MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Thyroid Cytopathology Cancer Diagnosis from Smartphone Images Using Machine Learning

Serge Assaad[a], David Dov[b,c], Richard Davis[c], Shahar Kovalsky[d], Walter T. Lee[c], Russel Kahmke[e], Daniel Rocke[e], Jonathan Cohen[e], Ricardo Henao[a,f], Lawrence Carin[a,f], Danielle Elliott Range[c,*]

[a] *Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina;* [b] *I-Medata AI Center, Tel Aviv Sourasky Medical Center, Tel Aviv-Yafo, Israel;* [c] *Department of Pathology, Duke University Medical Center, Durham, North Carolina;* [d] *Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina;* [e] *Department of Head and Neck Surgery and Communication Sciences, Duke University Medical Center, Durham, North Carolina;* [f] *King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*

ARTICLE INFO

Article history:
Received 7 October 2022
Revised 17 January 2023
Accepted 31 January 2023
Available online 13 February 2023

Keywords:
artificial intelligence
cytopathology
fine needle aspiration
machine learning
mobile imaging
thyroid

ABSTRACT

We examined the performance of deep learning models on the classification of thyroid fine-needle aspiration biopsies using microscope images captured in 2 ways: with a high-resolution scanner and with a mobile phone camera.

Our training set consisted of images from 964 whole-slide images captured with a high-resolution scanner. Our test set consisted of 100 slides; 20 manually selected regions of interest (ROIs) from each slide were captured in 2 ways as mentioned above.

Applying a baseline machine learning algorithm trained on scanner ROIs resulted in performance deterioration when applied to the smartphone ROIs (97.8% area under the receiver operating characteristic curve [AUC], CI = [95.4%, 100.0%] for scanner images vs 89.5% AUC, CI = [82.3%, 96.6%] for mobile images, $P = .019$). Preliminary analysis via histogram matching showed that the baseline model was overly sensitive to slight color variations in the images (specifically, to color differences between mobile and scanner images). Adding color augmentation during training reduces this sensitivity and narrows the performance gap between mobile and scanner images (97.6% AUC, CI = [95.0%, 100.0%] for scanner images vs 96.0% AUC, CI = [91.8%, 100.0%] for mobile images, $P = .309$), with both modalities on par with human pathologist performance (95.6% AUC, CI = [91.6%, 99.5%]) for malignancy prediction ($P = .398$ for pathologist vs scanner and $P = .875$ for pathologist vs mobile). For indeterminate cases (pathologist-assigned Bethesda category of 3, 4, or 5), color augmentations confer some improvement (88.3% AUC, CI = [73.7%, 100.0%] for the baseline model vs 96.2% AUC, CI = [90.9%, 100.0%] with color augmentations, $P = .158$). In addition, we found that our model's performance levels off after 15 ROIs, a promising indication that ROI data collection would not be time-consuming for our diagnostic system. Finally, we showed that the model has sensible Bethesda category (TBS) predictions (increasing risk malignancy rate with predicted TBS category, with 0% malignancy for predicted TBS 2 and 100% malignancy for TBS 6).

Introduction

There is a severe pathologist shortage in low- and middle-income countries (LMICs): the number of anatomical pathologists (per capita) is 50 times smaller in LMICs than in high-income

[*] Corresponding author.
E-mail address: danielle.range@duke.edu (D.E. Range).

0893-3952/$ - see front matter © 2023 United States & Canadian Academy of Pathology. Published by Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.modpat.2023.100129

countries.[1] Some authors have proposed using machine learning algorithms instead of human pathologists to diagnose diseases from the digital whole-slide images (WSIs).[2-4] However, digitizing slides requires expensive scanners costing up to $250,000,[5] a great financial burden for clinical centers in LMICs. Mobile phones have the potential to democratize slide digitization because they are ubiquitous and up to 1000 times less expensive than state-of-the-art slide scanners. Together with machine learning algorithms for cancer diagnosis, WSI capture with a mobile phone could be a vital innovation for LMICs with neither a high-resolution scanner nor an expert pathologist.

In this article, we propose a semi-automatic system to predict thyroid cancer from fine-needle aspiration biopsy (FNAB) slides using mobile phone images. In the workflow, we propose that regions of interest (ROIs) on the slide be manually selected and photographed (eg, by a cytotechnologist) using a smartphone attached to a microscope with an adapter. A classification neural network would then predict the final surgical pathology of each slide (benign or malignant) and the cytopathology diagnosis based on the Bethesda System for the Reporting of Thyroid Cytopathology (TBS).[6] The Bethesda System has 6 diagnostic categories (1 = nondiagnostic, 2 = benign, 3 = atypical, 4 = neoplastic, 5 = suspicious for malignancy, and 6 = malignant), each with an associated risk of malignancy (ROM, 5%—10%, 0%—3%, 6%—18%, 10%—40%, 45%—60%, and 94%—96%, respectively).

Establishing such a workflow raises several questions. First, is the quality of mobile images sufficient for accurate thyroid cancer classification? Second, what is the impact of training our neural network classifier using WSIs but deploying it on mobile phone images? Finally, because the smartphone camera field-of-view is small, is it practical and feasible to capture enough diagnostic material from each slide? How many ROIs are required to make reliable diagnoses?

We addressed the above challenges to demonstrate the successful application of machine learning for diagnostic cytopathology using smartphone images.

Materials and Methods

Data set

The training set consisted of 964 WSIs of alcohol-fixed Papanicolaou-stained thyroid FNAB slides, each with 1000 ROIs (each ROI is a 128 × 128 pixel RGB image). The WSIs consisted of all FNABs (with final pathology from surgical follow-up) from our institution's medical center from 2008 to 2016. The slides were scanned using a Leica AT-2 scanner (Leica Biosystems) at 40× magnification, then the resolution was down-sampled by a factor of 4. As in our previous work,[3] the ROIs for the training set were selected using an ROI detection network (based on the VGG-11 neural network architecture[7]) trained to detect follicular groups.

