Use of Machine Learning–Based Software for the Screening of Thyroid Cytopathology Whole Slide Images

David Dov, PhD; Shahar Z. Kovalsky, PhD; Qizhang Feng, MS; Serge Assaad, BS; Jonathan Cohen, MD; Jonathan Bell, MD; Ricardo Henao, PhD; Lawrence Carin, PhD; Danielle Elliott Range, MD

• *Context.*—The use of whole slide images (WSIs) in diagnostic pathology presents special challenges for the cytopathologist. Informative areas on a direct smear from a thyroid fine-needle aspiration biopsy (FNAB) smear may be spread across a large area comprising blood and dead space. Manually navigating through these areas makes screening and evaluation of FNA smears on a digital platform time-consuming and laborious. We designed a machine learning algorithm that can identify regions of interest (ROIs) on thyroid fine-needle aspiration biopsy WSIs.

***Objective.*—To evaluate the ability of the machine learning algorithm and screening software to identify and screen for a subset of informative ROIs on a thyroid FNA WSI that can be used for final diagnosis.**

***Design.*—A representative slide from each of 109 consecutive thyroid fine-needle aspiration biopsies was scanned. A cytopathologist reviewed each WSI and recorded a diagnosis. The machine learning algorithm screened and selected a subset of 100 ROIs from each WSI to present as an image gallery to the same cytopathologist after a washout period of 117 days.**

***Results.*—Concordance between the diagnoses using WSIs and those using the machine learning algorithm–generated ROI image gallery was evaluated using pairwise weighted κ statistics. Almost perfect concordance was seen between the 2 methods with a κ score of 0.924.**

***Conclusions.*—Our results show the potential of the screening software as an effective screening tool with the potential to reduce cytopathologist workloads.**

(*Arch Pathol Lab Med.* 2022;146:872–878; doi: 10.5858/arpa.2020-0712-OA)

Thyroid nodules are common, and an estimated 10% of the US population will have one in their lifetime. The majority of thyroid nodules are benign.[1] The routine evaluation of a thyroid nodule includes ultrasound imaging. Nodules that meet certain radiologic and clinical criteria are sampled, typically using a fine-needle aspiration biopsy (FNAB) technique.[2] The Bethesda System for the Reporting of Thyroid Cytopathology (TBS) comprises 6 diagnostic categories to classify thyroid FNABs, and each is associated with its own risk of malignancy (ROM) based on the surgical pathology results. The TBS categories (and their ROMs) are nondiagnostic (1%–4%), benign (BN; 0%–3%), atypia of undetermined significance (AUS; 10%–30%), follicular neoplasm (FN; 25%–40%), suspicious for malignancy (SUSP; 50%–75%), and malignant (MAL; 97%–99%).[3] Direct smears are made from cellular material obtained from the FNAB procedure by placing the specimen on a glass microscope slide and using mechanical pressure to spread the cells across the slide. The diagnostic material is typically present as single cells and small groups of cells, with relatively large intervening areas of blood, serum, colloid, and empty space. The screening and review of direct smears on a digital platform requires manually navigating through these acellular areas of a smear to find the cells of interest. This process can be time-consuming and laborious.[4,5] As the use of whole slide images (WSIs) becomes more popular for primary diagnosis in pathology, the cytopathologist will have to navigate such challenges. The use of computational approaches in the digital cytopathology arena can assist in overcoming some of these challenges.[6]

We designed and implemented a machine learning–based software that summarizes WSIs by generating an image gallery of automatically identified regions of interest (ROIs) containing follicular cells. Summarization is often used in the computer vision literature to describe algorithms that distill important information from large-volume data.[7] In this study, we investigate the adequacy of our summarization screening software in identifying diagnostic ROIs to assist in the digital screening and diagnosis of thyroid cytopathology.

Accepted for publication July 8, 2021.

Published online October 20, 2021.

From the Departments of Electrical and Computer Engineering (Dov, Assaad, Carin) and Biostatistics and Bioinformatics (Henao), Duke University, Durham, North Carolina; the Department of Mathematics, University of North Carolina at Chapel Hill (Kovalsky); the Department of Computer Science & Engineering, Texas A&M University, College Station (Feng); the Department of Head and Neck Surgery, Kaplan Medical Center, Rehovot, Israel (Cohen); and the Department of Pathology (Bell, Range), Duke University Medical Center, Durham, North Carolina. Carin is currently located in the Office of the Provost at King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia.

The authors have no relevant financial interest in the products or companies described in this article.

Corresponding author: Danielle Elliott Range, MD, Department of Pathology, Duke University Medical Center, 40 Duke Medicine Cir, Box 3712, Durham, NC 27705 (email: danielle.range@duke.edu).

