The American Journal of Pathology, Vol. 193, No. 9, September 2023

The American Journal of
PATHOLOGY
ajp.amjpathol.org

MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

Deep-Learning–Based Screening and Ancillary Testing for Thyroid Cytopathology

David Dov,*† Danielle Elliott Range,† Jonathan Cohen,‡ Jonathan Bell,† Daniel J. Rocke,§ Russel R. Kahmke,§ Ahuva Weiss-Meilik,* Walter T. Lee,§ Ricardo Henao,¶‖ Lawrence Carin,** and Shahar Z. Kovalsky††

From the I-Medata AI Center, Tel-Aviv Sourasky Medical Center, Tel Aviv-Yafo, Israel; the Departments of Pathology† and Head and Neck Surgery & Communication Sciences,§ Duke University Medical Center, Durham, North Carolina; the Department of Head and Neck Surgery,‡ Kaplan Medical Center, Rehovot, Israel; the Biological, Environmental Sciences and Engineering Division¶ and Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division,** King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia; the Department of Biostatistics and Bioinformatics,‖ Duke University, Durham, North Carolina; and the Department of Mathematics,†† University of North Carolina at Chapel Hill, Chapel Hill, North Carolina*

Accepted for publication
May 19, 2023.

Address correspondence to
David Dov, Ph.D., I-Medata AI
Center, Tel Aviv Sourasky
Medical Center, Tel Aviv-Yafo,
Israel.
E-mail: daviddov@tlvmc.gov.il.

Thyroid cancer is the most common malignant endocrine tumor. The key test to assess preoperative risk of malignancy is cytologic evaluation of fine-needle aspiration biopsies (FNABs). The evaluation findings can often be indeterminate, leading to unnecessary surgery for benign post-surgical diagnoses. We have developed a deep-learning algorithm to analyze thyroid FNAB whole-slide images (WSIs). We show, on the largest reported data set of thyroid FNAB WSIs, clinical-grade performance in the screening of determinate cases and indications for its use as an ancillary test to disambiguate indeterminate cases. The algorithm screened and definitively classified 45.1% (130/288) of the WSIs as either benign or malignant with risk of malignancy rates of 2.7% and 94.7%, respectively. It reduced the number of indeterminate cases ($N = 108$) by reclassifying 21.3% ($N = 23$) as benign with a resultant risk of malignancy rate of 1.8%. Similar results were reproduced using a data set of consecutive FNABs collected during an entire calendar year, achieving clinically acceptable margins of error for thyroid FNAB classification. *(Am J Pathol 2023, 193: 1185–1194; https://doi.org/10.1016/j.ajpath.2023.05.011)*

In recent years, significant progress has been made in applying machine-learning algorithms to analyze whole-slide images (WSIs). Such algorithms can reduce pathologists' workload, lower interreviewer variability, and provide a second level of diagnostic analysis. Studies have reported deep-learning algorithms that achieve expert (human)-level performance for the prediction and grading of prostate cancer,[1,2] the prediction of lung cancer,[3] survival predictions in patients with mesothelioma,[4] and the diagnosis of brain tumors.[5] These studies address the automated examination of surgical pathology histologic specimens that contain whole tissue sections.

The analysis of cytologic specimens presents two unique challenges in machine learning that differ from surgical pathology. The first relates to the acquisition of the cytologic biopsy material. A fine-needle aspiration biopsy (FNAB) takes a small needle to retrieve cellular material that is then mechanically spread across a glass microscope slide to generate a cytopathology smear. These smears comprise sparse aggregates of diagnostic, cellular material in a large background of diagnostically irrelevant areas composed of blood and empty space. This generates a computational challenge of separating the signal from the noise. As a result, many studies in computational cytopathology using FNABs are limited to small data sets, often artificially balanced to include, for example, equal numbers of malignant and benign cases.[6–14] Other studies require significant human intervention to manually select diagnostic regions during testing. In a previous study,[15] we addressed

Supported by NIH award number 1R21CA268428-01 (D.D., D.E.R., D.J.R., and W.T.L.).
D.D., D.E.R., and J.C. contributed equally to this work.
Disclosures: None declared.

https://doi.org/10.1016/j.ajpath.2023.05.011

this by presenting a deep-learning algorithm that first identifies the diagnostic material and, in turn, provides classification of malignancy.

The second challenge is the inherent, sizable diagnostic gray zone of atypical cases seen in all cytologic specimens. This uncertainty is a result of the small amount of diagnostic material and the limitations of examining individual cells without the benefit of tissue architecture. The ambiguous nature of the atypical diagnostic category and lack of a gold standard or ground truth pose challenges in defining both clinical-grade performance for automated algorithms and forming actionable protocols for clinical decision-making (eg, surgical management).[16] In the current paper, these challenges are addressed as they relate to thyroid cytopathology.

Thyroid nodules are a common condition—in the United States, there is an estimated 10% risk of acquiring a thyroid nodule in one's lifetime.[17] Thyroid FNABs are essential for the diagnosis and management of thyroid nodules. The Bethesda System for the Reporting of Thyroid Cytopathology (TBS) is a widely used, six-tiered system used to classify thyroid FNABs into diagnostic categories. Each TBS category is associated with a calculated risk of malignancy (ROM) based on final surgical pathology results: nondiagnostic (5% to 10% ROM), benign (1% to 3% ROM), atypical (10% to 30% ROM), neoplasm (25% to 40% ROM), suspicious (50% to 75% ROM), and malignant (97% to 99% ROM).[18] In this article, we collapse the TBS into three clinically relevant categories that align with management guidelines typically associated with the TBS diagnoses: i) benign: surgery is not required; ii) indeterminate (atypical and neoplasm): variable management; and iii) malignant (suspicious and malignant): surgery is recommended.[19,20] As many as 20% to 40% of the FNABs in the indeterminate categories (atypical and neoplasm) result in a benign postoperative diagnosis and a potentially unnecessary surgery.[21]

In this article, we propose a deep-learning–based ternary, rather than a classic binary, model, which classifies each FNAB scan into one of the three categories: benign, indeterminate, or malignant. At the expense of providing indeterminate classifications in some of the cases, the model is tuned to provide accurate low and high ROM for the benign and malignant categories, respectively. This approach allows us to apply the algorithm in two practical use cases while achieving clinical-grade performance: 1) screening to identify determinate cases (ie, providing definitive and reliable predictions that do not require further manual review by pathologists); and 2) ancillary testing for disambiguating and reducing the number of indeterminate cases, to help reduce unnecessary surgeries. The algorithm screens and definitively classifies 45.1% (130/288) of the scans as either benign or malignant, while providing human expert-level ROMs of 2.7% and 94.7%, respectively. The algorithm further reduces the number of indeterminate cases by definitively classifying 21.3% (23/108) with a ROM of 1.8%. These results were obtained on a test set comprising a WSI of a single representative slide containing the largest amount of diagnostic material from each FNAB. These results are reproduced on another test set containing multiple slides per FNAB, demonstrating that there is no need to manually select the representative slide. The same results are reproduced using a test set of consecutive FNABs collected during an entire calendar year, achieving the clinically acceptable margins of error for thyroid FNAB.

