-diagnostic or unsatisfactory (TBSRTC I); benign follicular nodules (BFN, TBSRTC II); AUS (TBSRTC III); suspicious for a follicular neoplasm (SFN, TBSRTC IV); suspicious papillary thyroid carcinoma (TBSRTC V); and malignant (TBSRTC VI), for which the malignancy includes papillary thyroid carcinoma, medullary thyroid carcinoma, and undifferentiated carcinoma. The definitions of the cell-level classification and WSI-level classification, which correspond to TBSRTC classification, are detailed in figure 2 and appendix 2 (pp 9–10).

Data were organised into six distinct datasets. One of these datasets was utilised as the training set, four served as validation sets, and one was used as a prospective validation set. Data in the training set were collected from Sun Yat-sen Memorial Hospital of Sun Yat-sen University (SYSMH). The internal validation set came from SYSMH, and three sets of external validation data came from Sichuan Cancer Hospital & Institute (SCHI), The First People's Hospital of Foshan (FPHF), and The Third Affiliated Hospital of Guangzhou Medical University (TAHGMU). The patients' information for each dataset is listed in the table. Detailed information on enrolment is described in appendix 2 (pp 2–3, 17).

Molecular testing is an ideal complementary method to diagnosing AUS samples. Specifically, *BRAF* testing, a reliable rule-in marker, significantly enhances the

	Total	SYSMH training set	SYSMH internal validation set	SCHI external validation set	FPHF external validation set	TAHGMU external validation set	SYSMH prospective set
Patients	7420 (100%)	4037 (54·4%)	1255 (17·0%)	746 (10·0%)	721 (9·7%)	192 (2·6%)	469 (6·3%)
Samples	8426 (100%)	4512 (53·5%)	1560 (18·5%)	816 (9·7%)	798 (9·5%)	203 (2·4%)	537 (6·4%)
WSIs	17 966 (100%)	11 254 (62·6%)	3821 (21·3%)	816 (4·6%)	798 (4·4%)	203 (1·1%)	1074 (6·0%)
Sex							
Male	1551 (20·7%)	831 (20·6%)	257 (20·5%)	162 (21·7%)	145 (20·1%)	43 (22·4%)	113 (21·0%)
Female	5937 (79·3%)	3206 (79·4%)	998 (79·5%)	584 (78·3%)	576 (79·9%)	149 (77·6%)	424 (79·0%)
Age of patients, years							
Mean	44	44	44	43	43	46	44
Range	4–85	4–82	8–81	14–80	13–87	22–85	17–79
TBSRTC categories							
Benign follicular nodules	4175 (49·4%)	2450 (54·3%)	808 (51·8%)	236 (28·9%)	255 (31·9%)	151 (74·4%)	275 (51·2%)
Atypical of undetermined significance	741 (8·9%)	372 (8·2%)	217 (13·9%)	63 (7·7%)	75 (9·4%)	3 (1·5%)	11 (2·1%)
Suspicious for a follicular neoplasm	122 (1·4%)	76 (1·7%)	23 (1·5%)	3 (0·4%)	14 (1·8%)	0	6 (1·1%)
Suspicious papillary thyroid carcinoma	395 (4·8%)	56 (1·2%)	17 (1·1%)	81 (9·9%)	225 (28·2%)	1 (0·5%)	15 (2·8%)
Papillary thyroid carcinoma	2933 (34·8%)	1519 (33·7%)	484 (31·0%)	428 (52·5%)	229 (28·7%)	47 (23·1%)	226 (42·0%)
Medullary thyroid carcinoma	47 (0·5%)	30 (0·7%)	10 (0·6%)	4 (0·5%)	0	1 (0·5%)	2 (0·4%)
Undifferentiated carcinoma	13 (0·2%)	9 (0·2%)	1 (0·1%)	1 (0·1%)	0	0	2 (0·4%)

Data are n (%) unless otherwise specified. FPHF=The First People's Hospital of Foshan. SCHI=Sichuan Cancer Hospital & Institute. SYSMH=Sun Yat-sen University. TAHGMU=The Third Affiliated Hospital of Guangzhou Medical University. TBSRTC=The Bethesda System for Reporting Thyroid Cytopathology. WSI=whole-slide image.

Table: **Baseline demographic and clinical characteristics of cohorts in this study**

diagnostic accuracy for these samples. Moreover, most *BRAF* testing samples are taken from the same matched FNA samples used for slide preparation and cytopathologic diagnosis, which eliminates the bias caused by the sampling procedure and makes them ideal for our integrated analysis. From the SYSMH training dataset, 186 AUS smears were collected, each with *BRAF* mutation information, of which 43 (23%) of 186 were found to carry the $BRAF^{V600E}$. The $BRAF^{V600E}$ mutation was determined by targeted DNA sequencing or amplification refractory mutation system PCR (AmoyDx Diagnostics, Xiamen, China).

The retrospective and prospective aspects of this study were approved by the Sun Yat-sen Memorial Hospital Institutional Review Board (number SYSKY-2022–241–01). All patients participating in the study provided informed consent. For those enrolled in the prospective study, written consent was obtained, while for the retrospective study, a waiver of consent was signed when applicable.