METHODS

After obtaining institutional review board approval, we searched our institutional files for all thyroidectomy specimens with a preceding FNAB from January 2008 to June 2016. Initial exclusions included nondiagnostic FNABs and thyroidectomies diagnosed as noninvasive thyroid follicular neoplasms with papillary-like nuclear features, which could not be placed in a benign or malignant category for study purposes.[8] Cases for which the biopsied nodule could not be correlated with the final surgical pathology result were also excluded. Fine-needle aspiration biopsies at our institution are prepared using both air-dried Diff-Quik–stained and alcohol-fixed Papanicolaou–stained slides. Rapid on-site assessments are variably performed using only Diff-Quik–stained slides. One alcohol-fixed, Papanicolaou-stained, direct smear from each FNAB procedure was selected for scanning. The selected slide represented the slide with the most follicular groups, regardless of associated clot, air bubbles, or other preanalytical artifacts. All slides were cleaned to remove pen marks and scanned with a ×40 objective at 9 focal planes. The WSIs were acquired as SVS files using a Leica AT-2 scanner. All cytologic (TBS) diagnoses were recorded for each nodule as documented in the electronic medical record (EMR). The final surgical pathology result was used as the ground truth and also recorded. The WSIs were divided into a training set and a test set. The test set slides comprised a subset of consecutive FNABs that were not analyzed by the machine learning algorithm (MLA) during training.

We used the training set to design the screening software, which has 2 parts: an MLA and a graphical user interface for the end user. The MLA itself comprises 2 components. The first is a screening MLA, which is based on a convolutional neural network and is designed to identify ROIs, that is, patches (image regions) containing follicular groups. To train this component of the MLA, we used a (fully) supervised learning method in which a cytopathologist used Aperio ImageScope software (Leica Biosystems, Inc) to annotate 4494 ROIs containing follicular cells on a subset of 145 WSIs from the training set. Because the vast majority of a slide does not contain nucleated cellular material, regions selected at random have a high likelihood of being uninformative. We used these randomly selected areas on the slides as examples of uninformative (ie, nondiagnostic) ROIs for training purposes. The screening MLA is based on VGG11 convolutional neural network architecture (implemented in PyTorch 0.4.1). The convolutional filters were initialized with parameters pretrained on ImageNet, which is a large and widely used data set in computer vision.[9]

After training the screening MLA, we used it to identify 1000 of the most informative ROIs on each WSI. These ROIs were used to train the second component of the MLA, termed the classifier MLA. The WSIs from the training set were labeled as benign or malignant, based on the final surgical pathology results, which were also used as the ground truth. The classifier MLA was trained to predict malignancy at the slide level, distinct from the ROI-level prediction.[10] In addition to the final pathology, the classifier MLA was trained to simultaneously predict the TBS category via an ordinal regression framework. The joint prediction serves as a regularization for the training process, providing improved classification accuracy. We refer the reader to Dov et al[11] for a detailed description of the training process.

In the testing phase, the screening MLA was tasked with identifying only the 100 most informative ROIs. These, in turn, were used by the classifier MLA for the prediction of malignancy. Specifically, the predicted malignancy label of the WSI was obtained by averaging the local ROI-level predictions. We found no advantage in using more than 100 ROIs for the automated malignancy prediction; these same 100 ROIs were also used in the image gallery presented to the reviewer. For the current paper, only the malignancy prediction of the final pathology, not the prediction of the TBS category, was used. Our previously published data show that the sensitivity and specificity of the MLA to predict malignancy on WSIs were 92% and 91%, respectively. For additional details of our training set, the engineering design, and thorough analysis of the performance of the MLA, we refer to our previous publications.[11–13]

The second part of our screening software is the graphical user interface, represented by a screenshot in Figure 1. When the software presents an FNAB for analysis, an image gallery is created that displays 100 ROIs automatically identified by the screening MLA as the most informative ROIs. These 100 ROIs correspond to 0.2% of the area on the slide. Typical ROIs have an average of 1 to 2 follicular groups, whereas larger groups may span more than one ROI. The individual ROIs are displayed at the equivalent of ×40 objective magnification. Each ROI includes a z-stack of 9 focal planes; the user may select a given ROI and use a mouse to scroll through the focal planes to focus the image. Only one z-stack was used to train the MLA; in isolated experiments, we found no difference in MLA performance with the use of additional z-stacks. For optimal viewing, when the user clicks on an ROI, the user is shown a composite, side-by-side view of the ROI at ×10 and ×40 magnification. In addition, the graphical user interface includes a view of the WSI at a ×1 objective magnification. An internal timer records the time to diagnosis for each case.