Materials and Methods

Data Set

After obtaining institutional review board approval, the authors compiled a data set of WSIs using archival FNABs with subsequent thyroid surgical specimens from January 2008 to December 2018, performed at Duke Health. All cytologic (TBS) and surgical pathology diagnoses were recorded as documented in the electronic medical record (EMR). The surgical pathology diagnosis served as the ground truth in this study. The authors excluded cases with the following: i) nondiagnostic FNABs in the EMR, defined by TBS as those cases containing less than six follicular groups; ii) an equivocal final pathology diagnosis of noninvasive follicular tumor with papillary-like features (NIFTP), which could not be placed into a benign or malignant ground truth category; iii) any biopsied nodule that could not be directly correlated with the same nodule in the surgical pathology report; iv) damaged WSIs including files that could not be opened or visualized because of scanning malfunction (ie, they were corrupted, partially/poorly scanned, or heavily pixelated); and v) FNAB slides in the training set from patients who were included in the test set. No case was excluded on the basis of diagnostic difficulty, poor visualization due to blood or clot, or because of artifacts typically seen in archival materials, such as air bubbles or faded stain.

The cohort comprised 2169 FNAB slides. The authors excluded FNABs diagnosed as nondiagnostic by the EMR cytopathologist (CP), which comprised 3.2% of the cases ($N = 70$) with an additional 13 cases classified as nondiagnostic by a CP reviewer participating in the study. Seventy-eight cases (3.6%) were excluded because the digital scan was corrupted/damaged ($N = 58$) or there was no correlation between the nodule biopsied and the surgical pathology diagnosis ($N = 20$). The latter typically included thyroids with multiple nodules in which one was malignant. Eighty cases were excluded because they were part of the training set. The final data set comprised 1928 WSIs from 1565 FNABs. To the authors' knowledge, this is the largest reported WSI data set in thyroid computational cytopathology that includes final surgical pathology diagnoses.

The final data set was divided into a training set of 964 FNABs and a test set of 601 FNABs. One representative alcohol-fixed, Papanicolaou-stained, direct smear was selected for scanning from each FNAB, by a medical student (J.B.). The slide with the largest amount of diagnostic material was considered representative. The training set

Table 1 TBS Category Breakdown in the Retrospective and the Consecutive Test Sets

TBS category	Retrospective test set (N = 288)	Consecutive test set (N = 313)
II: Benign	142 (49.3)	125 (40.0)
III: Atypical	85 (29.5)	104 (33.2)
IV: Neoplasm	23 (8.0)	26 (8.3)
V: Suspicious for malignancy	10 (3.5)	18 (5.8)
VI: Malignant	28 (9.7)	40 (12.8)

Data are given as number (percentage) of each group.
TBS, Bethesda System for the Reporting of Thyroid Cytopathology.

included only the representative slide (ie, one per FNAB). The test set included a retrospective set of 288 FNABs and a consecutive test set of 313 FNABs. The retrospective test set included FNABs collected from 2013 to 2016. The FNABs were selected by an arbitrary range based on scanning date. This yielded a fairly representative distribution of cases within each TBS category, as described in the literature and as seen in the consecutive test set (Table 1). The retrospective test set was further divided into two groups: a single-slide retrospective (SSR) group containing the one representative WSI from each FNAB case; and a multislide retrospective (MSR) group that contained all alcohol-fixed, Papanicolaou-stained slides (including the representative slide) from the same FNAB. For the SSR test set, the authors compared algorithm performance with the performance of three board-certified/eligible cytopathologists who reviewed the same slides and provided a cytologic diagnosis for each case using the TBS. CP1 (D.E.R.) is a board-certified cytopathologist with 20 years of post-graduate experience. CP2 is a board-certified cytopathologist with 4 years of post-graduate experience. CP3 is a board-eligible cytopathologist with less than one year of post-graduate experience.

The MSR test set contained an additional 363 slides, a total of 651 WSIs from 288 FNABs, and was used to demonstrate the ability of the algorithm to provide a single prediction per FNAB using multiple slides, without the need for manual selection of a representative slide. Because TBS categories are assigned in clinical practice based on review of all slides, the algorithm's performance for the MSR test set was compared with TBS diagnoses extracted from the EMR. The single-slide consecutive test set comprised one representative slide from all thyroid FNABs with surgical follow-up during the 2018 calendar year. The authors used consecutive cases to evaluate the reproducibility of the algorithm's performance on data with real-world distribution of cases across the TBS categories. Table 1 shows a comparison of the cases in each test set by TBS category.

All slides were cleaned and scanned with a 40× objective and nine levels of Z-stack on a Leica AT-2 scanner (Leica Biosystems, Nußloch, Germany). The authors used the middle Z-stack, which was further down-sampled by a factor of four in each dimension to reduce processing time. The authors did not find a quantifiable advantage in applying their algorithm to the Z-stacked scans with the full resolution (results not shown). On a subset of 145 scans from the training set, CP1 annotated 4494 regions containing diagnostic follicular cells. These were used to train the algorithm to distinguish informative regions of interest (ROIs; ie, patches containing follicular cells, from background areas containing white space, blood, and non-follicular cells), as detailed later.

In this article, the authors collapse the TBS into three clinically relevant categories that align with management guidelines typically associated with the TBS diagnoses: i) benign: surgery is not required; ii) indeterminate (atypical and neoplasm): variable management; and iii) malignant (suspicious and malignant): surgery is recommended.[19,20]

Algorithm

The proposed algorithm is inspired by the workflow of cytopathologists and comprises two convolutional neural networks (CNNs), as illustrated in Figure 1. The first network, termed informativeness CNN, discriminates thyroid follicular cells. These diagnostically relevant areas typically comprise only a tiny fraction of the entire slide, which is otherwise mostly occupied by irrelevant cellular material (eg, blood cells). The informativeness CNN mitigates this challenge by selecting only relevant areas, effectively reducing data dimensionality.