The test set consisted of 100 FNAB slides, each with 20 ROIs captured in 2 modalities: high-resolution scanner and mobile phone. Example test set ROIs are shown in Figure 1 (right). Every slide is assigned 2 labels: the (binary) postsurgery malignancy diagnosis ("final pathology") and the pathologist-assigned ordinal risk assessment (between 2 and 6) according to The Bethesda System (TBS),[3] extracted from the Electronic Medical Record (EMR). We limited the number of ROIs to 20 to expedite data collection, and we later showed that our model performance saturates at 15 ROIs (regardless of the ROI capture order).

The ROI pairs per test slide were selected as follows: for each slide, we used the ROI detection network to select 1000 ROIs containing diagnostic regions of follicular cells from the WSI (the detailed procedure is described in our previous work[3]). The ROI bounding boxes were then overlaid on the WSI using the Aperio ImageScope software (Leica Biosystems). A pathology resident (R.D.) reviewed the ROIs using the software and selected 20 ROIs. Specifically, R.D. was instructed to select regions that are in focus and contain groups of follicular cells. R.D. selected the ROIs only based on the WSI, without access to any other patient information. Moreover, R.D. selected ROIs only based on image quality and clarity of follicular groups, with no regard to the cancer characteristics of the follicular groups. R.D. then located the 20 ROIs in the tissue sample using a conventional microscope and captured them with a Redmi Note 10S mobile phone camera (Xiaomi Inc) attached to an Olympus BX46 microscope (Olympus Corporation) via a GoSky microscope lens adapter (GoSky Optics Inc). A picture of the mobile ROI capture setup is shown in Figure 2. This process created a paired data set of the scanner and mobile images to enable a one-to-one comparison. A limitation of this study is that, in favor of comparability, the process to capture the data set of ROI pairs somewhat differs from the way our system would be used in practice, namely, a cytotechnologist would select 20 ROIs directly using the microscope, without having access to the WSI. Using the

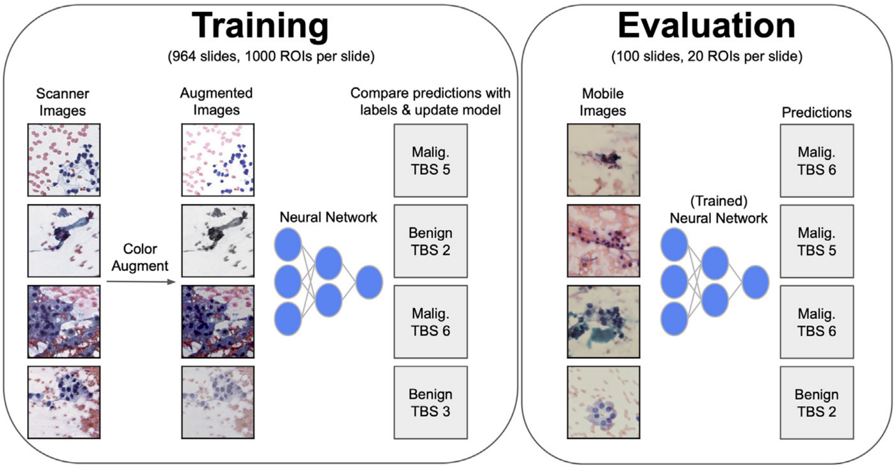

Figure 1.
Proposed procedures for training (left) and evaluation (right). During training, we applied color augmentations to the scanner images, made malignancy and sensible Bethesda System (TBS) category predictions using the neural network classifier, compared the predictions with the malignancy and TBS labels, and finally updated the model.

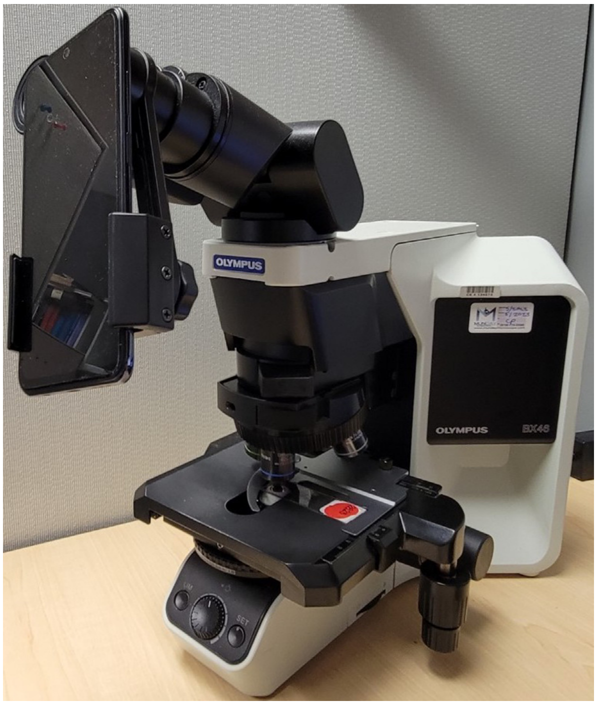

Figure 2.
Mobile region of interest (ROI) capture setup. For each slide, we captured 20 ROIs using a Redmi Note 10S camera attached to an Olympus BX46 via a GoSky microscope lens adapter.

detection network to first find 1000 ROIs and then displaying their bounding boxes on the ImageScope software was done purely to facilitate the collection of the paired ROI data set.

Classification Model Architecture

To classify malignancy, we used a MobileNetV2 model,[8] followed by global average pooling[9] and a linear layer. We based our classification model on MobileNetV2 (instead of VGG-11 as in our previous work[3]) because it is fast, lightweight, and explicitly designed to operate on mobile phones.

We trained the model to predict both the binary final pathology (using a cross-entropy loss) and the pathologist-assigned TBS (using an ordinal regression loss). In our previous work,[3] we found that this "2-endpoint" approach performed better than training the network using only the final pathology. We note that the slide-level malignancy and TBS labels were assigned to each ROI, an assumption that works well in practice if the ROIs contain follicular groups.[3]

Model Training

We trained the model for 1000 epochs (ie, 1000 passes over the training data set) using the AdamW optimizer[10] with a learning rate of 0.001 and weight decay parameter of 0.01. We used a batch size of 24 slides, with 12 ROIs per slide (randomly sampled at each epoch from the 1000 ROIs available per slide).