The screening software has the ability to integrate the selected ROIs into the existing WSI viewing software (ie, ImageScope). This integration works by marking the ROIs with bounding boxes, and it allows the user to navigate from one ROI to another simply by clicking on designated cells in a built-in annotation pane (Figure 2). This navigation feature facilitates more rapid review of the WSI by eliminating the need to scroll through uninformative areas of the WSI.

To assess the potential of the screening software as an assistive screening tool, we had an experienced board-certified cytopathologist blindly review and assign a TBS category to each WSI in the test set using Aperio ImageScope software (without the ROIs being presented). Using the image gallery created by the software, the same cytopathologist reviewed the test set 117 days later. The reviewer's WSI cytologic diagnoses (WSI-TBS), ROI-based diagnoses using the image gallery (ROI-TBS), and the EMR diagnoses were recorded for each FNAB. The final surgical pathology result documented in the EMR was categorized as benign or malignant and was used as the ground truth for each WSI. Concordance was calculated using pairwise weighted κ statistics.[14] Agreement was categorized as follows: 0 to 0.2, slight; 0.21 to 0.4, fair; 0.41 to 0.6, moderate; 0.61 to 0.8, substantial; and 0.81 to 1, almost perfect.[15] We used κ scores to evaluate concordance across TBS categories and across 3 malignancy risk groups: benign, intermediate risk, and high risk. The benign risk group included BN cytologic cases, the intermediate risk group included cases diagnosed as AUS and FN, and the high-risk group included the SUSP and MAL categories. Concordance between the WSI-TBS and the EMR and between the WSI-TBS and the ROI-TBS was assessed for both the TBS categories and the risk groups. The accuracy of predicting malignancy by the classifier MLA, the reviewer (using both methods), and the EMR was evaluated using receiver operating characteristic curves.

RESULTS

The cohort comprised a total of 908 FNABs; 799 FNABs were used for training of the MLA and 109 consecutive FNABs were used for the test set. Table 1 shows the distribution of the test set cases (n = 109) by cytologic diagnosis for the EMR and for the reviewer using WSIs and the ROI-based method; the number of malignant cases in each category is included along with the calculated risk of malignancy. Eighty-four cases were benign on final pathology and 25 were malignant. Table 2 summarizes the κ statistics of the 2 methods and the EMR diagnoses. When compared with the EMR, the WSI-TBS showed almost perfect concordance ($\kappa = 0.845$) across TBS categories and substantial agreement when restricted to just the risk groups ($\kappa = 0.669$). Intraobserver agreement between the ROI-TBS

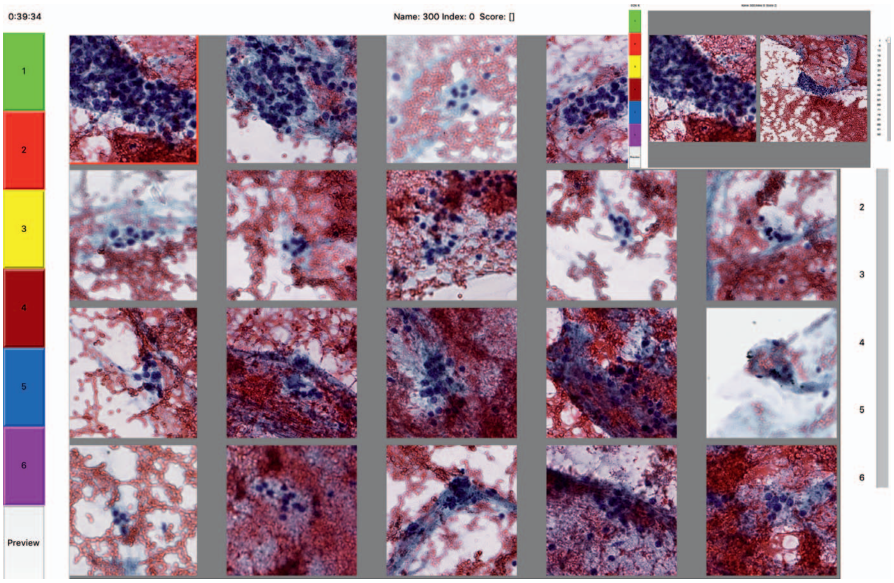


Figure 1. *Graphical user interface with image gallery. Twenty of the 100 region-of-interest (ROI) images (×40) selected by the screening machine learning algorithm are displayed at one time. Buttons labeled 1 through 6 (left side) can be clicked to record the Bethesda diagnosis based on ROIs. Elapsed time is displayed at top left. Any ROI can be selected to view the ×40 and ×10 views side by side (inset, top). A preview button (bottom left) returns the user to the image gallery. The right-side scroll bar navigates through all 5 sets of gallery images and the sixth position navigates to the ×1 view of the whole slide image (not shown).*

and the WSI-TBS across TBS categories and across the risk groups was almost perfect, yielding κ = 0.924 and κ = 0.834, respectively.