The second network, malignancy CNN, classifies FNABs into the three clinically relevant categories (benign, indeterminate, or malignant). The classification is based on ordinal regression whereby a scalar output of the network is compared with two learnable threshold parameters. During the training phase, the threshold parameters are tuned together with the parameters of the neural network via stochastic gradient descent. By the nature of ordinal regression, the three categories reflect increasing probability of malignancy.

Both CNNs are based on the widely used Visual Geometry Group (VGG) 11 architecture.[22] The convolutional layers are pretrained on Imagenet, a common data set of natural images,[23] and the networks are tuned (including the convolutional layers) using stochastic gradient descent with a learning rate of 0.001, momentum of 0.99, and weight decay of 10^{-7}.

Each of the RGB (red, green, blue) color channels of the scans was normalized to have 0 mean and variance 1. Then, the scans were tiled into patches of 128 × 128 pixels and fed into the informativeness CNN that predicts if they are informative (ie, contain thyroid follicular cells). During the training, the informative patches were sampled from the subset of regions manually marked by D.E.R. in the training set. Direct smears made from FNABs contain far more uninformative regions (eg, blank regions, blood cells, and artifacts) than informative ones. In some scans, which typically contain hundreds of thousands of patches, merely a few of them are informative. Because manually annotating

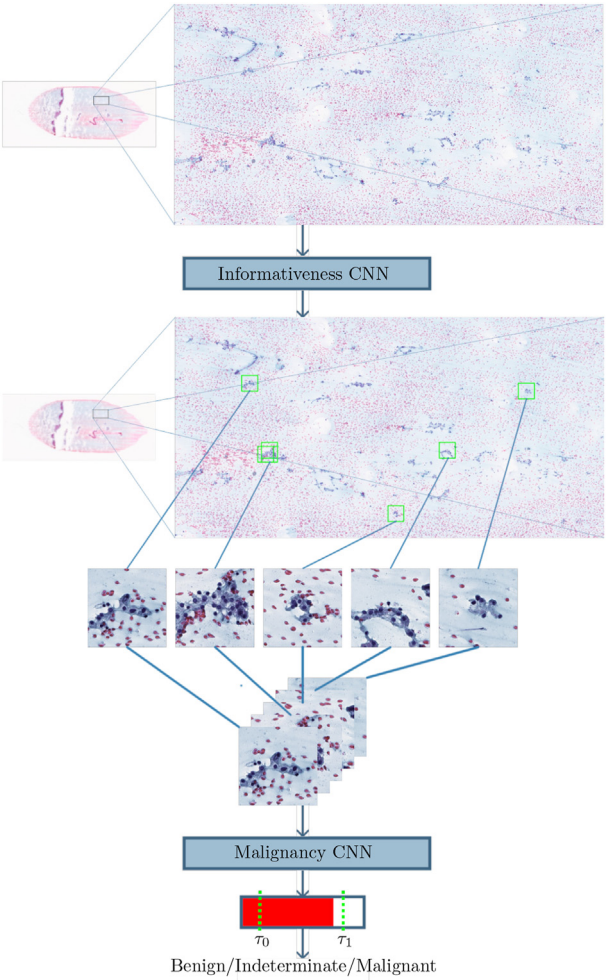

Figure 1 Illustration of the thyroid malignancy prediction algorithm. The informativeness convolutional neural network (CNN) identifies the most informative image regions. These are grouped into batches (sets) and processed by the malignancy CNN, which outputs the average of per-region predictions. The output is compared with learned thresholds providing classifications into the benign, indeterminate, and malignant classes.

regions in the scans is extremely time-consuming, the authors decided to devote this effort to only mark informative regions. The uninformative regions were sampled uniformly from the WSI given the overwhelmingly high likelihood of sampling background/negative areas. Despite the small chance that an informative region will be labeled as uninformative, in the experiments previously presented,[15] the informativeness CNN provided a high area under the receiver operating characteristic curve (AUROC) of 0.985.

After completing the training process, a sliding window sweeps over the WSI, and the CNN predicts the informativeness of each patch. For each WSI, the most informative patches are selected and organized into a set of patches of a fixed size, which are then fed into the malignancy CNN. The authors used 1000 patches per WSI for training the malignancy CNN, a number that provides a sufficiently large amount of data to train the neural network. In the test phase, the authors used 100 patches per scan, as there is no quantifiable advantage in using larger sets. The authors found this scheme efficient in extracting the informative regions, while filtering out white space and irrelevant material. The authors' patches selection strategy allowed selecting overlapping patches. Therefore, when the number of informative regions was smaller than the fixed number of selected patches (1000 in training and 100 in testing), the informativeness CNN usually selected overlapping regions. An alternative approach is selecting only patches with prediction value of the informativeness CNN higher than a certain threshold value. However, there is no straightforward way to select the threshold value, and this alternative approach did not provide improvement in early experiments.

The malignancy CNN provides predictions of the final surgical pathology diagnosis by averaging the predictions obtained from each patch in the set. To transform the algorithm's predictions into clinically relevant classifications of benign, indeterminate, and malignant categories, the authors used learnable threshold parameters to which they compare the (continuous) output of the malignancy CNN. Let $p \in [0, 1]$ be the (continuous) output of the malignancy CNN (after the sigmoid layer) and let $\tau_0, \tau_1, \tau_2, \tau_3 \in \mathrm{R}$ be the learnable thresholds. The thresholds divide the predictions into ranges associated with the different TBS categories, each with an increased risk of malignancy, according to Table 2.

In the training phase, the malignancy CNN simultaneously predicts one of the five diagnostic TBS categories (excluding the nondiagnostic category) via ordinal regression and the final surgical pathology label of benign or malignant, using the same single output P. This strategy allows the authors to automatically tune the threshold parameters as part of the training process while allowing the malignancy CNN to learn from the final pathology labels, which are the gold standard/ground truth. During testing, the

Table 2 Association of Threshold Parameters to Classification Categories for Ordinal Regression

The range of P value	TBS category	The range of P value	Classification
$p < \tau_0$	Benign	$p < \tau_0$	Benign
$\tau_0 < p < \tau_1$	Atypical		
$\tau_1 < p < \tau_2$	Follicular neoplasm	$\tau_0 < p < \tau_2$	Indeterminate
$\tau_2 < p < \tau_3$	Suspicious for malignancy		
$\tau_3 < p$	Malignant	$\tau_2 < p$	Malignant

Left, training phase; right, testing phase.
TBS, Bethesda System for the Reporting of Thyroid Cytopathology.