Our neural network classifier was trained entirely on high-resolution scanner images, because collecting a large enough data set and performing joint training or transfer learning on mobile images would be prohibitively slow.

Model Evaluation

During the evaluation, we first made ROI-level malignancy predictions using the classification model (for each of the 20 ROIs available per test slide). Then, we averaged the ROI-level predictions to obtain slide-level predictions. Finally, the slide-level predictions were compared with binary final pathology to compute the area under the receiver operating characteristic curve (AUC).

To improve performance, we obtained 5 different models using 5-fold cross-validation, and we used early stopping on the validation set to select the best model for each fold. During evaluation, we averaged the predictions of the 5 models to get an "ensemble" prediction.[11]

As additional evaluation steps, we examined our model's performance on "indeterminate cases" (slides with an EMR diagnosis of TBS 3, 4, or 5), as well as the observed ROM within the Bethesda categories predicted by the model.

For all reported AUC values, we also estimated 95% CIs based on DeLong's method.[12] Furthermore, for all AUC comparisons, we reported 2-sided P values using DeLong's test and use a significance level of $\alpha = 0.05$.

We examined our model's performance compared with the number of ROIs used. We averaged the model's performance across 1000 random permutations of the 20 ROIs to account for randomness in the ROI ordering. To assess convergence, we used a 2-sided Wilcoxon signed-rank test comparing the model's performance using 1–19 ROIs with the model's performance using 20 ROIs. We determined the convergence based on $P < .05/19$ (ie, a significance level of α=0.05 with a Bonferroni correction for $N = 19$ comparisons).

All our models were implemented in Python 3.9.7 with PyTorch 1.11.0. Each model was trained on a single Tesla V100 PCIE graphics processing unit with 16 GB of memory.

Data Augmentation

Data augmentation is a standard way to improve the training of computer vision algorithms. Data augmentation is a set of image manipulations (eg, changing the contrast/brightness, rotating), which do not fundamentally change the content of the image (the manipulations are said to be "label-preserving"[13]). In our problem, the characteristics of malignancy should be preserved after augmentation – slightly changing the brightness/contrast of an image does not change our decision about whether it should be classified as malignant (see Figure 3 for examples of data augmentations). The result of this augmentation process can be thought of in 2 ways. First, it effectively increases the size of the training set: each image/label pair appears multiple times in the training process, each time with a different image augmentation. Having a larger training set means that the classification model is less likely to overfit the training data. Second, augmentations make the algorithm more robust to small variations in image characteristics, which in turn improves the performance of the algorithm on mobile images, as we show in this article.

We based our data augmentation strategy on "TrivialAugment",[14] which recently achieved state-of-the-art performance on several computer vision benchmarks. TrivialAugment works as follows: for each example (ie, each ROI), a data augmentation operation is randomly sampled from the following list: brightness, saturation, contrast, sharpness, posterize, equalize, solarize, autocontrast, shear, rotation, and translation. Then, an augmentation