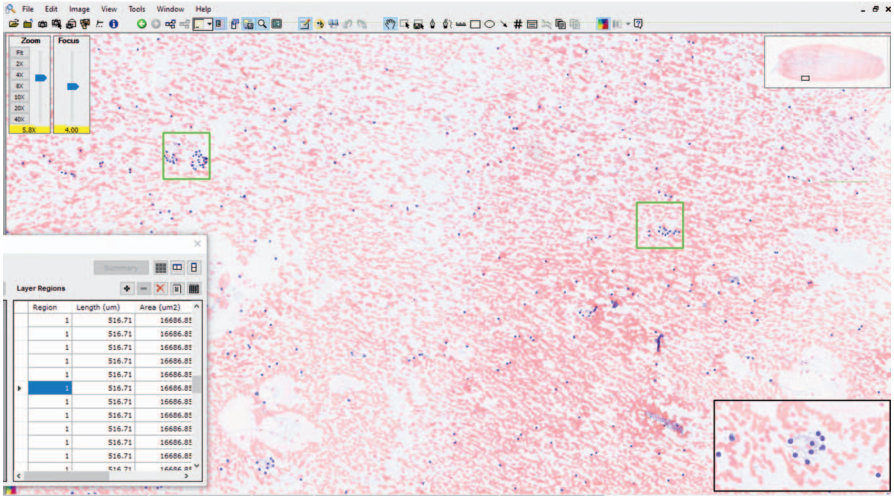
A total of 23 of the 109 cases (21.1%) were discordant between the WSI-TBS and ROI-TBS methods, and moved out of their original WSI-TBS category when the ROI-based method was used (Table 3). Sixteen cases moved between AUS and BN, and 5 moved from FN to AUS. The remaining 2 discordant cases were malignant and moved more than 1 level on ROI review, from FN to BN and SUSP to AUS. Among the 23 discordant cases, 14 (60.8%) were downgraded on ROI-based review and 9 (39.1%) were upgraded. Five of the 14 downgraded cases (35.7%) were malignant: 2 were downgraded from FN to AUS, 1 from SUSP to AUS, and the remaining 2 (14.3%) were inappropriately downgraded to BN from AUS and SUSP; both were follicular variant of papillary thyroid carcinoma on final pathology. All 9 of the upgraded cases moved from BN to AUS on ROI-based review; 1 of these cases was malignant.

The average time to diagnosis for ROI-TBS was 81.59 seconds (1.36 minutes) per case. Four cases had a time to diagnosis in excess of 1500 seconds and were excluded from this calculation because they were likely the result of an error from inadvertently leaving the software application open before recording the diagnosis.

Using the final pathology as the ground truth for the test set, the performances of the EMR and the reviewer in predicting malignancy were evaluated using a receiver operating characteristic curve. The areas under the curve for the EMR, WSI-TBS, and ROI-TBS were 0.931, 0.931, and 0.896, respectively. The performance of the classifier MLA (previously reported) yielded an area under the curve of 0.932 on the same test set.[11–13]

DISCUSSION

In a previous study,[11–13] we developed an MLA that demonstrated performance comparable with humans in the prediction of malignancy on thyroid FNABs. The current study tested the efficacy of this MLA to screen WSIs of thyroid FNABs for diagnostic follicular groups using an image gallery. The TBS diagnoses made by the reviewer using only the ROI selection automated by the software showed almost perfect concordance with TBS diagnoses made by the same cytopathologist on manual review of the



Region	Length (um)	Area (um2)
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85
1	516.71	16686.85

Figure 2. *Screenshot of Aperio ImageScope software with integrated regions of interest (green boxes) automatically identified by the screening machine learning algorithm. Clicking a cell on the annotation pane (bottom left) allows the user to quickly navigate to the next region of interest.*

Table 1. Distribution of the Test Set Cases (n = 109) and Malignant Cases (n = 25) Across Bethesda System for the Reporting of Thyroid Cytopathology (TBS) Categories as Diagnosed by the Electronic Medical Record (EMR) Pathologists and by the Reviewer Using Whole Slide Images (WSIs) and Regions of Interest (ROIs)