Quality control

We set up a standard procedure to control the image quality based on the TBSRTC. The participants self-reported that they were not pregnant, had no mental illness or cognitive impairment, and consented to thyroid liquid-based cytology for a definite diagnosis. Ethnicity data were unavailable. There was no age limit for eligible participants. Participants had a thyroid liquid-based preparation, and the smears were collected from hospitals. According to TBSRTC,[8] the criterion for satisfactory smears is that the FNA sample was located in the thyroid gland. The scanned smears should be focused and clear, and samples should be appropriately given the diagnosis category TBSRTC II, III, IV, V, or VI. The criterion for unsatisfactory smears was the diagnosis category TBSRTC I, and a smear was considered to be unsatisfactory by the evaluating cytopathologist if the scan was blurry.

For all training, validation, and test sets at the slide level, the labels were carefully reviewed by the senior cytopathologists panel. However, to ensure the independence of the external validation data, the labels for the three external validation sets were sourced directly from the pathology reports of the respective centres. All diagnoses from external centres also followed the TBSRTC guidelines. Senior cytopathologists (JuW, NZ, and HW) were defined as those who had obtained their professional pathology practice certificates, had more than 5 years of work experience, and on average, reviewed and reported on over 5000 cases annually. However, junior cytopathologists were those who have also obtained their professional pathology practice certificates but had less than 5 years of work experience, and had reviewed and reported on over 2000 cases annually on average. The "ground truth" refers to the initial reference to the histological diagnosis of the corresponding thyroid biopsy of the smear. In cases where smears did not have histological results, they were reviewed by three senior cytopathologists and diagnosed according to the TBSRTC standard. The diagnosis was based on the consensus of the three senior cytopathologists. If a consensus could not be reached, the smear was reviewed by a fourth expert cytopathologist with more than 10 years of experience.

See Online for appendix 2

Annotation strategy

In our labelling process, we considered annotations at both the cellular and WSI levels. WSI-level labels for training, validation, and test sets were carefully reviewed by our expert panel (NO, QJ, and SF) to ensure accuracy. However, for external validation data, labels were sourced directly from pathology reports to maintain independence. At cellular level, we used an AI-assisted annotation strategy to speed up the annotation process and increase data efficiency. The AI-assisted annotation strategy process is shown in appendix 2 (p 19). Initially, a subset of boxes in a select number of WSIs were annotated by senior cytopathologists. These classification models were trained on the basis of these annotations. These classification models were used for generating an AI-recommended region of interest (ROI). Any ROI that was classified as malignant or suspicious for malignancy in benign WSIs was unquestionably a false positive patch. Therefore, they were directly used as benign labels for training. This approach was inspired by the hard negative mining technique, which initially served in object detection tasks to prevent the model from being overwhelmed by easy samples.[29] The rest of the ROIs on WSIs in the cell-level classification dataset were sent to cytopathologists for confirmation and correction.

To summarise, this strategy generated three types of labels: (1) labels independently annotated by experts; (2) AI-recommended ROIs, which were then confirmed or corrected by medical experts; and (3) AI-generated benign labels for benign slides. The continuous process of expert confirmation and correction of the AI-suggested labels effectively created a continuous error correction loop for the AI. This strategy enhances the collection of samples that yield the most valuable insights, essentially working towards maximising learning while using fewer labelled examples, and is aligned with the concepts of "Human-in-the-loop" and "Active Learning".[30,31]

Model development

We developed an AI-assisted system to support cytopathologists in diagnosing thyroid nodules based on FNA WSIs. The system is a multi-category classifier with multi-stage feature extraction and a multi-model fusion approach. The training process of the ThyroPower system consisted of three stages: (1) training the Patch Attention Global Importance Pooling Network (PAGIN) for extracting cell-level features and locating abnormalities; (2) applying the trained PAGIN model to WSIs to obtain WSI-level feature representations; and (3) training WSI-level classifiers and fusing them together to make the final diagnostic decision. Other detailed procedures for AI development and optimisation are described in appendix 2 (pp 2–8).

Outcomes

The primary outcomes were the prediction sensitivity and specificity of the model in assisting pathologists in the cyto-diagnosis of thyroid nodules. Sensitivity was measured by the model's ability to correctly identify positive cases (true positives), and specificity was assessed by its ability to correctly identify negative cases (true negatives). Secondary outcomes included the consistency between the model's cytological diagnosis and postoperative histopathology results, as well as its effectiveness in identifying malignancy in cytological AUS samples.

Statistical analysis

To evaluate our model's performance, we use the metrics sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic (AUROC). The AUROCs for cell-level classification were computed using the one-versus-rest strategy, in which the performance of each class was compared to the rest. Specifically, a multi-class classification problem was separated into a series of binary classification problems, and it evaluated the model's ability to distinguish between each class and the rest of the classes and averaged to obtain the final AUROC. Moreover, we present micro and macro AUROC scores to capture different aspects of model performance. The micro AUROC aggregated the outcomes of all classes to compute a single measure, whereas the macro AUROC calculated the average of the AUROC scores for each class, treating all classes equally regardless of their size. In addition, we calculated Cohen's κ values to quantify the agreement between ThyroPower and senior cytopathologists.

To compute the AUROC and ROC curves for WSI-level classification, the non-benign probability threshold, 1−P (BFN), was varied, where P (BFN) denotes the probability of the BFN class. The AUROC for TBSRTC III+, including TBSRTC classes III, IV, V, and VI, was calculated, as was the AUROC for TBSRTC V+, which includes TBSRTC classes V, and VI. The latter metric was introduced to better show ThyroPower's ability to judge malignant tumours, as TBSRTC III and IV were classified with indeterminate diagnoses and had poor repeatability. The risk of malignancy for TBSRTC V+ was over 50%, whereas the risk for TBSRTC III and TBSRTC IV ranged between 10–30% and 25–40%,[32–34] respectively.