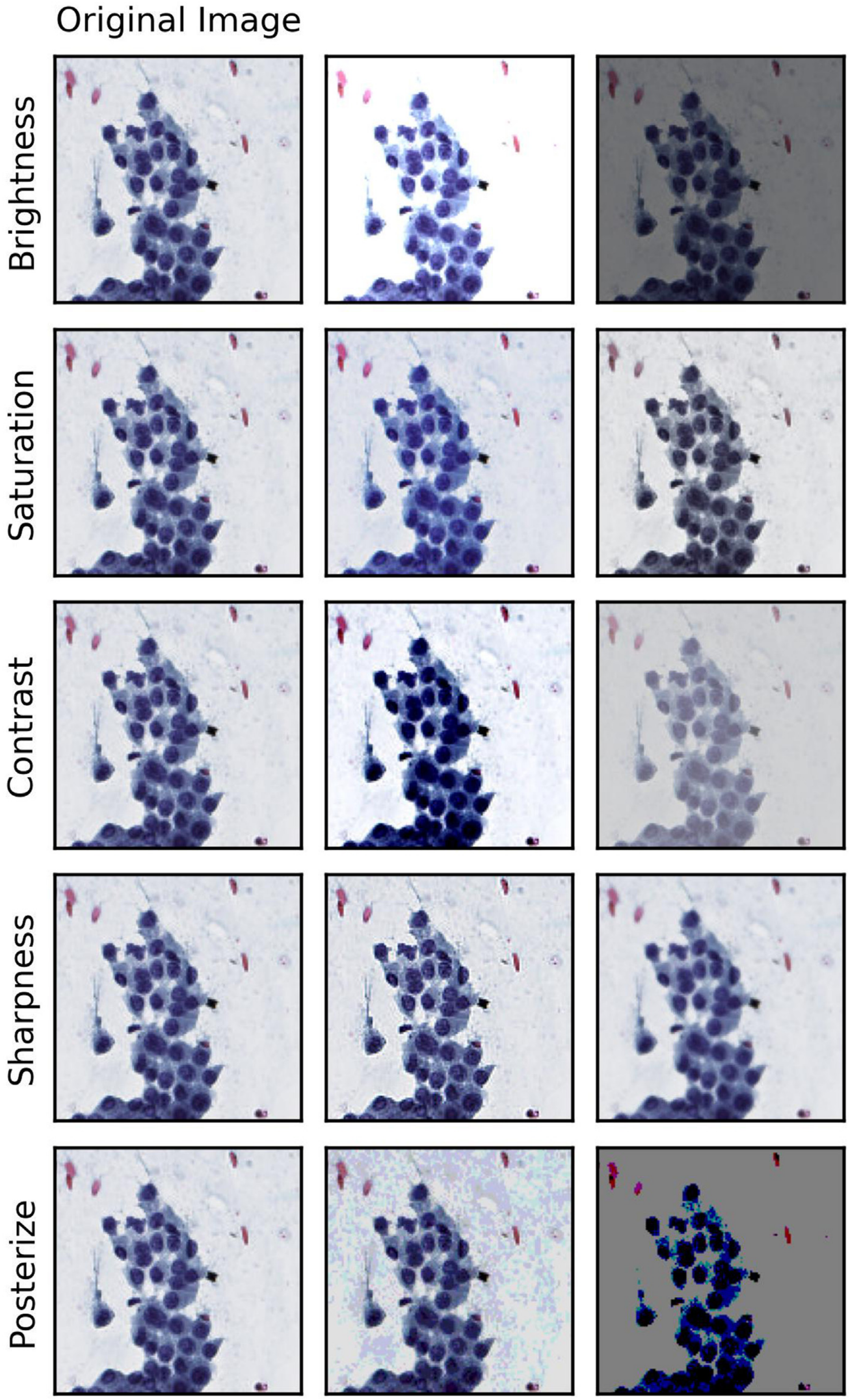

Figure 3.
Data augmentation examples, shown on the same regions of interest (ROIs) for ease of comparison. Each row is a different type of augmentation, and each column is an augmentation "strength." Left column shows the original ROI, and the middle and right columns show the augmented ROI (for 2 different strengths).

strength (eg, brightness level, contrast level, rotation angle) is randomly sampled, and finally, the augmentation is applied to the ROI. We made 2 changes to the TrivialAugment algorithm. First, we removed shears from the list of augmentations, because image shears are not characteristic-preserving for cytology images. Second, to facilitate analysis, we split the list of augmentations into 2 distinct types: color augmentations (brightness, saturation, contrast, sharpness, posterize, equalize, solarize, autocontrast — see Figure 3 for examples) and motion augmentations (rotations and translations). This split is useful because we wanted to specifically study the effect of color augmentations on model performance because we hypothesized that the main difference between the

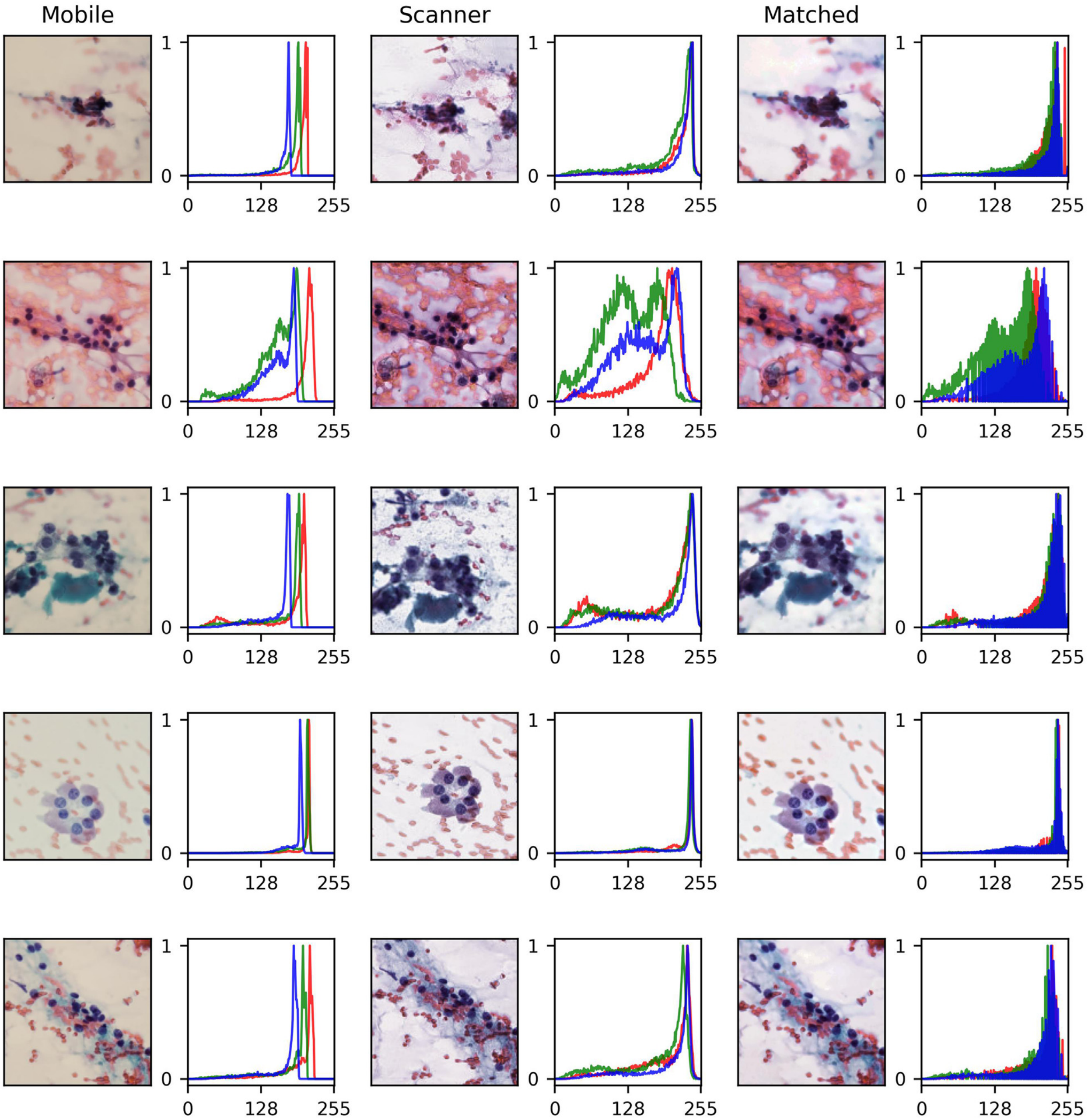

Figure 4.
Examples of regions of interest (ROIs) from the mobile test set (left) and the scanner test set (center), with their corresponding RGB histograms. “Matched” (right) is identical to the mobile image, but approximately matches the color histogram of the scanner image. Histograms are scaled to be between 0 and 1 for ease of visualization.

training and test sets is in their color compositions (see Figure 4). Note that the data augmentation procedure is applied only to the training set of scanner ROIs (Figure 1, left). During the evaluation of the paired mobile/scanner ROIs, we used the images as is, without any augmentation (Figure 1, right).