TBS Category	EMR Diagnosis, No. (%)	Malignant Cases, No. (ROM %)	WSI Diagnosis, No.	Malignant Cases, No. (ROM %)	ROI Diagnosis, No.	Malignant Cases, No. (ROM %)
BN	50 (45.9)	0 (0)	67	2 (3.0)	66	3 (4.5)
AUS	32 (29.4)	6 (18.8)	16	2 (12.5)	24	5 (20.8)
FN	9 (8.3)	3 (33.3)	12	7 (58.3)	6	4 (66.7)
SUSP	6 (5.5)	4 (66.7)	1	1 (100)	0	0 (0)
MAL	12 (11)	12 (100)	13	13 (100)	13	13 (100)

Abbreviations: AUS, atypia of undetermined significance; BN, benign; FN, follicular neoplasm; MAL, malignant; ROM, risk of malignancy; SUSP, suspicious for malignancy.

WSI. With refinement, this screening software can be used to screen WSIs of thyroid FNABs.

The selection of the 100 most informative regions of interest across a WSI of a thyroid FNAB was automated by the MLA and used by the ROI-based software. Our experiments show strong concordance between the WSI-TBS diagnosis and the ROI-TBS ($\kappa = 0.924$). A large majority (78.9%; 86 of 109) of the diagnoses remained unchanged between the 2 methods. The greatest change in number of cases was seen in the AUS category; 8 cases were added to that category with the ROI-based method, increasing the number of AUS cases from 16 to 24. The increased number of AUS cases suggests that the MLA may have presented the reviewer with more atypical regions. The fact that 9 of the 24 cases moved from BN to AUS based on ROIs also supports this theory. These 9 cases represent all of the discordant cases that were upgraded to AUS, and they appropriately include 1 malignant case. So, although the atypical rate may have increased, it had the added value of identifying a malignant case that was originally categorized as BN. Finding the most atypical groups is the overall goal of a screening tool, in which the threshold for atypia may be lower in order to ensure atypia is not missed and all positive cases are identified.

Given the differences noted among the discrepant cases, we wondered how collapsing the categories into risk groups, based on similar ROMs and clinical management, might affect concordance. BN remained as a separate category. The AUS and FN categories were combined because of their overlapping malignancy rates, 10%–30% and 25%–40%, respectively; similar cytologic features; and shared surgical pathology results.[3] In addition, their diagnoses may result in similar clinical management, including additional testing and/or surgical treatment.[2] The SUSP and MAL diagnoses were combined for similar reasons.[3] Concordance between the WSI- and ROI-based methods across all the risk groups remained high ($\kappa = 0.834$), but decreased. The overall number of intermediate risk cases (AUS and FN) was similar between the ROI-TBS and WSI-TBS methods (n = 30 versus n = 28, respectively). But the ROI-based intermediate risk group contained a higher proportion of AUS cases (80% versus 57%). As indicated above, 9 such cases moved from the BN category/risk group and likely explains why the κ score for the risk groups was lower than for the individual TBS categories.

More importantly, when one considers the concordance between WSI-TBS and EMR-TBS as a measure of interobserver agreement, we see that the ROI-based method is excellent ($\kappa = 0.845$) in eliciting the same diagnosis as the EMR. However, there are much smaller differences in diagnoses between WSI-TBS and ROI-TBS, as represented by the intraobserver agreement ($\kappa = 0.924$). This finding supports the implication that the use of the ROIs is a very effective screening tool when compared with the baseline performance of the interobserver agreement.

Among the 14 cases that were downgraded using the image gallery, 64.3% (n = 9) were benign and appropriately downgraded to either BN or AUS. However, the remaining 5 cases were malignant, and 2 of these were inappropriately downgraded to BN. Interestingly, all 5 of these downgraded cases were predicted to be malignant by the classifier MLA. Given this fact, it was unclear why the automated ROIs elicited a BN diagnosis for 2 cases by the reviewer. We reviewed the ROIs and WSIs from both of these cases, and in retrospect, the ROI-TBSs were reviewer errors. Both cases show sufficient atypia in the selected ROIs that the diagnosis should have been at least AUS.