Model diagnosis decisions were based on the probability outputs from the softmax function, which converts logits to probabilities, implemented in TensorFlow. The class with the highest probability was classified as the final class for WSI-level diagnosis. However, for identifying malignancies within the cytological AUS samples, we selected the optimal threshold value by maximising the F2-score. Based on the F2-score versus the threshold curve calculated on the internal validation cohort, the chosen threshold value was 0·083, and the corresponding optimal F2-score was 0·847.

To validate the effectiveness of our fusion model in comparison to its individual sub-models, Random Forest (RF) and Top-N Feature (TNF), we used two statistical

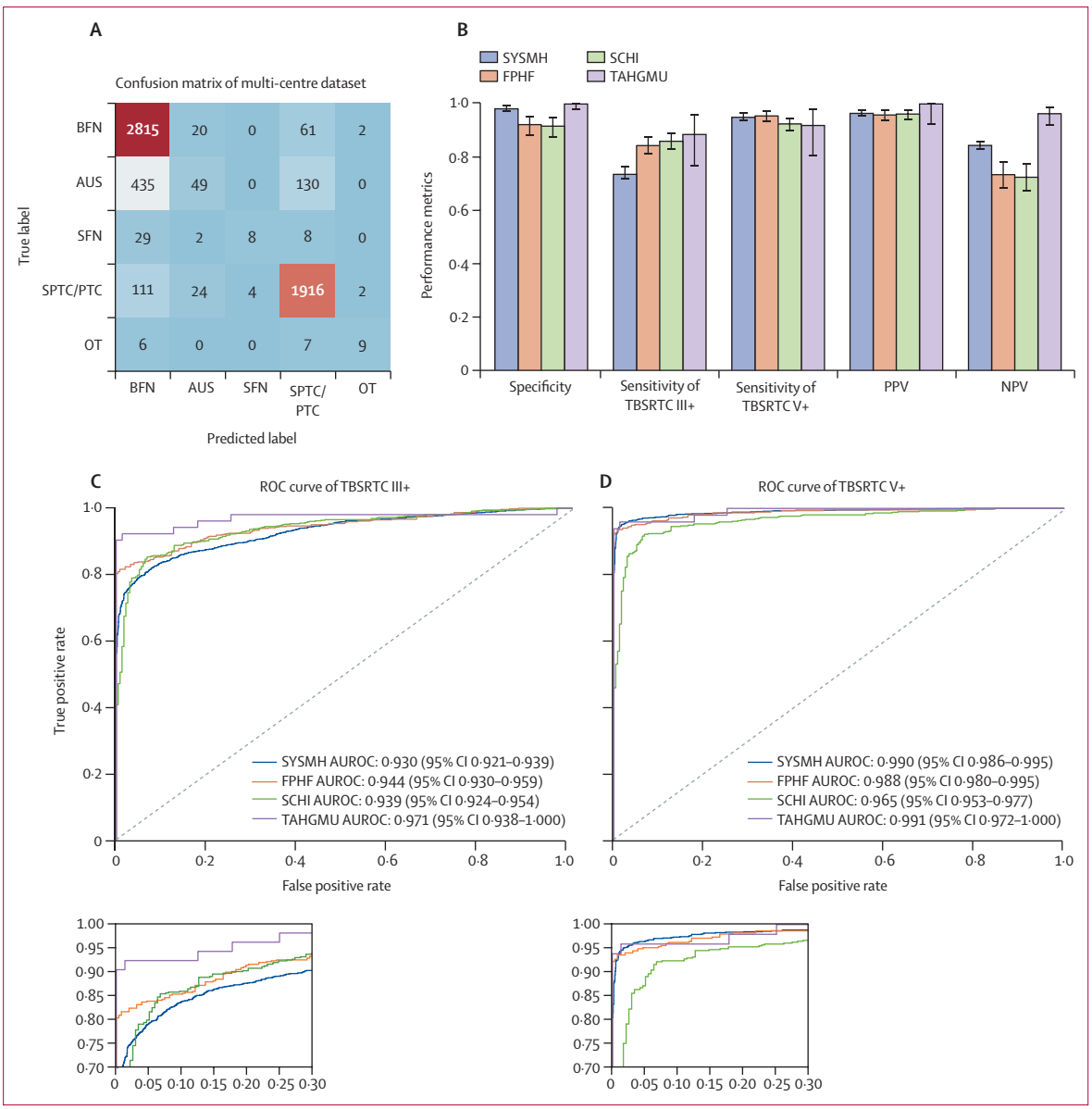


Figure 3: **Performance of the ThyroPower in the multi-centre dataset**
(A) Confusion matrix showing the detailed results of classification by AI, the summed numbers of the four centres are shown. Separate confusion matrixes of each centre are shown in appendix 2 (p 21). The heatmap colours of the grids represent the number of samples. The colours range from blue to red, with red indicating a larger number of samples and blue indicating a lower number of samples. (B) The performance of the ThyroPower on validation datasets from different centres. The specificity and sensitivity of TBSRTC III and TBSRTC V+, PPV, and NPV are shown; error bars are 95% CIs. The descriptions of different categories are listed in appendix 2 (pp 9–10). (C) AUROC curve showing the performance of ThyroPower in prediction of TBSRTC III plus samples from different medical centres. (D) AUROC curve showing the performance of ThyroPower in prediction of TBSRTC V plus samples from different medical centres. AUROC=area under the receiver operating characteristic. AI=artificial intelligence. AUS=atypia of undetermined significance. BFN=benign follicular nodule. FPHF=The First People's Hospital of Foshan. NPV=negative predictive value. OT=other tumours. PPV=positive predictive value. PTC=papillary thyroid carcinoma. SFN=suspicious for a follicular neoplasm. SCHI=Sichuan Cancer Hospital & Institute. SPTC=suspicious papillary thyroid carcinoma. SYSMH=Sun Yat-sen University. TAHGMU=The Third Affiliated Hospital of Guangzhou Medical University. TBSRTC=The Bethesda System for Reporting Thyroid Cytopathology.

tests. The DeLong test,[35] a statistical analysis designed for evaluating the differences between two AUROCs derived from the same set of individuals, was used to compare the AUROC curves of the models. The McNemar's test[36] was applied to assess differences in sensitivity and specificity between the models.