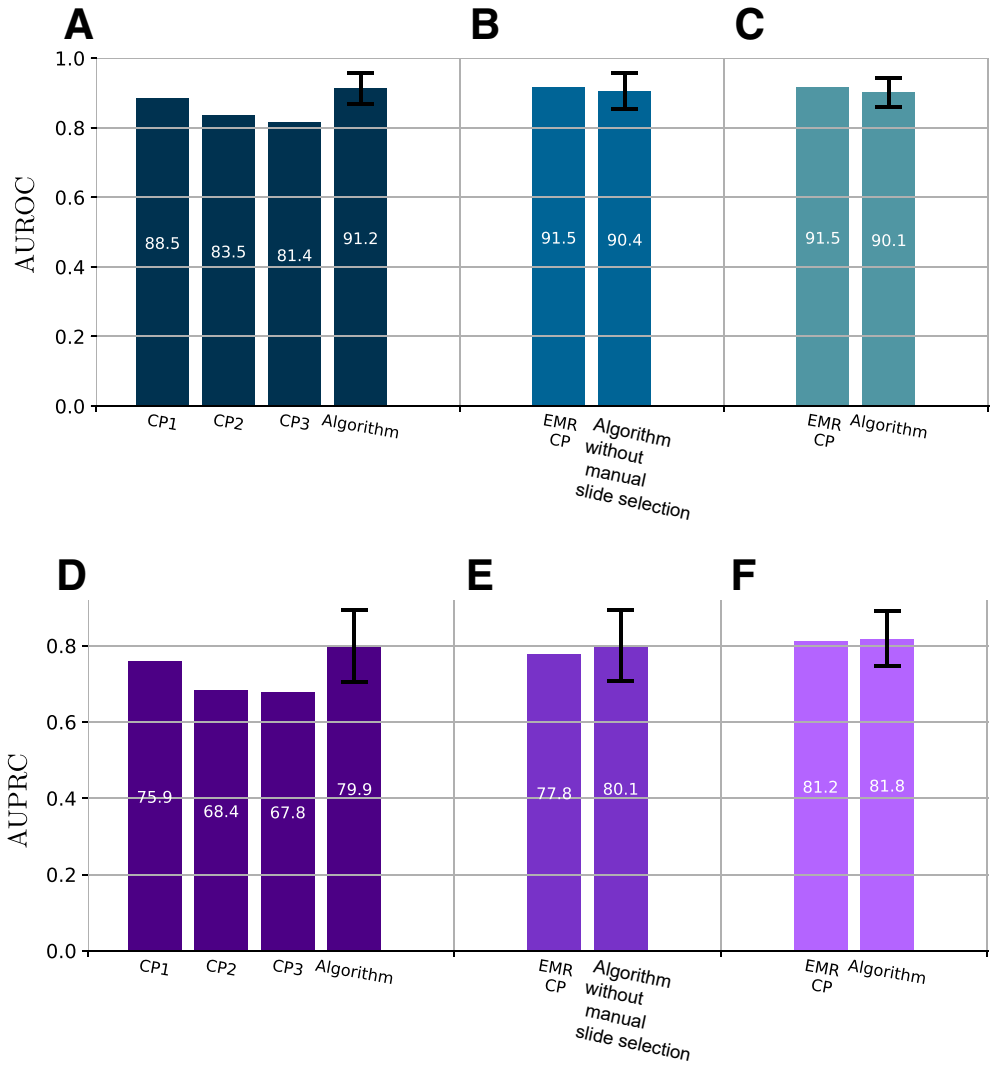

Figure 2 **A**–**F:** The predictive power of the algorithm in the form of area under the receiver operating characteristic curve (AUROC; **A**–**C**) and area under the precision-recall curve (AUPRC; **D**–**F**); higher is better. **A** and **D:** Single-slide retrospective test set. Prediction based on a single whole-slide image (WSI): the performance of the algorithm compared with three cytopathologists (CPs). **B** and **E:** Multislide retrospective test set. Prediction based on multiple slides without the need to manually select the representative WSI: algorithm compared with Bethesda System for the Reporting of Thyroid Cytopathology categories extracted from electronic medical record (EMR). **C** and **F:** Consecutive test set. Algorithm performance is presented with 95% CIs. $N = 288$ (**A** and **D**); $N = 650$ slides (**B** and **E**); $N = 313$ (**C** and **F**).

authors use the threshold parameters τ_0 and τ_2 to divide the prediction into the three clinically relevant categories: benign, indeterminate, and malignant. According to the ordinal regression model, a predicted value below the threshold τ_0 implies that the category is benign, and it is equivalent to predicting TBS2 category. When the predicted value is higher than the threshold τ_2, the predicted category is malignant and is equivalent to predicting that the TBS category is 5 or 6. Similarly, the intermediate category is equivalent to TBS 3 or 4. To classify the cases into the three categories, it is possible to train the algorithm using the threshold τ_0 and τ_2 only. Another alternative is to use a standard multiclass approach, where the model has separate outputs for each of the three classes: benign, malignant, and indeterminate. However, the authors used the ordinal regression approach with all thresholds as well as the postoperative surgical pathology for training because, in a previous study,[15] the authors showed that this strategy outperformed multiple alternative training approaches. More broadly, deep-learning models often perform better when they are trained to classify multiple labels of different types.[24] For the complete description of the training process, we refer the reader to our previous study.[15]

Hardware and Software

The experiments were performed on a server equipped with Intel Xeon E5-2699 version 4 processor (Intel Corp., Santa Clara, CA) and 4-T V-100 Peripheral Component Interconnect Express (PCIE) graphics processing units with 16

Table 3 AUC Values for the Classification of FNABs into Benign, Indeterminate, and Malignant for Two Sets of Cases in the SSR Test Set: All Cases for Which All Three CPs Agreed on the Classification ($N = 168$; Concordant Group) and All Cases for Which at Least One CP Disagreed With the Other Two ($N = 120$; Discordant Group)

Algorithm/CP	AUC for cases where all CPs agree across groups ($N = 168$)	AUC for cases where one CP disagrees across groups ($N = 120$)
CP1	0.96	0.72
CP2	0.96	0.61
CP3	0.94	0.57
Proposed algorithm	0.96 (95% CI, 0.93–1)	0.76 (95% CI, 0.64–0.88)

AUC, area under the curve; CP, cytopathologist; FNAB, fine-needle aspiration biopsy; SSR, single-slide retrospective.