Results

We trained a baseline model without augmentation, similar to a model from our previous work, which achieves human-level performance for the prediction of malignancy using WSIs of thyroid FNABs.[3] Although this baseline model provides high AUC (97.8%, CI = [95.4%, 100.0%]) for scanner ROIs, its performance is statistically worse for mobile images (89.5% AUC, CI = [82.3%, 96.6%], $P = .019$). Paired mobile and scanner ROIs show the same follicular group(s), but they have different color characteristics. Therefore, we hypothesized that the degradation in performance is likely due to color differences. To show this, we modified the mobile ROIs by matching their color histograms with their corresponding ROIs from the scanner set (Figure 4). Histogram matching improved performance on the mobile test set with 94.8% AUC, CI = [90.7%, 98.8%] vs 89.5% AUC, CI = [82.3%, 96.6%] for the baseline model (P = .115) (Figure 5). For reference, Figure 5 also shows the performance of the EMR pathologist (95.6% AUC, CI = [91.6%, 99.5%]) by comparing their assigned TBS

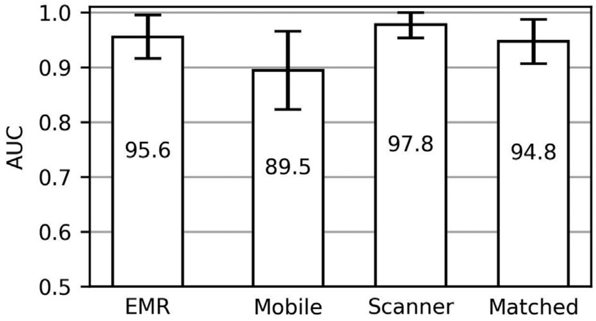

Figure 5.
Naïve model performance on mobile, scanner, and "matched" images. The naïve model is significantly worse on mobile images (89.5% AUC, CI = [82.3%, 96.6%]) compared with scanner images (97.8% AUC, CI = [95.4%, 100.0%], $P = .019$). Applying histogram matching ("Matched") improves performance (94.8% AUC, CI= [90.7%, 98.8%], $P = .115$) on mobile images. For reference, we also show the performance of the EMR pathologist's assigned TBS category (95.6% AUC, CI = [91.6%, 99.5%]). The error bars show the 95% confidence interval obtained from DeLong's method. AUC, area under the receiver operating characteristic curve.

category to ground-truth final pathology. Of course, histogram matching is not viable in practice because we will not have a corresponding scanner image to match histograms. However, the performance improvement using histogram matching alone suggests that the baseline model is overly sensitive to slight color variations and performs better on images whose color distributions are similar to the training data. This informed our next iteration of the model, in which we used data augmentation during training.

Figure 6 summarizes the model performance for different types of data augmentation. We found that data augmentation during training is effective in closing the performance gap between the mobile and scanner test sets. Adding color augmentations improved the performance on the mobile test set (96.0% AUC, CI = [91.8%, 100.0%] vs 89.5% AUC, CI = [82.3%, 96.6%] for the baseline model, $P = .048$) and did not significantly decrease performance on the scanner test set (97.6% AUC, CI = [95.0%, 100.0%] vs 97.8% AUC, CI = [95.4%, 100.0%] for the baseline model, $P = .546$). Furthermore, we found that motion augmentations improved performance on the mobile test set (92.7% AUC, CI = [87.7%, 97.7%] vs 89.5% AUC, CI = [82.3%, 96.6%] for the baseline model, $P = .205$) but not to the same level as color augmentations. Finally, we note that our proposed model (ie, with color augmentations) had an AUC that is statistically indistinguishable from the EMR pathologist's AUC (96.0%, CI = [91.8%, 100.0%] for the model vs 95.6% AUC, CI = [91.6%, 99.5%] for the EMR pathologist, $P = .875$).

Figure 7 shows how the performance of the model evolves as we use an increasing number of ROIs per test slide (starting at 1 ROI, and going up to 20 ROIs). To control for randomness in the ROI ordering, we reported the average performance and interquartile range for 1000 random permutations of the ROI ordering. For our proposed model (trained with color augmentations), we found that the AUC levels off at 15 ROIs for the mobile test set: using the Wilcoxon 2-sided signed-rank test compared with model performance at 20 ROIs, we obtained $P < 0.05/19$ for 1–14 ROIs and $P > 0.05/19$ for 15–19 ROIs. In Supplementary Figure S1, we plotted all P values obtained from this test for 1–19 ROIs.

Figure 8 shows the performance of our model in "indeterminate cases" (cases classified as TBS 3, 4, or 5 by the EMR pathologist). For the mobile test set, our proposed model (trained with color augmentations) achieved 96.2% AUC (vs 88.3% AUC for the baseline model, $P = .158$). We also note that the performance of our model on this "indeterminate" subset was better for mobile images than for scanner images but not significantly so (94.5% AUC, CI = [87.1%, 100.0%] for scanner images vs 96.2% AUC, CI = [90.9%, 100.0%] for mobile images, $P = .564$).

Finally, we measured the ROM in each of the model's predicted TBS category. The ROMs for TBS 2, 3, 4, 5, and 6 were 0.0%, 16.7%, 90.0%, 100.0%, and 100.0%, respectively (vs 0%–3, 6%–18, 10%–40, 45%–60, and 94%–96% reported in Cibas and Ali[6]).