In addition, we computed 95% CIs for all performance metrics to provide a precise measure of the metrics' estimated uncertainty. We used Matplotlib (version 3.5.1), R (version 4.2.1), and Seaborn (version 0.11.2) to generate all of the plots. The study followed the reporting and analysis guidelines of STARD.[37]

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

We developed AI models that imitated the diagnosis process of a cytopathologist; the models scanned the WSIs with our proposed deep learning framework PAGIN to extract useful diagnostic features (figure 1A, B), and then produced a diagnostic decision based on this information using a fusion model of two WSI-level classifiers (figure 1C).

During the development of the PAGIN model, we initially evaluated and compared the performance of different backbone networks (ie, various foundational architectures used within the PAGIN model for feature extraction): EfficientNet-b0,[38] ResNet50,[39] Incepcent-V3,[40,41] and VGG16,[42,43] all of which are top-performance networks in the computer vision field.[44] Our results showed EfficientNet-b0 had the best performance and efficiency (appendix 2 p 11), with an overall accuracy of 0·707 and a macro-average AUROC of 0·924. We further proposed a convolutional attention block module that plugged behind the model backbone. By adding the convolutional attention block module, the macro-average AUROC was improved from 0·924 to 0·936 (appendix 2 p 20).

The GradCAM in appendix 2 (p 18) shows that, in PAGIN, shallow layers tend to capture edges and very detailed pieces of information, deeper layers focus on more complicated features, and the deepest layers accurately concentrate on cells and clusters that contribute the most to the classification decisions.

To reach WSI-level classification, the ThyroPower system used a fusion of two sub-models: a classical machine learning model, RF, with customised hand-crafted statistical features, and a novel multi-instance-learning model referred to as the TNF model. The result showed that the fusion model has the highest AUROC score and significantly higher specificity and PPV simultaneously (appendix 2 pp 12, 20).

We assessed ThyroPower performance on TBSRTC classification across four medical centers (appendix 2 pp 2–3). ThyroPower showed good consistency in classification of different cell types with senior cytopathologists (figure 3A, B). We evaluated the performance on distinguishing benign from TBSRTC III+ and the AUROC was 0·930 (95% CI 0·921–0·939) for SYSMH internal validation and 0·944 (0·929–0·959), 0·939 (0·924–0·955), and 0·971 (0·938–1·000) for FPHF, SCHI, and TAHGMU external validation, respectively (figure 3C; appendix 2 p 13). For distinguishing benign from TBSRTC V+, the AUROC was 0·990 (95% CI 0·986–0·995) for SYSMH internal validation and 0·988 (0·980–0·995), 0·965 (0·953–0·977), and 0·991 (0·972–1·000) for FPHF, SCHI, and TAHGMU external validation, respectively (figure 3D; appendix 2 p 13). These performances were consistent across various subgroups, including different sexes and ages (appendix 2 p 16). Moreover, we calculated Cohen's κ values to quantify the agreement between ThyroPower and senior cytopathologists. The overall Cohen's κ value of TBSRTC III+ across all four medical centers was 0·729, suggesting agreement. Considering relatively lower agreement among cytopathologists for TBSRTC III and IV, we also calculated κ for TBSRTC categories V+. The overall Cohen's κ value for these categories across all four centres was 0·921. These κ values show good consistency of ThyroPower in classifying different TBSRTC categories with senior cytopathologists' assessments (appendix 2 p 13).