GB of memory each and 512 GB of RAM (NVIDIA, Santa Clara, CA). The algorithm was implemented in Python 3.5 using PyTorch 0.4.1 library for machine learning (*https://www.python.org*).

Statistical Analysis

The CIs of the AUROC and area under the precision-recall curve (AUPRC) scores of the algorithm in Figure 2 were calculated using bootstrapping[25] with 1000 iterations. To evaluate the statistical significance of the difference of AUROC and AUPRC scores between the algorithm and the cytopathologists, the authors calculated two-sided P values under the null hypothesis of equal distributions. To validate the results, the authors also calculated the CIs and the P values for the AUROC using the Delong test.[26] The authors obtained similar results to bootstrapping with the same statistical conclusions (the results were omitted from the article for brevity).

Results

Predictive Power

The authors evaluated the algorithm's ability to predict the final pathology using WSIs from the SSR test set of $N = 288$ (Figure 2, A and D). The performance was compared with three CPs, who reviewed the corresponding glass slide and assigned a TBS category. Performance was measured in the form of the AUROC and the AUPRC (precision and recall are positive predictive value and sensitivity, respectively). The authors calculated the AUROC and AUPRC for each cytopathologist by assigning their TBS diagnoses with a probability of malignancy. The AUROC and AUPRC of the algorithm (91.2 and 79.9, respectively) are comparable to the CPs. In particular, the CP with the highest scores and the most experience (CP1) falls within the 95% CIs of the algorithm, whereas the area under the curves for the algorithm are significantly higher than the less experienced CP2 and CP3 ($P = 0.003$ and $P < 0.001$, respectively, for AUROC; and $P = 0.017$ and $P = 0.005$, respectively, for AUPRC). No significant differences were found in the algorithm's performance compared with CP1. The CIs of the algorithm's performance in the SSR and the MSR test set overlap (Figure 2, B and E). In addition, the performance of the EMR pathologists lies within the CIs of the algorithm. This demonstrates that there is no quantifiable advantage or need to manually select representative slides. Last, the algorithm produces similar results with the single-slide consecutive test set (Figure 2, C and F).

As a secondary question, the authors wondered how the algorithm performed among cases for which the CPs showed some disagreement. These are presumably more difficult cases, which are less likely to be in the determinate categories of benign or malignant. Table 3 shows area under the curve values for the classification of FNABs into benign, indeterminate, and malignant for two sets of cases in the SSR test set: the concordant group includes all cases for which all three CPs agreed on the classification ($N = 168$), and the discordant group includes all cases for which at least one CP disagreed with the other two ($N = 120$). The algorithm outperformed all three CPs in the discordant group, suggesting that it may be better at malignancy predictions for cases that are more difficult for humans.

Screening

AUROC and AUPRC scores provide insight into the overall predictive power of the algorithm concerning the binary question of benign versus malignant. But they tell little about how the algorithm can be used and performs in a clinical setting where one of the most important applications is related to the indeterminate cases. Using the ternary system previously described, with the classifications of benign, indeterminate, and malignant, the algorithm essentially functions as a screening tool by readily identifying benign and malignant FNABs (determinate cases) with clinically acceptable ROMs to preclude subsequent evaluation by a cytopathologist. The remaining cases are predicted as indeterminate by the algorithm.

A clinically useful screening algorithm should perform at the expected ROM rate for the benign and malignant categories, 1% to 3% and 91%, respectively.[20,27] Figure 3 shows that the algorithm provides determinate classifications for 45.1% (130/288) of the cases in the SSR test set with 2.7% ROM and 94.7% ROM for the benign and malignant categories, respectively. This level of accuracy is comparable to or better than that of the cytopathologists and what is expected in the reported literature. Using the MSR

test set (Figure 3), the algorithm achieved a 5.4% and 100% ROM for the benign and malignant classifications, respectively, encompassing a total of 51% (147/288) of the cases. The authors used the algorithm without any tuning or refinement and applied it only once to the single-slide consecutive test set. For the consecutive test set, the algorithm assigned a determinate category to 39.6% (124/313) of the cases with 2.8% ROM for the benign category and 100% ROM for the malignant category, achieving human-level performance.

Ancillary Testing

The second use of the algorithm could be as ancillary testing, to assist with the disambiguation of cases diagnosed as indeterminate by CPs. The EMR CP diagnosed 108 cases in the SSR test set ($N = 288$) as indeterminate: 86 were benign and 22 were malignant on final pathology. The authors wanted to know which of these indeterminate cases might be reclassified as benign if the EMR diagnosis is augmented by the algorithm's prediction of malignancy. To this end, the authors proposed the following simple rule to apply to these indeterminate CP cases: if the algorithm's prediction is benign, then reclassify the case as benign, regardless of the EMR diagnosis. All other cases remain classified according to the original EMR diagnosis of indeterminate. Figure 4 shows that the use of this application reduces the number of indeterminate cases (augmented EMR) by 21.3% from 108 to 85. The augmented categories yield a 1.8% ROM for the benign category, which is at clinical-grade performance and lies within the range of reported ROM of 1% to 3%.[18] Similar results are seen with the MSR and single-slide consecutive test sets, with 2.9% and 2% ROM, for the benign category, respectively (Figure 4).

Discussion

Machine-learning applications in cytopathology present a unique set of challenges related to both acquisition and processing of FNAB slide images used to predict malignancy. In the current study, we presented an algorithm that addresses both of these challenges with the use of automated ROI detection and a ternary, clinically relevant, classification model.

