Original Article

Application of a Machine Learning Algorithm to Predict Malignancy in Thyroid Cytopathology

Danielle D. Elliott Range, MD [1]; David Dov, PhD[2]; Shahar Z. Kovalsky, PhD[3]; Ricardo Henao, PhD[2,4];
Lawrence Carin, PhD[2]; and Jonathan Cohen, MD, MHS[5]

BACKGROUND: The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) comprises 6 categories used for the diagnosis of thyroid fine-needle aspiration biopsy (FNAB). Each category has an associated risk of malignancy, which is important in the management of a thyroid nodule. More accurate predictions of malignancy may help to reduce unnecessary surgery. A machine learning algorithm (MLA) was developed to evaluate thyroid FNAB via whole slide images (WSIs) to predict malignancy. **METHODS:** Files were searched for all thyroidectomy specimens with preceding FNAB over 8 years. All cytologic and surgical pathology diagnoses were recorded and correlated for each nodule. One representative slide from each case was scanned to create a WSI. An MLA was designed to identify follicular cells and predict the malignancy of the final pathology. The test set comprised cases blindly reviewed by a cytopathologist who assigned a TBSRTC category. The area under the receiver operating characteristic curve was used to assess the MLA performance. **RESULTS:** Nine hundred eight FNABs met the criteria. The MLA predicted malignancy with a sensitivity and specificity of 92.0% and 90.5%, respectively. The areas under the curve for the prediction of malignancy by the cytopathologist and the MLA were 0.931 and 0.932, respectively. **CONCLUSIONS:** The performance of the MLA in predicting thyroid malignancy from FNAB WSIs is comparable to the performance of an expert cytopathologist. When the MLA and electronic medical record diagnoses are combined, the performance is superior to the performance of either alone. An MLA may be used as an adjunct to FNAB to assist in refining the indeterminate categories. ***Cancer Cytopathol* 2020;128:287-295**. *© 2020 American Cancer Society*.

KEY WORDS: Bethesda System for Reporting Thyroid Cytopathology; machine learning; malignancy prediction; neural network; thyroid fine-needle aspiration (FNA).

INTRODUCTION

Thyroid nodules are not uncommon, and an estimated 10% of the general population in the United States are expected to develop a thyroid nodule in their lifetime. Despite this prevalence, the risk of malignancy (ROM) among all thyroid nodules ranges from only 7% to 15%.[1,2] Fine-needle aspiration biopsy (FNAB) is the widely accepted modality for the evaluation of thyroid nodules. The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) is a standardized system used for the diagnosis of thyroid FNAB. TBSRTC comprises 6 diagnostic categories, each associated with its own ROM: nondiagnostic, benign (BN), atypia of undetermined significance or follicular lesion of undetermined significance (AUS/FLUS), follicular neoplasm (FN), suspicious for malignancy (SUSP), and malignant (MAL).[3] This associated ROM significantly affects the management of thyroid nodules.[4] There is an estimated surgical excision rate of 20% to 25% for all thyroid nodules, but less than half of them will be malignant.[5] This results in unnecessary surgery for a large number of patients with thyroid nodules. Also, the bulk of such cases fall within the indeterminate

Corresponding Author: Danielle D. Elliott Range, MD, Department of Pathology, Duke University School of Medicine, 40 Duke Medicine Cir, DUMC Box 3712, Durham, NC 27710 (danielle.range@duke.edu).

[1]Department of Pathology, Duke University School of Medicine, Durham, North Carolina; [2]Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, North Carolina; [3]Department of Mathematics, Trinity College of Arts and Sciences, Duke University, Durham, North Carolina; [4]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina; [5]Department of Head and Neck Surgery and Communication Sciences, Duke University School of Medicine, Durham, North Carolina

Received: October 21, 2019; **Revised:** December 11, 2019; **Accepted:** December 12, 2019

Published online February 3, 2020 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cncy.22238, wileyonlinelibrary.com

TBSRTC categories, which include atypia of undetermined significance (AUS), FN, and SUSP. We created a machine learning algorithm (MLA) to analyze whole slide images (WSIs) of thyroid FNAB slides for the detection of regions of interest (ROIs) and the prediction of malignancy and to assist in the refinement of the indeterminate TBSRTC categories.

MATERIALS AND METHODS

After obtaining institutional review board approval, we searched the institutional files for all thyroidectomy specimens with a preceding FNAB from January 2008 to June 2016. Initial exclusions included the following: cases with nondiagnostic FNAB and cases with equivocal final pathology diagnoses. For example, final pathology results that included noninvasive thyroid FNs with papillary-like nuclear features or uncertain malignant potential were excluded because cases with this diagnosis cannot be placed in a benign or malignant category.[6,7] Similarly, those cases for which the biopsied nodule could not be correlated with the final surgical pathology results were also excluded. 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 clots, air bubbles, or other artifacts. All slides were cleaned and scanned at a 40× objective focal plane. The WSIs were acquired as SVS files with a Leica AT-2 scanner. All cytologic and surgical pathology diagnoses were recorded for each nodule as documented in the electronic medical record (EMR). A random subset of consecutive FNABs was designated as the test set to be exclusively used for evaluating the performance of the proposed approach. An experienced head and neck pathologist and board-certified cytopathologist (D.R.) blindly reviewed the WSIs in the test set and provided a TBSRTC category for each. The remaining cases were used for training of the MLA.