Discussion

In this study, we evaluated whether a machine learning algorithm trained on WSIs of thyroid FNABs could be effectively applied to smartphone images. Although a few recent works[3,15-17] have also used similar approaches for thyroid cytopathology WSIs (see Kezlarian and Lin[18] for a comprehensive review), this study is (to our knowledge) the first to use machine learning to classify malignancy from mobile images of thyroid FNABs. Our approach is particularly relevant to the practice of cytopathology in LMICs. On a data set of 100 slides (with 20 ROIs per slide) selected from the WSI and captured with a smartphone, we examined the performance gap between mobile phone-based and scanner-based imaging for the diagnosis of cancer. Our study shows that mobile phone image capture, coupled with machine learning, can be used to classify thyroid FNABs and aid in the prediction of malignancy. Our approach raises several important questions that we addressed in this work.

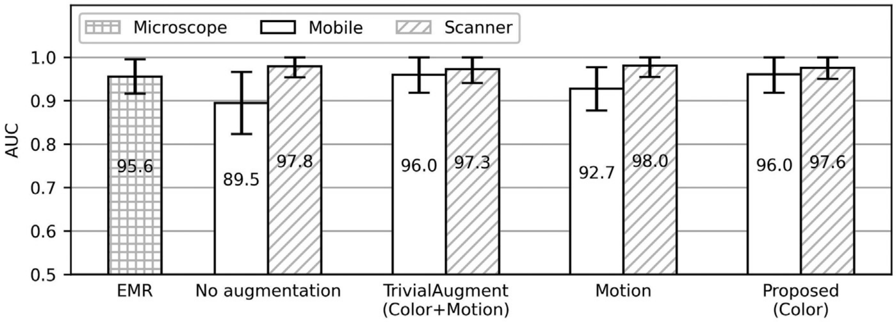

Figure 6.
Effect of different training data augmentations on model performance. The proposed color augmentations (right) yield the most improvement in mobile test set AUC (from 89.5% AUC, CI = [82.3%, 96.6%] to 96.0% AUC, CI = [91.8%, 100.0%], $P = .048$). The hatching patterns show the modality used for diagnosis. The error bars show the 95% CI obtained from DeLong's method. AUC, area under the receiver operating characteristic curve.

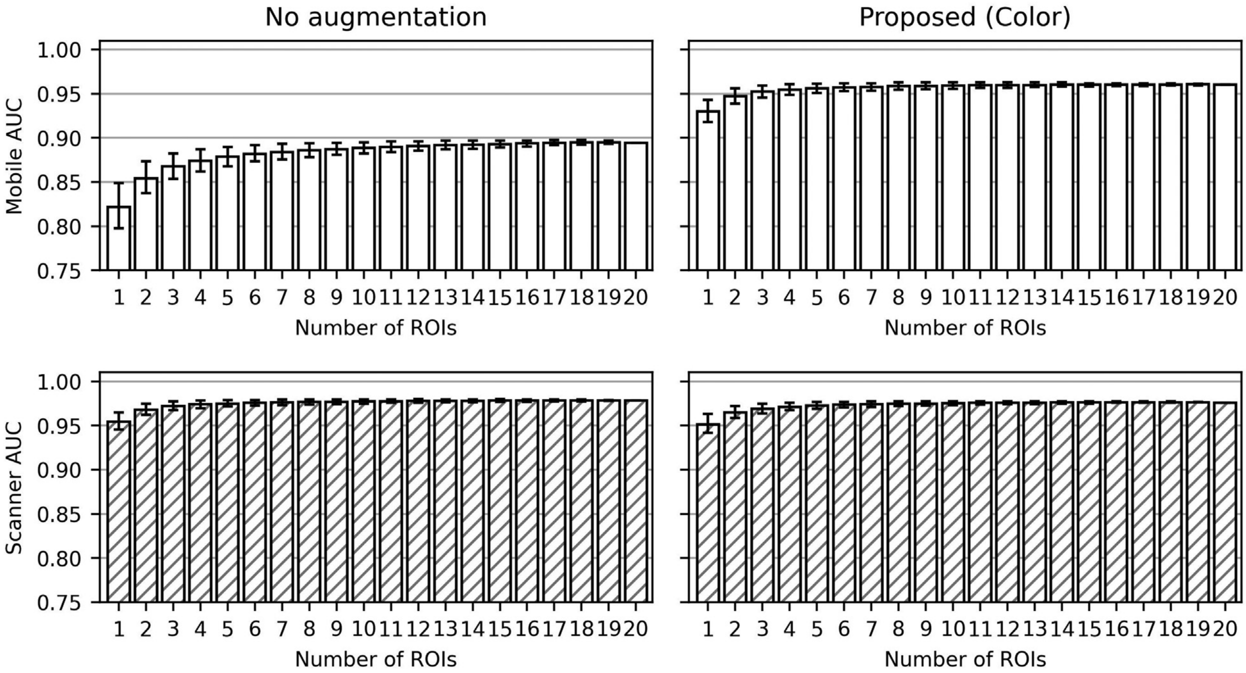

Figure 7.
Mobile and Scanner AUC vs number of ROIs at test time for the naïve model (no augmentation) and the proposed model (color augmentations). The error bars are the interquartile range over 1000 random permutations of ROI orderings. The proposed model converges at 15 ROIs (determined using a 2-sided Wilcoxon signed-rank test comparing with performance at 20 ROIs with significance level $\alpha = 0.05$ and a Bonferroni correction for $N = 19$ comparisons). AUC, area under the receiver operating characteristic curve; ROI, region of interest.

First, what is the difference in the diagnostic quality of mobile phone images, as compared with WSIs obtained from a high-resolution scanner? Skandarajah et al[19] analyzed the optical quality (eg, resolution, image distortion, illumination variation) of mobile phone cameras for cell imaging. They showed that 5-megapixel mobile phone cameras have enough resolution to operate near the diffraction limit in magnification ranges relevant for cell imaging. They noted that the typical convenient automation of mobile phone cameras (eg, autofocus, exposure) can hinder accurate color capture. Rivenson et al[20] used deep learning to correct differences in image characteristics between mobile images and scanner images for lung tissue, pap smears, and blood smears reliably (as measured by reconstruction error). However, we believe that diagnostic quality is best assessed by measuring the accuracy of the deep learning algorithm to predict malignancy, which directly quantifies the gap in diagnostic quality between scanner and mobile images. Specifically, we compared our model's malignancy against final surgical pathology. Our initial assessment showed an 8.3% absolute AUC decrease in performance ($P = .019$) between the scanner and mobile images using the baseline algorithm/neural network. By noting that color histogram matching improved performance (Figure 5), we showed that our baseline algorithm is overly sensitive to slight variations in color. However, matching color histograms between the scanner and mobile images is not feasible in low-resource settings where scanner images would not be available.