WSIs of direct smears of FNAB of thyroid nodules contain a large amount of useful and useless information. Our algorithm can separate the ROIs of interest from the background noise via an automated approach that does not require human input. This automated selection of ROIs is then used to predict malignancy. We have shown that this process results in slide-level predictions of malignancy that approach human levels. We used two different data sets to demonstrate human-level predictions of malignancy. Algorithm performance using single-slide and multislide retrospective data sets showed no differences, illustrating the algorithm's ability to select the most relevant ROIs across all slides in a given FNAB case. This process is an important task that effectively identifies ROIs in a noisy

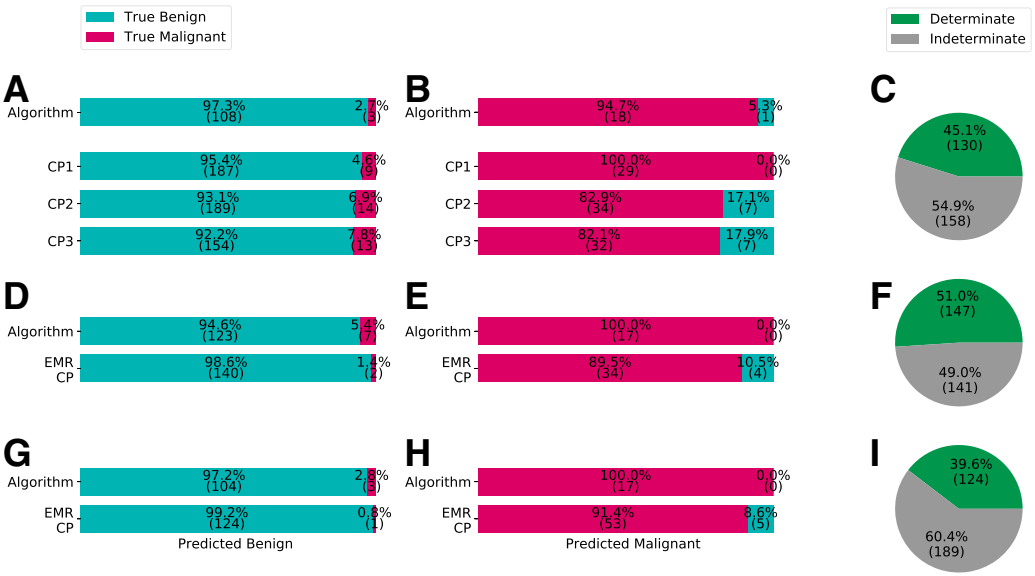

Figure 3 Screening for thyroid cancer. Comparisons of the algorithm's performance in predicting benign and malignant cases. **A**–**C:** Use of the single-slide retrospective data set (SSR) to compare algorithm performance with three cytopathologists (CPs). **D**–**F:** Use of multislide retrospective data set (MSR; 650 whole-slide images) to compare algorithm performance with that of the electronic medical record (EMR). **G**–**I:** Use of the consecutive data set (SSC) to compare performance with the EMR. **A**, **D**, and **G:** Cases predicted benign. **B**, **E**, and **H:** Cases predicted malignant. Cyan indicates cases with benign postoperative diagnosis (true negative). Pink indicates cases with malignant postoperative diagnosis (true positive). **C**, **F**, and **I:** Distribution of conclusive and inconclusive predictions. Green indicates conclusive decisions by the algorithm that need no further review. $N = 288$ slides (**A**–**C**); $N = 288$ FNABs (**D**–**F**); $N = 313$ slides (**G**–**I**).