We designed an MLA based on 2 convolutional neural networks (CNNs),[8] one to identify follicular groups (screening MLA) and the other to simultaneously predict the TBSRTC category and the final pathology (classifier MLA). To design a screening MLA that could identify follicular groups among all the cellular material on a direct smear, a cytopathologist (D.R.), using Aperio ImageScope software (Leica Biosystems, Inc), annotated ROIs containing follicular cells on a subset of WSIs from the training set. For the identification of uninformative regions or negative regions of interest (NROIs), we used a random selection of areas on the scans. Because the vast majority of regions on a direct smear do not contain any nucleated cellular material, a random selection of regions has a high probability of yielding an NROI. We used these negative examples and the annotated ROIs to train the screening MLA. Once it was trained, we applied the screening CNN to each WSI and extracted 1000 ROIs with the highest prediction values for being informative. These ROIs were saved offline and were, in turn, used for the training and evaluation of the classifier MLA.

The classifier MLA was trained on labeled WSIs from the training set to generate local predictions of the final pathology (benign or malignant) for each ROI. These local predictions were then aggregated into a global, single prediction for each WSI. This approach is referred to as multiple instance learning.[9,10] In addition to assigning binary labels to the final pathology, we simultaneously trained the algorithm to predict the TBSRTC category via an ordinal regression framework, in which the output of the neural network was compared with a trainable set of cutoffs to determine the correct category.[11]

For both CNNs, we used the same architecture based on VGG11 (implemented in PyTorch 0.4.1).[12] The convolutional filters were initialized with parameters pretrained on ImageNet, which is a large and widely used data set in computer vision.[13] To avoid overfitting, we used the following criterion for the screening CNN: training was stopped when performance on the validation set did not improve between epochs. This typically occurred after 1 to 5 epochs. The classifier MLA was trained for 100 epochs. The performance was evaluated on the validation set after each epoch. In turn, the network parameters were chosen according to the highest performance achieved during training. In our experiments, we used a computing server with 4 Tesla V-100 PCIE graphics processing units (GPUs) with 16 GB of memory each. Each model was trained with a single GPU. We previously published a detailed description of the engineering, principles, and performance of both the screening and classifier MLAs.[14]

TABLE 1. Electronic Medical Record Cytologic Diagnoses for the Training and Test Sets With Associated ROMs Based on the Final Pathology

	Training Set (n = 799)		Test Set (n = 109)	
TBSRTC (ROM[3])	Cases, No. (%)	ROM, %	Cases, No. (%)	ROM, %
BN (0%-3%)	380 (47.6)	2.2	50 (45.9)	0
AUS (10%-30%)	231 (28.9)	19.9	32 (29.4)	18.8
FN (25%-40%)	73 (9.1)	34.2	9 (8.3)	33.3
SUSP (50%-75%)	32 (4.0)	81.3	6 (5.5)	66.7
MAL (97%-99%)	83 (10.4)	100	12 (11)	100

Abbreviations: AUS, atypia of undetermined significance; BN, benign neoplasm; FN, follicular neoplasm; MAL, malignant; ROM, risk of malignancy; SUSP, suspicious for malignancy; TBSRTC, The Bethesda System for Reporting Thyroid Cytopathology.

In addition to evaluating the performance of the algorithm in predicting the final pathology, we also considered in our experiments a combined MLA-human approach in which the MLA was used to enhance human decisions. Specifically, we used a combination rule that used the EMR cytologic diagnoses for FNABs categorized as BN or MAL and the MLA classifier for the remaining cases in the indeterminate categories (AUS, FN, and SUSP).

RESULTS

A total of 916 cases were available for scanning after all exclusion criteria were applied. Eight WSIs were further excluded from the final cohort because of suboptimal scanning characteristics (typically poor focus or incomplete scanning with noticeable portions of the slide missing). This resulted in a final cohort of 908 WSIs from 659 patients. We split the cohort into a training set of 799 WSIs and a test set of 109 consecutive FNABs. The reviewing cytopathologist annotated 145 scans from the training set to identify ROIs containing follicular cells for a total of 4494 ROIs. Table 1 summarizes the TBSRTC diagnoses in the EMR for the training and test sets along with the calculated ROMs.