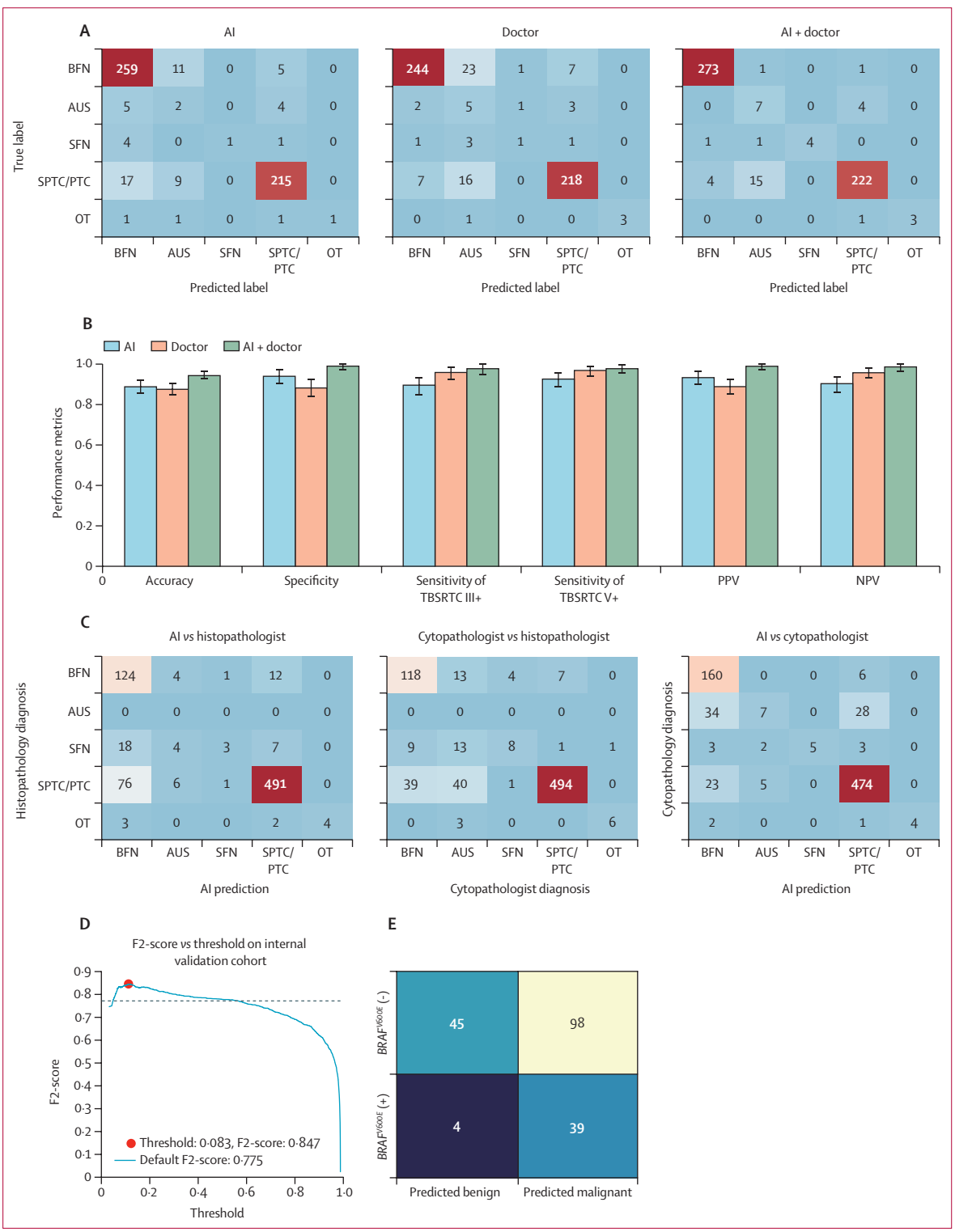
To investigate the performance of ThyroPower in a real-world situation, we conducted a prospective validation study. A total of 1064 WSIs were obtained from 537 smears were collected from 469 patients. The AUROC of distinguishing benign from TBSRTC III+ and TBSRTC V+ were 0·977 and 0·981, respectively (appendix 2 p 14).

We further evaluated the ThyroPower system's ability to assist junior cytopathologists in diagnostics. All samples were initially diagnosed by junior cytopathologists and then the results of AI prediction, and the indication of the ROIs were displayed. With our ThyroPower graphic user interface, cytopathologists could easily and quickly find the ROIs with potential malignant cells (appendix 2 p 24). The junior cytopathologist re-evaluated their judgement and made an updated decision with the assistance of the ThyroPower system. ThyroPower boosted the accuracy from 0·877 (95% CI 0·846–0·904) to 0·948 (0·926–0·965), and improved the specificity from 0·887 (95% CI

Figure 4: **Performance of the ThyroPower in AI-assisted situation on prospective dataset and consistency of *BRAF* mutation status on retrospective dataset**
(A) Confusion matrix showing the detailed classification results of AI, doctor (junior cytopathologists), and AI + doctor (cytopathologists with AI assisting). The descriptions of different categories are listed in appendix 2 (pp 9–10). The heatmap colors of the grids represent the number of samples. The colours range from blue to red, with red indicating a larger number of samples and blue indicating a lower number of samples. (B) The performance of AI, doctor (junior cytopathologists), and AI + doctor (cytopathologists with AI assisting). The specificity, sensitivity of TBSRTC III and TBSRTC V+, PPV, and NPV are shown, and the error bars are 95% CIs. (C) Using samples with postoperative histopathology diagnosis, the diagnosis results of cytopathologists and AI were compared. Confusion matrix showing the detailed results of AI *vs* true label, cytopathologists *vs* true label, and AI *vs* cytopathologists. (D) An F2-score-versus-threshold curve showing the relationship between F2-score and threshold, with the star indicating the highest F2-score (threshold: 0·087, F2-score: 0·847). This threshold was used to make predictions from AUS samples. (E) The confusion matrix showing AI prediction results and $BRAF^{V600E}$ mutation status indicates a high degree of consistency between predicted malignant and $BRAF^{V600E}$-positive samples. As a rule in biomarker, $BRAF^{V600E}$-positive samples were considered malignant. $BRAF^{V600E}$-negative samples cannot provide definite information of malignant or benign status. AI=artificial intelligence. AUS=atypia of undetermined significance. BFN=benign follicular nodule. NPV=negative predictive value. OT=other tumours. PPV=positive predictive value. PTC=papillary thyroid carcinoma. SFN=suspicious for a follicular neoplasm. SPTC=suspicious papillary thyroid carcinoma. TBSRTC=The Bethesda System for Reporting Thyroid Cytopathology.

0·844–0·922) to 0·993 (0·974–0·999; figure 4A, B; appendix 2 p 14).