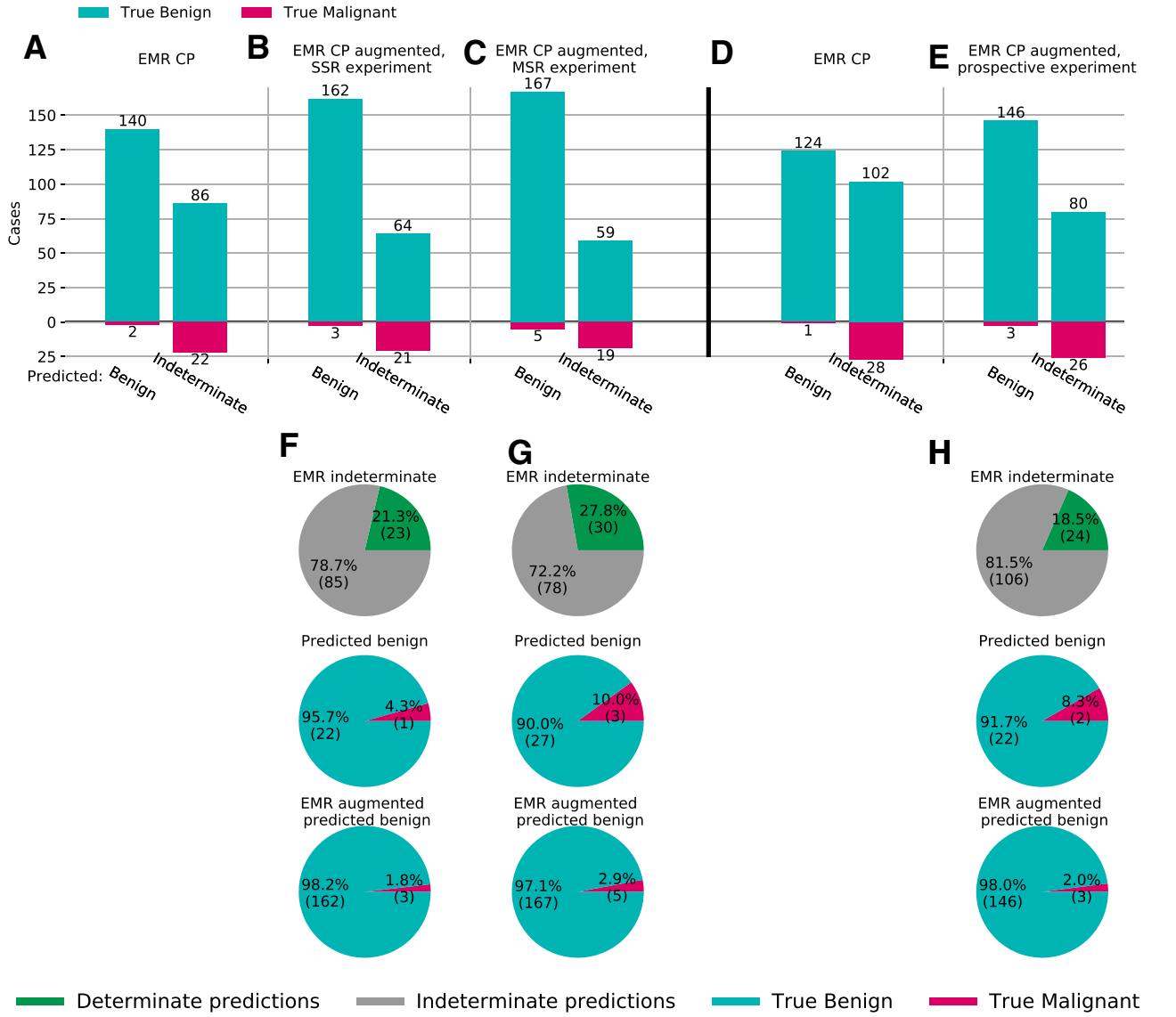

Figure 4 Ancillary testing to reduce indeterminate rates. **A**—**E:** Augmentation of electronic medical record (EMR) cytopathologist (CP) predictions by the algorithm reduces the indeterminate cases. **A**—**E:** EMR CP benign and indeterminate diagnoses (**A**), EMR diagnoses augmented by the algorithm using the single-slide data set (**B**) and using the multislide data set (**C**), EMR diagnoses in the consecutive experiment (**D**), and augmented EMR diagnoses in the consecutive experiment (**E**). **F**—**H: Top panels:** Cases reported inconclusive by EMR CP, predicted benign (green) and indeterminate (gray) by the algorithm. **Middle panels:** Indeterminate EMR cases, predicted benign by the algorithm. **Bottom panels:** EMR cases predicted benign after augmentation by the algorithm. Cyan indicates benign postoperative cases. Pink indicates malignant postoperative cases. **F:** Single-slide retrospective (SSR) data set. **G:** Multislide retrospective (MSR) data set. **H:** Consecutive test data set.

background without human input, a task which can be tedious and time-consuming, and can limit the amount of data that are ultimately processed.

Similar to histopathology, cytopathology diagnosis has a benign and malignant category. However, unlike histopathology, the atypical diagnosis makes up a sizable component of cytopathology diagnoses. In fact, this indeterminate category is inherent to cytopathology and exists across all specimen types. Rather than attempting to apply a binary system to thyroid cytopathology machine learning, as is often used in histopathology, we leveraged the use of a ternary model (ie, benign, indeterminate, or malignant), to generate a more clinically relevant classification system. Within this system, we attempted to address two clinical problems: pathologist workload and reductions in the number of indeterminate cases diagnosed by the pathologist.

The ideal screening tool for thyroid FNABs would classify as many cases as possible into a definitive category (ie, benign and malignant), while keeping the number of false-negative and false-positive cases low. In our case, we aimed to train the algorithm to keep ROM rates at human levels, as established in the literature with the use of the TBS (see studies by Cibas and Ali[18] and Faquin et al[21]). This was best demonstrated with the comparison of the algorithm's

classification of benign cases against the classification of the three cytopathologists and the EMR cytopathologist.

Across all three data sets, the algorithm provided determinate classifications (benign or malignant) in 45.1% to 51% of the cases, implying the potential to significantly reduce cytopathologists' workload by obviating the need for CP review (Figure 3). These classifications maintained an ROM within the clinically relevant range for the benign category of 1% to 3%. This supports our claim that the use of our algorithm can serve to reduce cytopathologists' workloads by screening and classifying more cases as definitively benign, without sacrificing accuracy. Moreover, this screening tool performs at clinically expected human levels with maintenance of ROM in the benign category and little increase in false-negative rates. In the MSR experiment, the determinate category increased compared with the SSR experiment (130 to 147) at the expense of an increase in the ROM to 5.4%. Yet, it is possible to achieve the clinical-grade $<3\%$ ROM by further tuning the threshold value that separates the benign and indeterminate categories. As a limitation, more cases received a determinate classification by the cytopathologists than by the algorithm.

We recognize that a large proportion of tumors in the indeterminate category are follicular-patterned lesions for which carcinomas and adenomas cannot be readily distinguished on FNABs. And given the fact that papillary thyroid carcinoma represents $>85\%$ of all thyroid malignancies, performance among follicular-patterned carcinomas may be overshadowed. Our retrospective cohort contained 61 carcinomas (31 classic papillary thyroid carcinomas and 30 follicular-patterned carcinomas). The fact that almost half (49%) of all the carcinomas were follicular patterned suggests that the performance of the algorithm in this subset of cases is not skewed by the proportion of classic papillary thyroid carcinoma cases.

NIFTP cases would be included in the follicular-patterned lesions. We excluded FNABs with a final pathology diagnosis of NIFTP because this entity, by definition, has no ground truth and is neither benign nor malignant. From a practical perspective, the exclusion of NIFTP cases in our study could be considered a weakness. Since NIFTP, by most studies,[28] is likely a benign entity, we believe it would have served to reduce the ROM in our indeterminate category, and the real-world prediction would have been benign, as is the case for the prediction for most lesions in the indeterminate category. In addition, the prevalence of NIFTP at our institution is low (2.5%)[29] and would be unlikely to have a significant impact on our current data. Still, a more detailed analysis of all follicular-patterned lesions with performance characteristics would be interesting for a future a study.

Finally, we propose that our algorithm can also be applied as an ancillary tool to aid in the reduction of the number of indeterminate cases. This application can have a significant clinical impact as the indeterminate category drives increases in diagnostic surgeries, as well as expensive molecular tests.[19] In addition to these considerations are the effects on patients and their physicians as they attempt to navigate the uncertainties of an indeterminate diagnosis. We have shown that the augmented results of the algorithm, when applied to the indeterminate cases diagnosed by EMR pathologists, resulted in 23 to 30 cases moving from indeterminate to a determinate category, an 18.5% to 27.8% reduction across all three data sets. Similar to the screening application, we saw only small increases in ROM in the indeterminate category using the algorithm, from 20% to 24%, well within the clinically acceptable range.[18] The increase in ROM with the application of the algorithm can also be explained by the fact that the number of true benign is reduced, leaving more true malignant cases in the indeterminate category. This is a reasonable change that assigns more importance to the indeterminate category as a category that should trigger additional testing. We note in this context as a limitation of this study, that we tried in our experiments also to use the algorithm to reclassify indeterminate cases into the (determinate) malignant category. The ROM, however, did not meet the human level of 91% (results are not presented). This implies that the proposed algorithm has the potential to be used as a rule-out rather than a rule-in ancillary test, similar to Afirma (Veracyte, Inc., South San Francisco, CA), which is valued primarily as a rule-out test.