We used a 5-fold cross-validation procedure in which the training data (145 slides for the screening MLA and 799 slides for the classifier MLA) were split for training (80%) and validation procedures (20%). In this setting, we trained 5 models for each of the 2 CNNs, and the predictions of the 5 models were averaged. On the validation data, the screening MLA provided an area under the receiver operating characteristic (ROC) curve of 0.985 for the distinction between ROIs and NROIs. In addition, true, verified ROIs that were annotated by the cytopathologist in the training set yielded an average prediction of 0.97 by the MLA in the validation process. Figure 1 shows 2 smears with different amounts and distributions of cellularity (Fig. 1A,D) along with the corresponding heat maps (Fig. 1B,E) generated by the ROIs identified by the screening MLA. The heat maps were created by the application of the screening classifier to the WSI, which was divided into image regions with 75% overlap along each axis. The corresponding histograms represent the predicted probability of regions being informative (ROIs) on each smear (Fig. 1C,F). The majority of the regions received low prediction values, and this indicates that they are uninformative NROIs. The informative ROIs are represented by a significantly smaller group of regions that are clustered at the opposite end of the histograms with predictions between 0.9 and 1.

The final pathology of the test set was benign for 84 cases and malignant for 25 cases. Table 2 summarizes the predictions of malignancy made by the classifier MLA with respect to the final pathology (ground truth). Ninety percent of the benign nodules and 92% of the malignant nodules were correctly predicted to be benign and malignant by the MLA, respectively. The MLA was able to predict malignancy of the final surgical pathology with a sensitivity of 92.0% and a specificity of 90.5%. Figure 2 depicts ROC curves comparing final pathology predictions. The area under the curve (AUC) reflects the performance for the MLA (0.932), the EMR pathologists (0.931), and the reviewing cytopathologist (0.931) in predicting the final pathology in the test set. When the EMR diagnoses for cases categorized as BN and MAL were combined with the MLA predictions for the indeterminate EMR diagnoses of AUS, FN, and SUSP, the AUC (combined) increased to 0.962.