Accuracy in predicting malignancy was comparable between the 2 methods, though reduced for the ROI-based

Table 2. Diagnostic Concordance Between Whole Slide Image (WSI)– and Region of Interest (ROI)–Based Diagnoses With Diagnostic Accuracy

	WSI-TBS		
	TBS Categories κ	Risk Groups κ	ROC Curve AUC
EMR	0.845	0.669	0.931
ROI-TBS	0.924	0.834	0.896
WSI-TBS	. . .	. . .	0.931

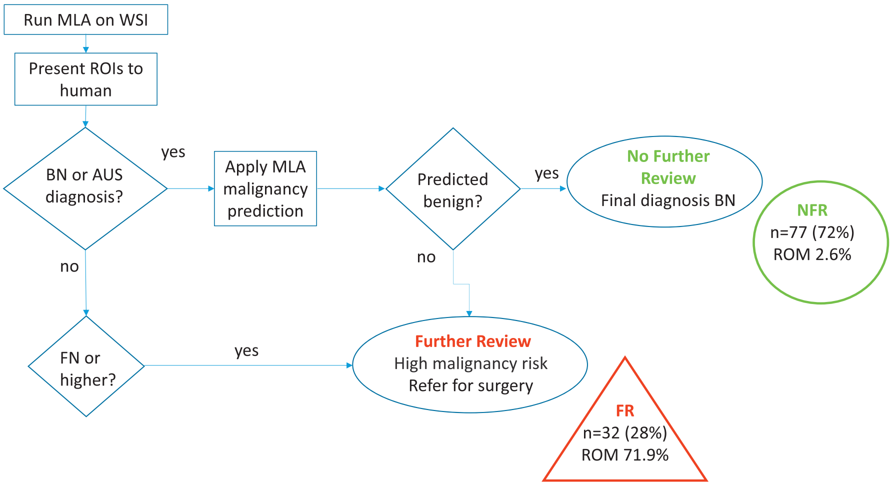
Abbreviations: AUC, area under the curve; EMR, electronic medical record; ROC, receiver operating characteristic; ROI-TBS, TBS diagnosis based on ROIs; TBS, the Bethesda System for the Reporting of Thyroid Cytopathology; WSI-TBS, TBS diagnosis based on WSI.

Table 3. Bethesda System for the Reporting of Thyroid Cytopathology (TBS) Categories Based on Whole Slide Images (WSIs) Versus Regions of Interest (ROIs)

WSI-Based TBS Diagnosis	ROI-Based TBS Diagnosis				
	Benign	Atypical	FN	Suspicious	Malignant
Benign	58	9	0	0	0
Atypical	7	9	0	0	0
FN	1	5	6	0	0
Suspicious	0	1	0	0	0
Malignant	0	0	0	0	13

Abbreviation: FN, follicular neoplasm.

Figure 3. *An example of the proposed workflow for use of the machine learning algorithm (MLA) and image gallery as an assistive tool for pathologists. (1) The whole slide image (WSI) is analyzed by the screening MLA selecting 100 most informative regions of interest (ROIs). (2) The ROIs are presented to the human screener (eg, cytotechnologist or trainee). (3) The classifier MLA predicts malignancy, which is provided to the human screener after review of the ROIs. The screener would independently review the 100 ROIs and provide a cytologic diagnosis. This Bethesda diagnosis based on ROIs is combined with the MLA malignancy prediction along a decision tree to determine if further review is required by a cytopathologist. Abbreviations: AUS, atypia of undetermined significance; BN, benign; FN, follicular neoplasm; FR, further review; NFR, no further review.*



method, with areas under the curve of 0.896 and 0.931. An ideal screening tool would identify all true-positive cases while limiting the number of false positives. So, we conducted an experiment to test the potential use of the image gallery as a screening tool to be used along with the MLA final pathology predictions. The workflow we propose below leverages our knowledge of the MLA's ability to reduce indeterminate diagnoses when combined with human decisions[13] and mitigates the limitations seen in using the image gallery with respect to the inappropriately downgraded cases.

Figure 3 demonstrates the proposed workflow in which the WSI would be analyzed by the MLA for automated selection of the 100 most informative ROIs, which would then be presented to the human screener (eg, cytotechnologist or trainee). At the same time, a prediction of malignancy would be made by the MLA and provided to the human screener after review of the ROIs. The screener would independently review the 100 ROIs and provide a cytologic diagnosis. This ROI-TBS could then be combined with the MLA malignancy prediction along a decision tree that would separate the cases into 2 groups: those that require further review (FR) by a cytopathologist and those that do not. Further review might include a review of (1) additional ROIs on the WSI using the navigation feature of the software or (2) the glass slide.

When we applied this digital workflow, 32 cases were in the FR category, which yielded a ROM of 71.9% and included 23 malignant and 8 benign cases (Table 4). The remaining 77 cases that fell into the no FR (NFR) category included only 2 malignant cases. This digital workflow essentially divides the cases into 2 screening groups: (1) a high-risk group requiring FR for diagnostic confirmation, with a calculated ROM (71.9%) between that reported for the SUSP and MAL TBS categories, and (2) a low-risk group (NFR) with a calculated ROM (2.6%) equivalent to that reported for the BN TBS category.[3]

Table 4. Distribution of Cases Using a Screening Workflow That Combines Region of Interest (ROI)–Based Diagnoses and Machine Learning Algorithm (MLA) Malignancy Predictions

	ROI + MLA Diagnosis, No. (n = 109)	Malignant Cases, No. (ROM %) (n = 25)
No further review	77	2 (2.6)
Needs further review	32	23 (71.9)

Abbreviation: ROM, risk of malignancy.