The above-mentioned assessments and comparisons were largely based on the judgement of senior cytopathologists. However, postoperative histopathology serves as the golden standard for the diagnosis of thyroid cancer, as it allows for the acquisition of more cell numbers and information regarding tissue structure.[45]



To make a more objective and accurate evaluation, we compared the diagnosis results of ThyroPower and senior cytopathologists using the results of postoperative histopathology information. 757 WSIs from 651 patients in the internal validation set that had postoperative histopathology diagnoses were selected. Both ThyroPower and senior cytopathologists made solid predictions with high sensitivity and specificity (figure 4C; appendix 2 p 23). ThyroPower had a slightly lower sensitivity but higher specificity compared with senior cytopathologists, the sensitivity of AI and senior cytopathologists was 0·842 and 0·922, respectively; the specificity of AI and senior cytopathologists was 0·880 and 0·831, respectively (appendix 2 pp 15, 23).

We explored ThyroPower's capability to identify malignant nodules from AUS samples with *BRAF* mutations using 186 AUS samples from 186 patients with *BRAF* mutation information.

Although the ThyroPower is a multi-class classifier matching the TBSRTC guidelines, we also explored an additional thresholding step to aid the further diagnosis of AUS samples. We applied the thresholding on the output probability of the BFN class, since it indicates the likelihood of a sample being benign. The threshold was selected based on the F2-score-versus-threshold curve calculated on the internal validation cohort (figure 4D).

The F-beta measure is a weighted harmonic mean of sensitivity and precision (PPV), and the F2-score indicates more weight is given to sensitivity.[46] Considering cytological AUS samples' low-risk nature and the lesser degree of atypia, malignancies in AUS samples are more challenging to determine. Therefore, the F2-score was adopted for threshold selection.

We chose the optimal threshold values with the maximal F2-score (figure 4D), cytological AUS samples with a benign probability lower than the selected threshold were suggested as malignant and might be recommended for further clinical management. The confusion matrix in figure 4E shows that 39 (90·7%) of 43 of *BRAF* mutation-positive samples were correctly identified by our thresholding strategy.

Discussion

The cytological diagnosis of FNA biopsy is vital for surgical therapeutic intervention. With the large number and rapid growth of patients with thyroid nodules, the shortage of experienced cytopathologists is becoming more pronounced, especially in resource-poor or remote areas. In this study, we developed an AI system for assisted diagnostics of thyroid nodules with a total of 17 966 WSIs of 8426 smears from 7420 patients from four centres. To the best of our knowledge, this is the largest thyroid FNA dataset under the supervision of experienced cytopathologists. Previous studies showed large datasets allow for designing a network with more trainable parameters, which leads to better classification performance,[47] whereas size reduction results in decreasing performance with respect to all metrics for almost all classifiers.[48] We consider the data in this study to be representative. The medical centres involved in our study are geographically diverse within China, and their patient populations come from all over the country. The patients were randomly selected to ensure the representativeness of our sample. Furthermore, the sex ratio in our study is approximately 1:4 (male patients to female patients), which aligns with previous studies.[49] With innovative deep learning models and a large dataset, the ThyroPower system had good performance on both internal and external validation datasets, as well as subgroups such as different sexes and different ages, collecting more patient data can further improve the performance. Therefore, we have successfully developed an innovative deep learning diagnostic system that has shown promising and reliable results on training data.

We acknowledge that while a patient might have multiple smears, each representing distinct thyroid nodules and thereby maintaining their independence. The analytical design is smear-based rather than patient-based. This approach aligns with the common scenario in clinical practice.

The inter-rater and intra-rater variability among cytopathologists could affect the performance of the AI models. Although the inter-rater variability among senior cytopathologists was less than 5% based on our routine clinical practice (unpublished data), both inter-rater and intra-rater variability are inevitable. We used several strategies to mitigate the effects of this variability in our study. Apart from engaging experienced cytopathologists to minimise variability, we conducted our study using a large sample size drawn from various medical centres, to reduce the influence of variability on our models' robustness. Additionally, to further evaluate the performance of ThyroPower, we compared the diagnosis results of ThyroPower with those of senior cytopathologists, using the gold standard postoperative histopathology results. The predictions of ThyroPower showed good consistency with both histopathology and cytopathologic results.

The internal validation cohort had a higher population of AUS and BFN samples (table), resulting in a drop in performance for TBSRTC III+ sensitivity, which was was attributed to the increased proportion of these ambiguous and less definitive categories within the cohort. This drop in performance was attributed to the increased proportion of these ambiguous and less definitive categories within the cohort. Nevertheless, the ThyroPower system achieved a robust performance on samples from different data sources. Moreover, the ThyroPower system accurately identified both TBSRTC II samples and TBSRTC V and VI samples, showing its ability to avoid under-diagnosis and over-diagnosis.

This study has some limitations. Firstly, we found that the most false negative diagnoses of both ThyroPower and cytopathologists were from patients with SFN. The

ThyroPower model has lower accuracy for this challenging category, so it is necessary to accumulate more training data to improve the diagnosis accuracy in this category. Secondly, the prevalence of the dataset has a significant effect on the predictive values (NPV and PPV). As the SYSMH retrospective validation dataset shows, most patients with a cytological diagnosis of BFN did not have a postoperative histopathology result, resulting in a prevalence that is higher than in the actual population. This finding leads to a bias in the NPV and PPV, with the numbers being lower or higher than the true values. Thirdly, we excluded a number of WSIs from SCHI and the prospective set because the WSIs were blurry, colour fading, or the WSI was not of thyroid nodule origin. These exclusion samples were determined based on the need to maintain data quality and ensure that only reliable and relevant images were included in our analysis. We acknowledge that the exclusion samples for these centres were higher than desired, which might have resulted in a decrease in the number of certain types of thyroid smears, potentially influencing the results. Finally, the majority of patients are of Chinese Han ethnicity, as the Han ethnic group comprises over 90% of the Chinese population. However, detailed ethnicity data are unavailable, as it is not mandatory for patient registration, leaving the performances of different ethnic subgroups unknown.