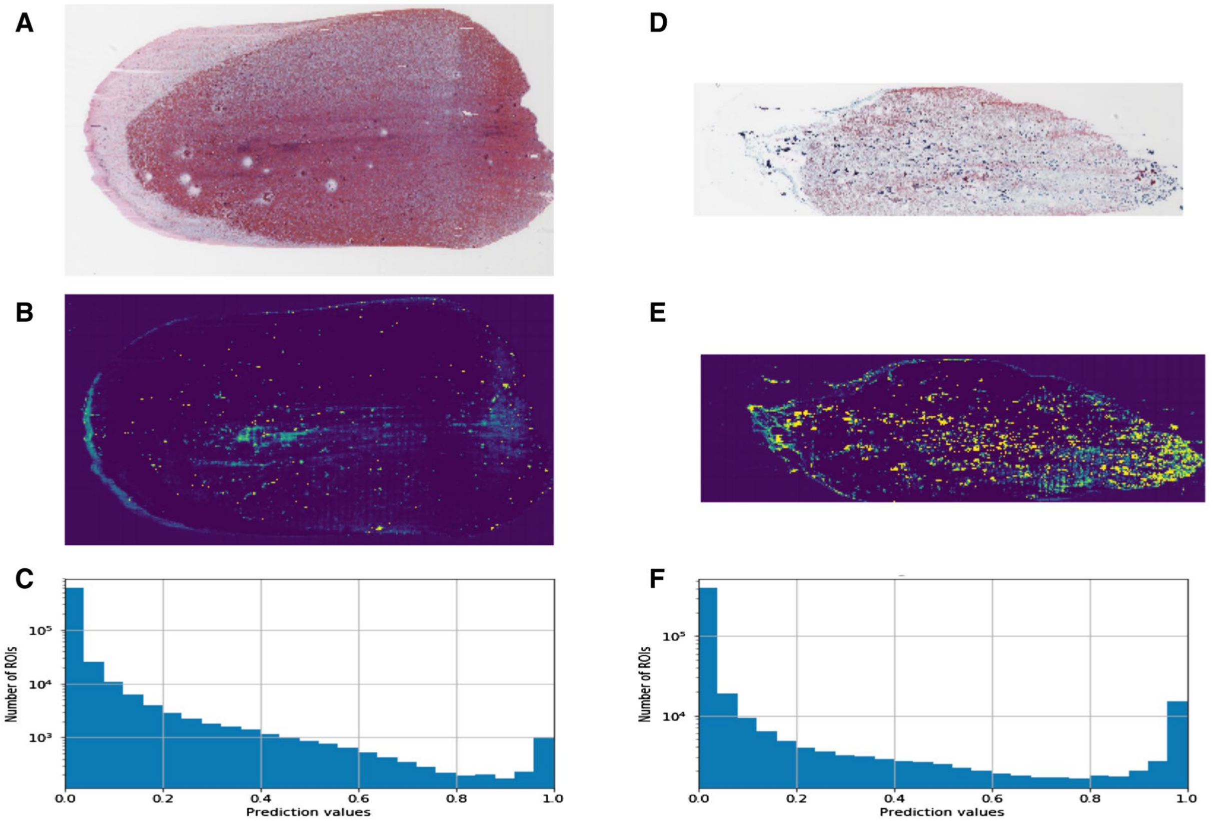


Figure 1. Direct smears of (A) moderately cellular and (B) very cellular fine-needle aspiration biopsies with (B,E) corresponding heat maps highlighting ROIs identified by the machine learning algorithm. (C,F) Histograms for each smear show large regions with prediction values less than 0.9, which correlate with uninformative regions. ROI indicates region of interest.

TABLE 2. MLA Predictions of Malignancy in Comparison With the Final Pathology

Ground Truth (Surgical Pathology)	Predicted by MLA, No.	
	Benign	Malignant
Benign (n = 84)	76	8
Malignant (n = 25)	2	23

Abbreviation: MLA, machine learning algorithm.

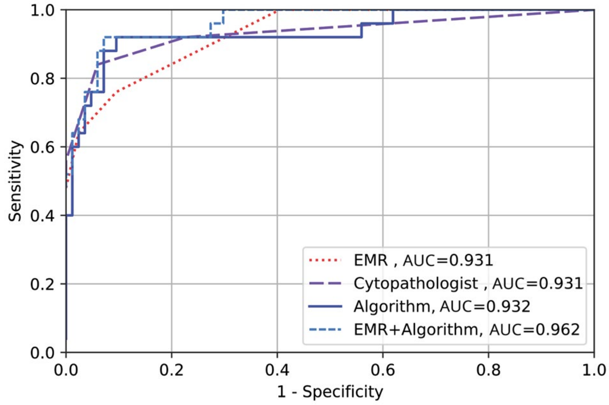


Figure 2. Receiver operating characteristic curves for the prediction of malignancy for the EMR, the cytopathologist (reviewer), the MLA alone, and the MLA using EMR diagnoses for benign and malignant TBSRTC categories (combined). AUC indicates area under the curve; EMR, electronic medical record; MLA, machine learning algorithm; TBSRTC, The Bethesda System for Reporting Thyroid Cytopathology.

DISCUSSION

We designed an MLA to apply to WSIs of thyroid FNAB to 1) identify ROIs containing follicular cells and 2) use those ROIs to predict the final pathology as either benign or malignant. The MLA performance was comparable to human levels with AUCs of 0.932 and 0.931, respectively. To our knowledge, no other studies have used an MLA to distinguish benign and malignant thyroid nodules on cytopathology WSIs.

The main challenge in training the MLA was obtaining a sufficient amount of labeled data for each CNN. For the screening CNN, the use of randomly selected NROIs for training allowed us to obtain large amounts of labeled image regions while saving a significant amount of manual effort by annotating only a few informative ROIs. To train

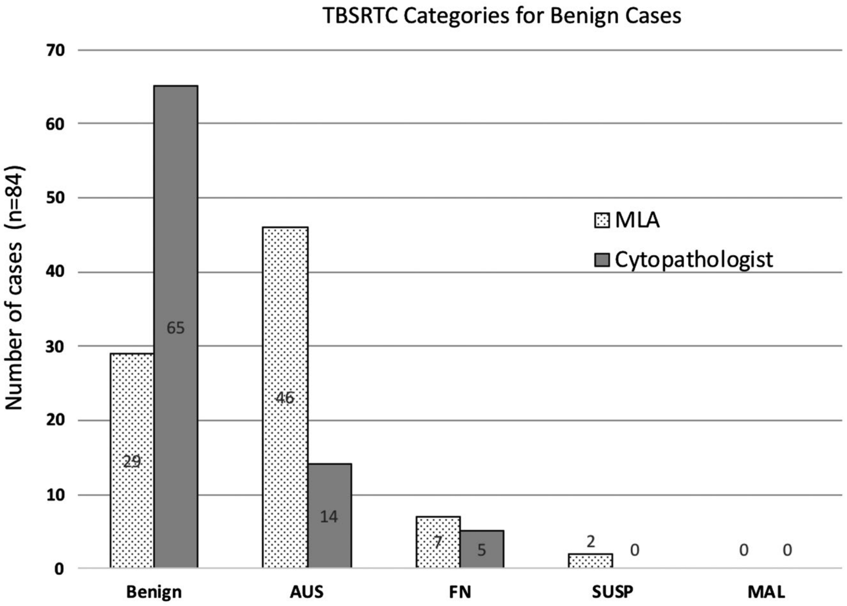


Figure 3. Distribution of benign cases across the TBSRTC categories. AUS indicates atypia of undetermined significance; FN, follicular neoplasm; MAL, malignant; MLA, machine learning algorithm; SUSP, suspicious for malignancy; TBSRTC, The Bethesda System for Reporting Thyroid Cytopathology.