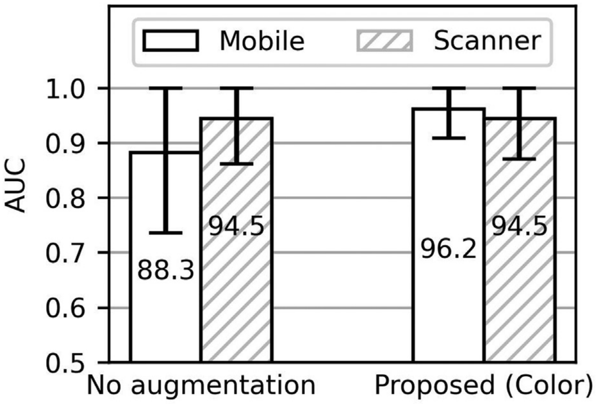

Figure 8.
AUC of the naïve model (no augmentation) and the proposed model (color augmentations) on indeterminate cases (cases assigned TBS 3, 4, and 5 by the EMR pathologist). The hatching patterns show the modality used for diagnosis. The error bars show the 95% confidence interval obtained from DeLong's method. AUC, area under the receiver operating characteristic curve.

This brings us to the second challenge in implementing our workflow: how can we practically address color and image quality differences between mobile and scanner images? De Haan et al[21] built an end-to-end system for the automatic diagnosis of sickle cell disease from mobile phone-based microscope images. They used an "enhancement network" to correct differences between the mobile phone images and those of a benchtop microscope. This network explicitly models the differences between a specific scanner and a specific mobile phone. Such a process, in our case, would require a large set of paired mobile/scanner images to train the "enhancement network." Our approach simplifies the process: rather than changing the smartphone images with an additional enhancement network, we used data augmentation (only during training) to make our classification model less sensitive to differences in image characteristics. This is also an important consideration from a clinical perspective where direct manipulation of digital pathology images for diagnostic purposes may lead to regulatory issues. In addition, this data augmentation approach does not require the time-consuming collection of paired ROIs for enhancement network training. The approach results in a robust, more platform-agnostic model (ie, a model which is not tied to specific phone or scanner models), as evidenced by the model's performance on a test set of mobile images (having been trained *only* using WSIs).

Applying color augmentations during training improved the algorithm's performance on the mobile test set to be statistically indistinguishable from its performance on the scanner test set across all TBS diagnoses (97.6% AUC, CI = [95.0%, 100.0%] on scanner test set vs 96.0%, CI = [91.8%, 100.0%] on mobile test set, $P = .309$) and among the indeterminate cases (96.1% AUC, CI = [90.5%, 100.0%] for mobile vs 94.3%, CI = [86.7%, 100.0%] for scanner, $P = .718$). Additionally, color augmentations improved performance on indeterminate (TBS 3, 4, and 5) mobile images (88.3% AUC, CI= [73.7%, 100.0%] for the baseline model vs 96.2% AUC, CI= [90.9%, 100.0%] with color augmentations, $P = .158$). This is a promising indication that mobile phones could serve as a viable slide digitization mechanism for diagnostic use.

Another practical consideration for this workflow is the amount of time it takes to collect the ROIs from the slide. In the proposed workflow, a trained laboratory technologist would review FNAB slides under the microscope and use a smartphone mounted to the microscope to capture 20 ROIs. Our machine learning algorithm would then be applied to these ROIs and provide a prediction of malignancy and TBS diagnosis. In our previous work,[3] we used hundreds of ROIs from the WSI to make a malignancy prediction for each slide. This is not practical in a real-world setting because capturing hundreds of ROIs with a mobile phone would be time-consuming. We analyzed the performance of our proposed model as a function of the number of ROIs used at the test time to determine how many ROIs needed to be manually collected. We showed that our model's performance across all TBS categories converges at 15 ROIs for the mobile test set (Figure 7). In terms of time spent, capturing 20 ROIs with the mobile phone and microscope setup was comparable to manually highlighting 20 ROIs on the Aperio ImageScope platform (between 5 and 10 minutes per slide). By attaching the camera phone to the microscope, the user (R.D.) integrated ROI capture with standard microscopic techniques (ie, coarse focus, fine focus, switching objectives, and use of a slide stage) familiar from routine pathology practice. In this respect, the camera phone system is more like routine practice than the use of WSIs.

Although the primary goal of the model is to predict the final pathology, the model's predicted TBS category also provides valuable insight into its mechanism. First, the observed ROM increased with the predicted TBS category. Moreover, the ROMs (for all categories except TBS 4) were close to observed ROMs in clinical practice.[22] We note that the ROMs in our model's predicted TBS categories were skewed higher for TBS 4 and 5 than our institutional ROMs.[23] The TBS 4 ROM of the model was 90.0% compared with the institutional ROM of 35%. This skew is likely because our model was trained to predict the final pathology, so it attempts to separate benign and malignant cases distinctly. In doing so, the model assigns most malignant cases into TBS categories 4 or higher. The high ROM for the model's predicted TBS 4 category is clinically acceptable because surgery is indicated in most cases assigned TBS 4 and above.[24] A similar effect is noted for TBS 5, although the difference with institutional ROM is not as stark (100% for the model vs 79.5% institutional ROM).