Although the ROM for the NFR category is higher than that for the BN category in the EMR (Table 1), it comes with some benefits. First, the 27 additional cases that require NFR can aid in workload reductions for cytopathologists. In addition, the binary result of FR and NFR effectively eliminates the indeterminate category. This could also result in easier clinical management; with such a high malignancy rate in the FR category (71.9%), one may consider direct referral to surgery. The ROM can also be increased in this group if FR is done using the WSI or glass slide or by soliciting a second opinion to successfully weed out the benign cases in this category. As previously reported, a review of the only 2 malignant cases in the NFR category revealed a slide selection bias. Both of these cases were predicted as benign by the classifier MLA and diagnosed as BN when reviewed by 2 additional cytopathologists, suggesting the selected slide was not representative of the lesion.[13] The EMR pathologist categorized these 2 cases as AUS and FN. We believe this workflow could serve as a means to screen and classify the majority of cases as BN, potentially reducing workload, while reducing the atypia and indeterminate rates using a combination of human and machine predictions. Further study and validation of this process with a larger cohort and reviewers with various levels of experience will be needed in the future.

Studying the time to diagnosis was not a goal of this project, but we noted the average time was 1.36 minutes per case. As a rough comparison, Hanna et al[16] reported an average time to diagnosis of 1.6 minutes for each nongynecologic WSI from various specimen sources. We, like others, found the ability to view the ROIs in the visual context of their surroundings on the WSI helpful for diagnostic interpretation.[16] Use of the navigation feature in our screening software, alone or as a tool for FR, allows the reviewer to view ROIs across the WSI more quickly and in the context of their surroundings, saving screening time and time to diagnosis. Unlike other systems, the screening software automates the selection of relevant ROIs for the image gallery, and requires no human input in this initial process.[17–22]

There are some limitations to the current study. The obvious one is our small test set. Although we believe the proportion of cases in each TBS category is representative of our patient population, small variations and discrepancies in our data are difficult to extrapolate to a larger study. There is an inherent bias in this study, as the training of the screening and classifier MLAs was partially based on supervised learning performed by the study reviewer. However, these supervised learning instances included annotated ROIs from only 145 WSIs (18.1%) of the total 799 used for training, and did not include any cases from the test set.

Other limitations to our study are likely a result of limitations in the software. The reviewer had no way to evaluate particular follicular groups in the context of their surroundings—for example, the ability to compare follicular groups to other neighboring groups or in the context of regional artifact such as air drying. In addition, the reviewer could not view additional ROIs if desired. And lastly, identification of more subtle examples of colloid or lymphocytes is limited because both may be located between the ROIs. This may also explain the increase in AUS ROI-TBS diagnoses. Five cases of chronic thyroiditis (21.7%) were among the 23 discordant cases and represented a third of the discordant cases in the AUS ROI-TBS category. The image gallery was designed to present only the ROIs to evaluate their adequacy for diagnosis, but these limitations can be overcome with use of the navigation feature in the software when WSI review is deemed necessary.

Finally, the screening software may not have consistently identified the most atypical groups. This was illustrated by the 2 cases that were inappropriately downgraded from AUS and FN to BN by the reviewer. The MLA predicted these cases to be malignant and one would have expected the automated ROIs to reflect this prediction. The WSI-TBS for both cases was concordant with that of at least one of the other reviewers who reviewed the discordant cases. This suggests that the automated ROIs may not have adequately represented the WSI in rare cases. However, retrospective review of the ROIs in these 2 cases showed enough cytologic atypia that the reviewer should have classified them as, at least, AUS. Because 5 of the 6 malignant cases in the discordant group were predicted as malignant by the classifier MLA, combining the MLA malignancy predictions with the reviewer screening results can mitigate this limitation. Combining the 2 using the proposed workflow results in a noninferior method to classify the BN cases (Table 3).