the classifier MLA, we made the assumption that any 1 ROI on a given slide is representative of all the ROIs. In practical terms, this assumption holds true for FNAB cases in general, for which the standard assumption is that all or most follicular groups in a given FNAB are representative of the targeted lesion. This assumption allowed us to generate a training set of 799,000 labeled ROIs (1000 ROIs from each WSI in the training set). However, both strategies for identifying NROIs and ROIs may introduce a certain amount of inaccuracy into the labels, also known as label noise, where some informative ROIs may be randomly sampled and wrongly labeled as NROIs or NROIs may be wrongly included among the 1000 ROIs selected. This training strategy, based on the use of noisy labels, is closely related to a widely studied subfield in machine learning termed weakly/semisupervised learning.[15] Specifically, CNNs are known to be resistant to such inaccuracies in the labels, and as we found in our experiments, this approach indeed led to the accurate identification and classification of the ROIs.

The classifier MLA was designed to ultimately predict the final pathology as either benign or malignant. However, it was also trained to simultaneously predict a TBSRTC category. Cutoff values between 0 (benign) and 1 (malignant) were automatically established by the MLA and correlated with each of the 5 TBSRTC categories. The predicted TBSRTC category offers a window into how the MLA predicts malignancy by offering a familiar, ordinal, and easily understood surrogate. This training strategy, by which a CNN is trained to predict multiple labels, was recently found to improve classification performance in other medical image analysis tasks.[15,16] We found in our study that it did indeed improve the malignancy prediction. Figures 3 and 4, where we can see that the overwhelming majority of benign cases were categorized as BN or AUS by the MLA, illustrate this point. Moreover, the overwhelming majority of the malignant cases were categorized as FN, SUSP, or MAL by the MLA. All FNABs predicted as BN and MAL were benign and malignant, respectively, on final pathology. Likewise, among all the malignant nodules, the MLA did not place any into the BN cytologic category, and among all benign nodules, the MLA did not place any in the MAL category. This illustrates the tight correlation between the predicted TBSRTC category and the final pathology prediction, where the former predicts the latter. The TBSRTC cutoff values and predictions were not used to evaluate overall performance because the end goal was to predict the final surgical pathology result.

We found that the best way to evaluate the classifier MLA performance was to compare its ability to predict malignancy with the ability of humans. The human prediction of malignancy is reflected in the chosen TBSRTC

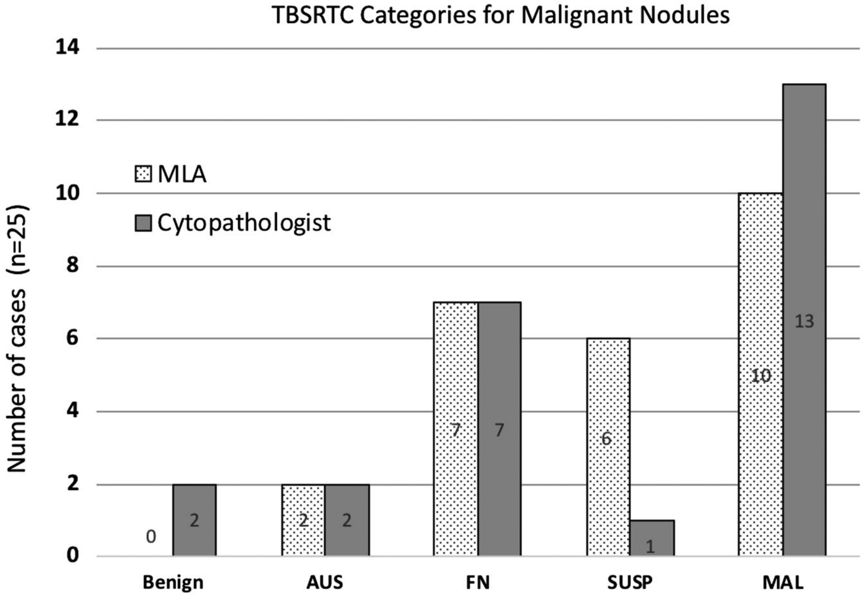


Figure 4. Distribution of malignant cases across the TBSRTC categories. AUS indicates atypia of undetermined significance; FN, follicular neoplasm; MAL, malignant; MLA, machine learning algorithm; SUSP, suspicious for malignancy; TBSRTC, The Bethesda System for Reporting Thyroid Cytopathology.

category, each of which has a defined prediction, or ROM. Thus, a pathologist's application of TBSRTC is an indirect reflection of the prediction of malignancy for a given case. Figure 2 compares the prediction of malignancy (ie, the TBSRTC category) of the reviewing cytopathologist and the EMR pathologist to that of the MLA. The AUCs in Figure 2 show that the MLA performance is comparable to that of the pathologists in the EMR, who represent as many as 11 general pathologists and board-eligible/board-certified cytopathologists with years of experience ranging from 1 to more than 20 years. The MLA performance was slightly worse than that of a single experienced cytopathologist.