Indeterminate thyroid nodules account for about 10–30% of thyroid FNA biopsies, creating a dilemma for both patients and doctors. Molecular testing could help with diagnosing these samples, but it is only feasible for a portion of patients for various reasons. Theoretically, AI can search the very details of every single cell, which could capture many cellular features neglected by cytopathologists. Using samples informed with *BRAF* mutation status, we found that by adding a simple thresholding step, the ThyroPower could provide a reliable recommendation for identifying malignancies in AUS samples. In future studies, we aim to collect more samples with molecular testing information to train AI systems to achieve a more sensitive and specific prediction of AUS samples.

Besides high performance, AI models for medicine should also be interpretable,[50] and our ThyroPower system pipeline has followed this principle. Our model pipeline imitates the diagnosis process of a pathologist, which is to scan through the whole slide to see if there are any abnormalities and then give a diagnosis based on what has been found. We do so by training a patch-level classifier to locate those cells or clusters that are suspicious of malignancy, then building our whole slide-level classifiers based on these abnormal cells or clusters that we found. We also mimic the scenario when the diagnosis was made by multiple experts by fusing two classifiers at the diagnostic phase. Furthermore, the training data were annotated by senior cytopathologists according to TBSRTC, which is a standard widely accepted in various countries. This approach ensures the quality of our training data. Finally, our systems are expected to enhance rather than replace human experts' capabilities.[50] The ROIs were displayed along with the AI's final decision. Cytopathologists could efficiently go through these ROIs and make the final diagnostic decision. In summary, this study demonstrated the potential of AI in cytological diagnosis by effectively assisting cytopathologists in the decision-making process, thereby improving diagnostic accuracy and efficiency. This system has significant applications in real-world clinical practice, particularly in resource-limited or remote areas, where it can substantially alleviate the shortage of pathologists.

Contributors

YZ, NO, JiW, and JuW conceptually designed the study. JuW, NZ, HW, SJ, XZ, and Q J performed the cytopathologic analysis. JuW, NZ, HW, SJ, XZ, and SF did the annotation of slides. QY, XC, SL, and RC built the models of the AI. JuW, NZ, HW, SJ, XZ, SF, GH, BL, and Q J provided the thyroid cytology slides. JiW, QY, YZ, and JuW analysed the data. YZ, JiW, JuW, QY, and NO wrote the manuscript. YZ, QY, JiW, JuW, NZ, HW, SJ, XZ NO, and Q J discussed and reviewed the manuscript. All the authors read and approved the final manuscript. All authors had full access to the data and had final responsibility for the decision to submit for publication. YZ, JuW, and QY accessed and verified the data. JuW, NZ, HW, QY, SJ, and XZ contributed equally to this work.

Declaration of interests

We declare no competing interests.

Data sharing

After publication, the data will be made available upon reasonable request to the corresponding author. The algorithm source codes used in this study are available at https://github.com/cellsvision/Thyriod-FNA.

Acknowledgments

We thank Lisha Yan and Xingqun Cai for their help with cytopathologic analysis and Wenwen Zhao for helping to search for the slides. This work was supported by grants from the Guangdong Science and Technology Department (grant number 2023A1515010411 awarded to YZ, and grant numbers 2020B1212060018 and 2020B1212030004).

References

1 Uppal N, Collins R, James B. Thyroid nodules: global, economic, and personal burdens. *Front Endocrinol (Lausanne)* 2023; **14:** 1113977.
2 Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur J Clin Invest* 2009; **39:** 699–706.
3 Alexander EK, Cibas ES. Diagnosis of thyroid nodules. *Lancet Diabetes Endocrinol* 2022; **10:** 533–39.
4 Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; **71:** 209–49.
5 Baskin HJ. Ultrasound-guided fine-needle aspiration biopsy of thyroid nodules and multinodular goiters. *Endocr Pract* 2004; **10:** 242–45.
6 Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* 2016; **26:** 1–133.
7 Filetti S, Durante C, Hartl D, et al. Thyroid cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2019; **30:** 1856–83.
8 Haddad RI, Bischoff L, Ball D, et al. Thyroid Carcinoma, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2022; **20:** 925–51.
9 Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. *Thyroid* 2017; **27:** 1341–46.

10 Wu RI, Hatlak K, Monaco SE. Trends in cytopathology fellowship positions and vacancies over the past decade. *J Am Soc Cytopathol* 2021; **10:** 471–76.

11 Leong AS, Leong FJ. Strategies for laboratory cost containment and for pathologist shortage: centralised pathology laboratories with microwave-stimulated histoprocessing and telepathology. *Pathology* 2005; **37:** 5–9.

12 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:** 713–22.

13 Lodewijk L, Vriens MR, Vorselaars WM, et al. Same-day fine-needle aspiration cytology diagnosis for thyroid nodules achieves rapid anxiety decrease and high diagnostic accuracy. *Endocr Pract* 2016; **22:** 561–66.

14 Yang Z, Zhao X, Zhu Z, Fu Y, Hu Y. How patients with an uncertain diagnosis experience intolerance of uncertainty: a grounded theory study. *Psychol Res Behav Manag* 2021; **14:** 1269–79.