The ideal clinical implementation of this technology might begin with a custom graphical user interface that is directly linked to the WSI and the navigation feature. A screener, such as a cytotechnologist or trainee, would view the automated ROIs. If desired, the screener could use the navigation feature to evaluate select ROIs in order to glean additional information such as colloid or lymphocytes. Once satisfied with the initial evaluation, the screener would assign the TBS category and then follow the proposed workflow indicated in Figure 3. If the result led to FR, the case would be flagged for review by a pathologist. The pathologist would then have the option to use any or all of the available modalities to render a final TBS diagnosis, including use of the navigation feature to assist in the review of the WSI. This practical implementation of the screening software as an assistive tool for pathologists would result in fewer cases for the pathologist to review and allow for more time to be spent on challenging cases. In addition, the pathologist would have the benefit of the malignancy prediction of the classifier MLA to aid in arriving at the final cytologic diagnosis.

In conclusion, using an MLA we created software that automatically generates an image gallery for the screening and identification of ROIs for thyroid FNAB WSIs. We used this image gallery to assess whether the detected ROIs were sufficient to render a TBS diagnosis. Our results demonstrated almost perfect concordance between TBS diagnoses made using the image gallery and those based on the WSI alone. Our results suggest that the screening software can be used to effectively screen thyroid FNABs. We believe this software, in conjunction with the MLA malignancy predictions and additional navigation features, can reduce indeterminate rates in thyroid FNAB diagnoses, aid in the classification and triage of FNABs, improve the accuracy of the screening process, and help reduce pathologist workloads on digital platforms.

References

1. Dean DS, Gharib H. Epidemiology of thyroid nodules. *Best Pract Res Clin Endocrinol Metab*. 2008;22(6):901–911.

2. 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):1–133.

3. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. *Thyroid*. 2017;27(11):1341–1346.

4. Girolami I, Marletta S, Pantanowitz L, et al. Impact of image analysis and artificial intelligence in thyroid pathology, with particular reference to cytological aspects. *Cytopathology*. 2020;31(5):432–444.

5. Evered A, Dudding N. Accuracy and perceptions of virtual microscopy compared with glass slide microscopy in cervical cytology. *Cytopathology*. 2011; 22(2):82–87.

6. Landau MS, Pantanowitz L. Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. *J Am Soc Cytopathol*. 2019;8(4):230–241.

7. Potapov D, Douze M, Harchaoui Z, Schmid C. Category-specific video summarization. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. *European Conference on Computer Vision*. Zurich, Switzerland: Springer International Publishing; 2014:540–555. *Lecture Notes in Computer Science*; vol 8694.

8. Seethala RR, Baloch ZW, Barletta JA, et al. Noninvasive follicular thyroid neoplasm with papillary-like nuclear features: a review for pathologists. *Mod Pathol*. 2018;31(1):39–55.

9. Deng J, Dong W, Socher R, Li L, Kai L, Li F. ImageNet: a large-scale hierarchical image database. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20–25, 2009; Miami, FL.

10. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal*. 2019;54:280–296.

11. Dov D, Kovalsky SZ, Assaad S, et al. Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. *Med Image Anal*. 2021;67:101814.

12. Dov D, Kovalsky SZ, Cohen J, Range DE, Henao R, Carin L. Thyroid cancer malignancy prediction from whole slide cytopathology images. In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. 2019;106:553–570. *Proceedings of Machine Learning Research*; vol 106.

13. Elliott Range DD, Dov D, Kovalsky SZ, Henao R, Carin L, Cohen J. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. *Cancer Cytopathol*. 2020;128(4):287–295.

14. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.

15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.

16. Hanna MG, Monaco SE, Cuda J, Xing J, Ahmed I, Pantanowitz L. Comparison of glass slides and various digital-slide modalities for cytopathology screening and interpretation. *Cancer Cytopathol*. 2017;125(9):701–709.

17. Gopinath B, Shanthi N. Development of an automated medical diagnosis system for classifying thyroid tumor cells using multiple classifier fusion. *Technol Cancer Res Treat*. 2015;14(5):653–662.

18. Wright AM, Smith D, Dhurandhar B, et al. Digital slide imaging in cervicovaginal cytology: a pilot study. *Arch Pathol Lab Med*. 2013;137(5):618–624.

19. Chantziantoniou N, Mukherjee M, Donnelly AD, Pantanowitz L, Austin RM. Digital applications in cytopathology: problems, rationalizations, and alternative approaches. *Acta Cytol*. 2018;62(1):68–76.

20. Collins BT, Collins LE. Assessment of malignancy for atypia of undetermined significance in thyroid fine-needle aspiration biopsy evaluated by whole-slide image analysis. *Am J Clin Pathol*. 2013;139(6):736–745.

21. Chain K, Legesse T, Heath JE, Staats PN. Digital image-assisted quantitative nuclear analysis improves diagnostic accuracy of thyroid fine needle aspiration cytology. *Cancer Cytopathol*. 2019;127(8):501–513.

22. 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(1).