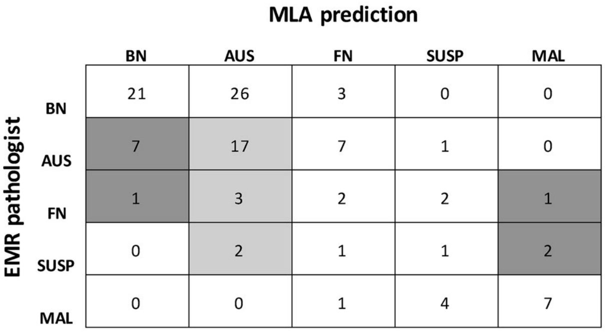
A total of 10 cases (9%) were incorrectly classified by the MLA. Two additional cytopathologists blindly reviewed these 10 cases to determine whether the discrepancies in diagnoses were due to slide selection or variation in interpretation. Neither of the 2 additional reviewers was the signing pathologist on the case. The 2 false-negative cases were the only malignant nodules predicted to be AUS by the MLA. They were diagnosed as AUS and FN in the EMR, and all 3 reviewers categorized these 2 cases as BN. The difference in TBSRTC classification was minimal (BN vs AUS), but the fact that all reviewers made the same BN cytologic diagnosis suggests that the single slide reviewed for the study may not have been wholly representative of the FNAB. Perhaps the other slides, not reviewed, showed some degree of atypia yielding an AUS or FN diagnosis by the EMR pathologist. Interestingly, both cases were follicular variant of papillary thyroid carcinoma, an entity well known to be difficult to classify on cytopathology.

Eight benign nodules were incorrectly predicted by the MLA to be malignant. Six were predicted by the MLA to be FN, and 2 were predicted to be SUSP; as expected, none of these were categorized as BN or AUS. Similarly to the MLA, at least 1 reviewer or 2 humans (a reviewer and an EMR pathologist) diagnosed 5 of the 8 cases as FN, SUSP, or MAL. Interestingly, 2 of the predicted FN cases were chronic lymphocytic thyroiditis (CLT), a known pitfall in this category when background lymphocytes are missed. Among the 3 remaining false-positive cases, 2 were categorized as BN by at least 2 reviewers, and 1 was diagnosed as AUS by all reviewers; this suggests that the MLA severely misclassified all 3 of these cases. This is an area that will require additional study and training because it highlights a limitation of the current study. The MLA was not trained to recognize other cell types such as lymphocytes and macrophages. It also was not trained to make specific diagnoses that may aid in clinical management, such as CLT and medullary thyroid carcinoma.

TABLE 3. Calculated ROMs Based on the Final Surgical Pathology for MLA Cytologic Predictions in the Test Set (n = 109)

TBSRTC Prediction	Cases, No.	ROM, %
BN	29	0
AUS	48	4.2
FN	14	50
SUSP	8	75
MAL	10	100

Abbreviations: AUS, atypia of undetermined significance; BN, benign neoplasm; FN, follicular neoplasm; MAL, malignant; MLA, machine learning algorithm; ROM, risk of malignancy; SUSP, suspicious for malignancy; TBSRTC, The Bethesda System for Reporting Thyroid Cytopathology.



EMR pathologist \ MLA prediction	BN	AUS	FN	SUSP	MAL
BN	21	26	3	0	0
AUS	7	17	7	1	0
FN	1	3	2	2	1
SUSP	0	2	1	1	2
MAL	0	0	1	4	7

Figure 5. Comparison of MLA predictions and EMR pathologists' cytologic diagnoses. Dark gray boxes indicate indeterminate EMR cases moved into definitive categories (BN or MAL) by the MLA combined rule. Light gray boxes indicate indeterminate EMR cases that would move into the adjusted BN category. AUS indicates atypia of undetermined significance; BN, benign neoplasm; EMR, electronic medical record; FN, follicular neoplasm; MAL, malignant; MLA, machine learning algorithm; SUSP, suspicious for malignancy.

Table 3 compares the ROMs generated by the MLA on the basis of the TBSRTC predictions. Even though the MLA predicted a larger number of cases to be AUS in comparison with the EMR, the predicted AUS category had a much lower ROM of 4.2% in comparison with that of the EMR (18.8%). Such a low ROM in the predicted AUS category is comparable to published malignancy rates of 0% to 3% in the BN category. Therefore, when the MLA is used, an AUS prediction could be considered predictive of a benign final pathology.[5] Taken another way, the predicted TBSRTC labels are simply automated cutoff values created by the MLA in the training process. These values could be further adjusted to yield more clinically relevant labels. For example, by adjusting the cutoff used for the BN label to encompass those values used to predict AUS, we could effectively collapse the 2 categories into 1 "adjusted" BN category in which 97.4% of cases would be benign on final pathology (ROM, 2.6%).