15 Alexander EK, Kennedy GC, Baloch ZW, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. *N Engl J Med* 2012; **367:** 705–15.

16 Jensen CB, Saucke MC, Francis DO, Voils CI, Pitt SC. From overdiagnosis to overtreatment of low-risk thyroid cancer: a thematic analysis of attitudes and beliefs of endocrinologists, surgeons, and patients. *Thyroid* 2020; **30:** 696–703.

17 Dedhia PH, Saucke MC, Long KL, Doherty GM, Pitt SC. Physician perspectives of overdiagnosis and overtreatment of low-risk papillary thyroid cancer in the US. *JAMA Netw Open* 2022; **5:** e228722.

18 Alexander EK, Doherty GM, Barletta JA. Management of thyroid nodules. *Lancet Diabetes Endocrinol* 2022; **10:** 540–48.

19 Nikiforov YE, Ohori NP, Hodak SP, et al. Impact of mutational testing on the diagnosis and management of patients with cytologically indeterminate thyroid nodules: a prospective analysis of 1056 FNA samples. *J Clin Endocrinol Metab* 2011; **96:** 3390–97.

20 Poller DN, Glaysher S. Molecular pathology and thyroid FNA. *Cytopathology* 2017; **28:** 475–81.

21 Halverson CME, Wessinger BC, Clayton EW, Wiesner GL. Patients' willingness to reconsider cancer genetic testing after initially declining: mention it again. *J Genet Couns* 2020; **29:** 18–24.

22 Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. *Med Image Anal* 2021; **67:** 101813.

23 Saldanha OL, Quirke P, West NP, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat Med* 2022; **28:** 1232–39.

24 Schuhmacher D, Schörner S, Küpper C, et al. A framework for falsifiable explanations of machine learning models with an application in computational pathology. *Med Image Anal* 2022; **82:** 102594.

25 Duan W, Gao L, Liu J, et al. Computer-assisted fine-needle aspiration cytology of thyroid using two-stage refined convolutional. *Neural Netw* 2022; **11:** 4089.

26 Li LR, Du B, Liu HQ, Chen C. Artificial intelligence for personalized medicine in thyroid cancer: current status and future perspectives. *Front Oncol* 2021; **10:** 604051.

27 Lin YJ, Chao TK, Khalil MA, et al. Deep learning fast screening approach on cytological whole slides for thyroid cancer diagnosis. *Cancers (Basel)* 2021; **13:** 3891.

28 Bini F, Pica A, Azzimonti L, et al. Artificial intelligence in thyroid field-a comprehensive review. *Cancers (Basel)* 2021; **13:** 4740.

29 Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27–30, 2016: 761–76.

30 Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal* 2021; **71:** 102062.

31 Ren P, Xiao Y, Chang X, et al. A survey of deep active learning. *ACM Comput Surv* 2021; **54;** 1–40.

32 Krane JF, Vanderlaan PA, Faquin WC, Renshaw AA. The atypia of undetermined significance/follicular lesion of undetermined significance:malignant ratio: a proposed performance measure for reporting in The Bethesda System for thyroid cytopathology. *Cancer Cytopathol* 2012; **120:** 111–16.

33 Cibas ES, Baloch ZW, Fellegara G, et al. A prospective assessment defining the limitations of thyroid nodule pathologic evaluation. *Ann Intern Med* 2013; **159:** 325–32.

34 Stanek-Widera A, Biskup-Frużyńska M, Zembala-Nożyńska E, Półtorak S, Śnietura M, Lange D. Suspicious for follicular neoplasm or follicular neoplasm? The dilemma of a pathologist and a surgeon. *Endokrynol Pol* 2016; **67:** 17–22.

35 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44:** 837–45.

36 McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12:** 153–57.

37 Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016; **6:** e012799.

38 Tan M, Le Q, eds. Efficientnet: rethinking model scaling for convolutional neural networks. International Conference on Machine Learning; 2019: PMLR.

39 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition; June 27–30, 2016 (abstr 90).

40 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, eds. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27–30, 2016.

41 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:** 4876.

42 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014; published online Sept 4. https://arxiv.org/abs/1409.1556 (preprint).

43 Teramoto A, Yamada A, Kiriyama Y, et al. Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. *Inform Med Unlocked* 2019; **16:** 100205.

44 Elharrouss O, Akbari Y, Almaasees N, Al-Maadeed S. Backbones-review: feature extraction networks for deep learning and deep reinforcement learning approaches. *arXiv* 2022; published online June 16. https://arxiv.org/pdf/2206.08016 (preprint).

45 Gąsiorowski O, Leszczyński J, Kaszczewska J, et al. Comparison of fine-needle aspiration cytopathology with histopathological examination of the thyroid gland in patients undergoing elective thyroid surgery: do we still need fine-needle aspiration cytopathology? *Diagnostics (Basel)* 2024; **14:** 236.

46 Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York, NY: ACM press, 1999.

47 Zhang C, Bengio S, Hardt M, Recht B, Vinyals OJCotA. Understanding deep learning (still) requires rethinking generalization. *Commun ACM* 2021; **64:** 107–15.

48 Althnian A, AlSaeed D, Al-Baity H, et al. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Appl Sci (Basel)* 2021; **11:** 796.

49 Yin T, Zheng B, Lian Y, et al. Contrast-enhanced ultrasound improves the potency of fine-needle aspiration in thyroid nodules with high inadequate risk. *BMC Med Imaging* 2022; **22:** 83.

50 Kundu S. AI in medicine must be explainable. *Nat Med* 2021; **27:** 1328.