In summary, we evaluated our algorithm in single-slide and multislide settings. We demonstrated the use of the algorithm in a fully automated scenario and validated the results on data collected during an entire year at an academic medical center. The algorithm screened the SSR test set and classified 45.1% (130/288) of the scans as determinate (benign or malignant) with human-expert level ROM rates of 2.7% and 94.7%, respectively. The algorithm functioned well as an ancillary test for the same test set and reduced the number of indeterminate cases by 21.3% (from $N = 108$ to 85); the ROM for this reclassified set was 1.8%, well within the clinically acceptable range. These single-slide results were reproduced using two additional test sets. One test set contained multiple slides per FNAB, demonstrating that there is no need to manually select one representative slide. And the other test set comprised consecutive slides collected during the 2018 calendar year, achieving similar clinically acceptable margins of error for thyroid FNAB diagnoses.

In future research, we plan to exploit the potential use of the algorithm to replace genetic sequencing tests, such as Thyroseq (Sonic Healthcare USA ThyroSeq Laboratory, Rye Brook, NY), that are often done in indeterminate cases but are expensive with high false-positive rates.[30] We also plan to evaluate the algorithm in a multicenter study to assess how the algorithm's performance is influenced by differences between scanners and slide preparation techniques across institutions.

References

1. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ: Clinical-grade

computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019, 25:1301−1309
2. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, van der Laak J, Hulsbergen-van de Kaa C, Litjens G: Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol, 2020, 21:233−241
3. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A: Classification and mutation prediction from non−small cell lung cancer histopathology images using deep learning. Nat Med 2018, 24:1559−1567
4. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, Manceron P, Toldo S, Zaslavskiy M, Le Stang N, others: Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med 2019, 25:1519−1525
5. Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, et al: Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nat Med 2020, 26: 52−58
6. Filipczuk P, Fevens T, Krzyżak A, Monczak R: Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. IEEE Trans Med Imaging 2013, 32:2169−2178
7. Pouliakis A, Margari C, Margari N, Chrelias C, Zygouris D, Meristoudis C, Panayiotides I, Karakitsos P: Using classification and regression trees, liquid-based cytology and nuclear morphometry for the discrimination of endometrial lesions. Diagn Cytopathol 2014, 42: 582−591
8. Tosun AB, Yergiyev O, Kolouri S, Silverman JF, Rohde GK: Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. Cytometry A 2015, 87:326−333
9. Panicker RO, Soman B, Saini G, Rajan J: A review of automatic methods based on image processing techniques for tuberculosis detection from microscopic sputum smear images. J Med Syst 2016, 40:1−13
10. Gilshtein H, Mekel M, Malkin L, Ben-Izhak O, Sabo E: Computerized cytometry and wavelet analysis of follicular lesions for detecting malignancy: a pilot study in thyroid cytology. Surgery 2017, 161: 212−219
11. Vaickus LJ, Suriawinata AA, Wei JW, Liu X: Automating the Paris system for urine cytopathology—a hybrid deep-learning and morphometric approach. Cancer Cytopathol 2019, 127:98−115
12. Fragopoulos C, Pouliakis A, Meristoudis C, Mastorakis E, Margari N, Chroniaris N, Koufopoulos N, Delides AG, Machairas N, Ntomi V, Nastos K, Panayiotides IG, Pikoulis E, Misiakos EP: Radial basis function artificial neural network for the investigation of thyroid cytological lesions. J Thyroid Res 2020, 2020:5464787
13. Lu J, Sladoje N, Stark CR, Ramqvist ED, Hirsch J-M, Lindblad J: A deep learning based pipeline for efficient oral cancer screening on whole slide images. Int Conf Image Anal Recognit 2020, 12132: 249−261
14. Sanyal P, Mukherjee T, Barui S, Das A, Gangopadhyay P: Artificial intelligence in cytopathology: a neural network to identify papillary carcinoma on thyroid fine-needle aspiration cytology smears. J Pathol Inform 2018, 9:43
15. Dov D, Kovalsky SZ, Assaad S, Cohen J, Range DE, Pendse AA, Henao R, Carin L: Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. Med Image Anal 2021, 67:101814
16. Landau MS, Pantanowitz L: Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. J Am Soc Cytopathol 2019, 8:230−241
17. Popoveniuc G, Jonklaas J: Thyroid nodules. Med Clin 2012, 96: 329−349
18. Cibas ES, Ali SZ: The 2017 Bethesda system for reporting thyroid cytopathology. Thyroid 2017, 27:1341−1346
19. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L: 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
20. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW: The Bethesda system for reporting thyroid cytopathology: a meta-analysis. Acta Cytol 2012, 56:333−339
21. Faquin WC, Wong LQ, Afrogheh AH, Ali SZ, Bishop JA, Bongiovanni M, Pusztaszeri MP, VandenBussche CJ, Gourmaud J, Vaickus LJ, Baloch ZW: Impact of reclassifying noninvasive follicular variant of papillary thyroid carcinoma on the risk of malignancy in the Bethesda system for reporting thyroid cytopathology. Cancer Cytopathol 2016, 124:181−187
22. Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. ArXiv 2014, [Preprint] doi: 10.48550/arXiv.1409.1556
23. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L: Imagenet: a large-scale hierarchical image database. Edited by 2009 IEEE Conf Comput Vis Pattern Recognit. IEEE, 2009. pp. 248−255
24. Draelos RL, Dov D, Mazurowski MA, Lo JY, Henao R, Rubin GD, Carin L: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Med Image Anal 2021, 67:101857
25. Efron B, Tibshirani RJ: An Introduction to the Bootstrap. Boca Raton, FL: CRC Press, 1994
26. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988, 44:837−845
27. Cibas ES, Ali SZ: The Bethesda system for reporting thyroid cytopathology. Thyroid 2009, 19:1159−1165
28. Range DE, Jiang XS: An update on noninvasive follicular thyroid neoplasm with papillary-like nuclear features. Curr Opin Oncol 2018, 30:1−7
29. Elliott Range D, Jiang XS: Noninvasive follicular thyroid neoplasm with papillary-like nuclear features and the risk of malignancy in the Bethesda system for the reporting of thyroid cytopathology. Diagn Cytopathol 2020, 48:531−537
30. Balentine CJ, Vanness DJ, Schneider DF: Cost-effectiveness of lobectomy versus genetic testing (Afirma) for indeterminate thyroid nodules: considering the costs of surveillance. Surgery 2018, 163: 88−96