The ROMs for the predicted categories of SUSP and MAL were 75% and 100%, respectively. Both were on par with EMR and published ROM rates. This leaves us with the predicted FN category as the effective indeterminate category in the MLA predictions because this category had the highest amount of uncertainty with an ROM of 50%. However, the MLA categorized only 12.8% of the test cases as FN, in contrast to 29.4% of the cases diagnosed as AUS in the EMR.

Recognizing the superior human performance in predicting malignancy among BN and MAL FNAB cases, we created a rule using a combination of MLA and human decisions to leverage this performance. The rule instructed the algorithm to use the EMR cytologic diagnosis if it was BN or MAL and to provide predictions only for cases in the indeterminate categories (AUS, FN, and SUSP). Figure 2 shows that the MLA performance improved when the rule was applied, with the AUC increasing from 0.931 to 0.962 and with the specificity increasing from 90.5% to 92.9%. Looking specifically at the individual cases, Figure 5 shows the movement of cases when the MLA was combined with the human EMR decisions. All predicted BN cases that were diagnosed as indeterminate by the EMR pathologist (n = 8) moved into the BN category. Likewise, all predicted MAL cases that were diagnosed as indeterminate by the EMR pathologist (n = 3) moved into the MAL category. This movement reduced the indeterminate category of the EMR pathologist by 55% (from 47 to 26). Moreover, if we were to use an adjusted MLA cutoff for the BN category as noted previously, an additional 22 indeterminate EMR cases that were predicted to be AUS would also be moved into the BN category. These results show the potential for the use of the MLA as an adjunctive tool for cytopathologists to aid in reducing the indeterminate cases.

Unlike many artificial intelligence studies using thyroid cytopathology, our methods required relatively little human effort or expertise in the training process.[17,18] Only 145 WSIs were annotated for follicular cells with an average of 38 ROIs per scan. The annotations required minimal expertise and relatively few hours of physician time. The assignment of a TBSRTC category did require some expertise but only for a limited number of cases in the training set (n = 145). The training process did not necessitate any complex tasks beyond those that are part of an average cytopathologist's routine workflow.[19,20]

Manual acquisition of morphometric data and semiqualitative features of individual cells was not needed in the development of our MLA.[21-23] Finally, the use of WSIs precluded the need for manual acquisition and analysis of image regions while allowing for the use of large quantities of cells for analysis. This type of analysis using WSIs has been largely unaddressed in the machine learning literature and has positive implications for generalizability.

The current study highlights the potential for future clinical use. Our MLA could be used as a screening tool to identify follicular groups in practice settings that use digital pathology for the primary diagnosis and to save time for the cytopathologist and/or cytotechnologist. Additional studies to test and compare the identified ROIs against traditional slide review would be necessary. As shown by the performance of the combined ROC curve, this MLA could also be applied to individual FNABs with an indeterminate cytologic diagnosis, such that a prediction of adjusted BN would be a strong prediction of a benign final pathology. Alternatively, we could use the manually adjusted BN cutoff to yield a clinically relevant label, as discussed previously. The manually adjusted threshold would then have to undergo a rigorous validation process to optimize sensitivity and specificity for each category. An ideal classifier MLA would take into consideration not only the FNAB but also the sonographic and clinical data. The current MLA could be paired with other MLAs that use these parameters to create a more powerful tool for the prediction of thyroid malignancy.

They are some limitations to this study worth mentioning. Machine learning requires known data sets for training, and the ground truth for all of our cases used the cytologic gold standard of surgical pathology follow-up. This creates a selection bias that may favor more complex or malignant cases. However, the composition of our cohort was very similar to that of various studies examining ROM in both the frequency of cases and the ROMs across all 5 TBSRTC categories.[24] We are currently performing a prospective study to assess the performance of the MLA among all FNAB cases with and without surgery in an attempt to address this bias. The number of cases in the indeterminate category was relatively small, and a larger test set of such cases is needed to more thoroughly gauge the performance of the MLA in this category.

Each WSI averaged more than 30 GB of data, and this created challenges for storage and analysis. Therefore, we limited our slide selection to a single slide and did not use the remaining alcohol-fixed slides or the air-dried slides routinely obtained for thyroid FNABs at our institution. Although we chose the slide with the most follicular groups, the chosen slide may not have been representative of the cytologic diagnosis. This probably resulted in sampling errors for at least 2 of the false-positive cases for which the diagnoses of the MLA and the reviewers varied widely in comparison with those of the EMR pathologists, who had the benefit of reviewing all slides and the clinical history.

Our data set included 9 z-stacks per WSI, but we opted to use only the middle z-plane for our study. We did perform a limited experiment with a subset of training cases using all of the alcohol-fixed slides and more than 1 z-stack and saw no difference in performance. Using more than 1 z-stack would require the use of fewer WSIs in each training minibatch because of GPU memory capacity. We leave the study of this tradeoff in data usage and the construction of minibatches to a future study. Finally, the presence of colloid and lymphocytes are 2 features for which we did not train. The identification of colloid may help to reduce the higher number of predicted AUS cases (n = 48), many of which were diagnosed as BN by the EMR and the reviewing cytopathologist. Likewise, the recognition of lymphocytes in 2 of the false-positive CLT cases could have led to a more accurate categorization of BN and a final prediction of benign instead of FN and malignant.

To the best of our knowledge, our cohort, which is terabytes in size, is the largest cohort of thyroid FNAB WSIs reported in the literature. We used this cohort to train and test an MLA to predict the final pathology on the basis of FNAB. Some expertise was required in the training process to identify follicular cells and to categorize a subset of WSIs with TBSRTC. However, this type of training is part of a cytopathologist's routine workflow. Unlike other image analysis studies, this one allows for the use of this MLA in a variety of practice settings, including potentially low-technology environments.[25] Our MLA effectively identified ROIs and performed at human levels in the prediction of the final pathology.

FUNDING SUPPORT

Duke University faculty funds supported this study.

CONFLICT OF INTEREST DISCLOSURES

The authors made no disclosures.

AUTHOR CONTRIBUTIONS

Danielle D. Elliott Range: Conceptualization, data curation, investigation, methodology, resources, validation, visualization, writing–original draft, and writing–review and editing. **David Dov:** Conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing–original draft, and writing–review and editing. **Shahar Z. Kovalsky:** Conceptualization, data curation, formal analysis, funding acquisition, investigation, project administration, supervision, validation, and writing–review and editing. **Ricardo Henao:** Data curation, methodology, supervision, and writing–review and editing. **Lawrence Carin:** Conceptualization, formal analysis, funding acquisition, investigation, methodology, resources, supervision, and writing–review and editing. **Jonathan Cohen:** Conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, and writing–review and editing.

REFERENCES

1. Tamhane S, Gharib H. Thyroid nodule update on diagnosis and management. *Clin Diabetes Endocrinol*. 2016;2:17.
2. Popoveniuc G, Jonklaas J. Thyroid nodules. *Med Clin North Am*. 2012;96:329-349.
3. Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. *Thyroid*. 2017;27:1341-1346.
4. 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.
5. Faquin WC, Wong LQ, Afrogheh AH, et al. 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.
6. Eskander A, Hall SF, Manduch M, Griffiths R, Irish JC. A population-based study on NIFTP incidence and survival: is NIFTP really a "benign" disease? *Ann Surg Oncol*. 2019;26:1376-1384.
7. Xu B, Tallini G, Scognamiglio T, Roman BR, Tuttle RM, Ghossein RA. Outcome of large noninvasive follicular thyroid neoplasm with papillary-like nuclear features. *Thyroid*. 2017;27:512-517.
8. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60:84-90.
9. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. *arXiv*. 2018:1802.04712.
10. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*. 2016;32:i52-i59.
11. Dorado-Moreno M, Gutierrez PA, Hervas-Martinez C. Ordinal classification using hybrid artificial neural networks with projection and kernel basis functions. In: Rodriguez ESC, Snasel V, Abraham A, Wozniak M, Grana M, Cho SB, eds. Hybrid Artificial Intelligent Systems. Springer; 2012:319-330. Lecture Notes in Computer Science; vol 7209.
12. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*. 2015:1409.1556v6.
13. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. *IEEE Conf Comput Vision Pattern Recognit*. 2009:248-255.
14. Dov D, Kovalsky S, Cohen J, Range D, Henao R, Carin L. Thyroid cancer malignancy prediction from whole slide cytopathology images. *arXiv*. 2019:1904.00839.
15. 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.
16. Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. *arXiv*. 2017:1705.10694.
17. 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;127:501-513.
18. Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. *Diagn Cytopathol*. 2018;46:244-249.
19. 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.
20. 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.
21. Daskalakis A, Kostopoulos S, Spyridonos P, et al. Design of a multi-classifier system for discriminating benign from malignant thyroid nodules using routinely H&E-stained cytological images. *Comput Biol Med*. 2008;38:196-203.
22. Gopinath B, Shanthi N. Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images. *Australas Phys Eng Sci Med*. 2013;36:219-230.
23. Ozolek JA, Tosun AB, Wang W, et al. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Med Image Anal*. 2014;18:772-780.
24. 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.
25. Varlatzidou A, Pouliakis A, Stamataki M, et al. Cascaded learning vector quantizer neural networks for the discrimination of thyroid lesions. *Anal Quant Cytol Histol*. 2011;33:323-334.