The overall performance of our method (96.0% AUC on mobile images), its performance in indeterminate cases (96.2% AUC on mobile images), its data efficiency (AUC convergence at 15 ROIs per slide), and its sensible TBS predictions support our hypothesis that the proposed workflow can be effective in thyroid cancer diagnosis. To our knowledge, this study is the first application of machine learning to thyroid cytopathology using a smartphone. We believe that this proof-of-concept is promising for low-resource settings, where neither high-resolution scanners nor expert pathologists are always available. It is a first step toward deploying an automated system needing only a microscope and a smartphone equipped with a machine learning algorithm. In particular, the malignancy prediction algorithm we used, MobileNetV2, was specifically designed to operate on mobile phones given its light memory footprint and fast runtime.

Several limitations of our work and directions for future work should be addressed. First, there is a difference between the way the model was evaluated and its intended deployment. During evaluation, the pathology resident selected 20 ROIs from a pool of 1000 scanner ROIs identified by a neural network. Then, they found the corresponding ROIs on a microscope and captured them using a mobile phone – this was done only to speed up the creation of the paired test data set of scanner and mobile ROIs. In practice, we envision that a human cytotechnologist would select the 20 ROIs using a microscope and capture them with a mobile phone.

All the models presented in this study were trained exclusively on scanner images, not mobile images. This is purely due to data collection constraints: we did not have enough mobile samples to construct a reasonably sized training set, given that collecting thousands of mobile ROIs for training would be time-consuming. Although this is a limitation, it shows that our model is robust to differences in image modalities. Given more mobile samples, we suspect that incorporating mobile images into the training set would further improve the algorithm's performance and robustness.

The approach we propose could also potentially be fully automated. For example, the diagnosis could be performed entirely using a smartphone equipped with neural networks for ROI detection and classification, but this would require the development of hardware to control ROI capture. Finally, the scale of the evaluation could be larger, including more slides (currently there are 100), more mobile devices (currently we use a Redmi Note 10S phone), and more clinical centers to collect data from (currently all the cases are from a single institution). Also, it is worth noting that our algorithm uses only a single z-stack depth from the WSI (we default to the "middle" depth out of 9 z-stack depths), which may result in some ROIs being blurry and slightly decrease our algorithm's performance. Some of the results presented are not statistically significant (eg, the effect of color augmentations on model performance among indeterminate cases with $P = .158$, effect of histogram matching on naïve model performance with $P = .118$), but we conjecture that differences are likely to become significant with a larger sample. Finally, because our data set consists of FNABs with surgical follow-up, it is likely biased toward clinically concerning nodules requiring resection.

Acknowledgments

This work is supported in part by the National Cancer Institute, National Institutes of Health, Grant/Award Number: 1R21CA268428-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

S.A. was responsible for conceptualization, writing code, running experiments, and writing the manuscript; D.D. was responsible for conceptualization, data curation, writing code, and editing the manuscript; R.D. and J.B. were responsible for data

collection; J.C. and S.K. were responsible for conceptualization and data collection; R.H., D.R., and R.K. were responsible for conceptualization and reviewing the manuscript; W.T.L., L.C., and D.E.R. were responsible for conceptualization, supervision, and editing the manuscript. All authors read and approved the final paper.

Data Availability

The data that support the findings of this study are available from the Duke Office of Translation and Commercialization (OTC), but restrictions apply to the availability of these data, which were used under a pending license agreement for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of Duke OTC.

Funding

This work was supported in part by the National Cancer Institute, National Institutes of Health, Grant/Award Number: 1R21CA268428-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Competing Interest

The authors declare no competing interests.

Ethics Approval and Consent to Participate

After the Institutional Review Board approval was obtained, we searched institutional files for all thyroidectomy specimens with preceding fine-needle aspiration biopsy. We used them to construct the training set and test set used in the study.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.modpat.2023.100129

References

1. Wilson ML, Fleming KA, Kuti MA, Looi LM, Lago N, Ru K. Access to pathology and laboratory medicine services: a crucial gap. *Lancet*. 2018;391(10133): 1927–1938. https://doi.org/10.1016/S0140-6736(18)30458-6
2. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309. https://doi.org/10.1038/s41591-019-0508-1
3. 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. https://doi.org/10.1016/j.media.2020.101814
4. Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med*. 2019;2(1):48. https://doi.org/10.1038/s41746-019-0112-2
5. Isse K, Lesniak A, Grama K, Roysam B, Minervini MI, Demetris AJ. Digital transplantation pathology: combining whole slide imaging, multiplex staining and automated image analysis. *Am J Transplant*. 2012;12(1):27–37. https://doi.org/10.1111/j.1600-6143.2011.03797.x
6. Cibas ES, Ali SZ. The 2017 Bethesda system for reporting thyroid cytopathology. *Thyroid*. 2017;27(11):1341–1346. https://doi.org/10.1089/thy.2017.0500
7. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Preprint. Published online Sep 4, 2014. arXiv:14091556 https://arxiv.org/abs/1409.1556
8. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. *Mobilenetv2: Inverted residuals and linear bottlenecks*. 2018:4510–4520. https://doi.org/10.1109/CVPR.2018.00474
9. Lin M, Chen Q, Yan S. Network in network. Preprint. Published online Dec 16, 2013. arXiv:13124400. https://doi.org/10.48550/arXiv.1312.4400
10. Loshchilov I, Hutter F. Decoupled weight decay regularization. Preprint. Published online Nov 14, 2017. arXiv:171105101. https://doi.org/10.48550/arXiv.1711.05101
11. Dietterich TG. Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer; 2000:Lecture Notes in Computer Science; 1857.
12. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988:837–845. https://doi.org/10.2307/2531595
13. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):1–48. https://doi.org/10.1186/s40537-019-0197-0
14. Müller SG, Hutter F. Trivialaugment: tuning-free yet state-of-the-art data augmentation. Preprint. Published online Mar 18, 2021. arXiv:2103.10158v2. https://doi.org/10.48550/arXiv.2103.10158
15. 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):43. https://doi.org/10.4103/jpi.jpi_43_18
16. Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. *Diagn Cytopathol*. 2018;46(3):244–249. https://doi.org/10.1002/dc.23880
17. Guan Q, Wang Y, Ping B, et al. Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. *J Cancer*. 2019;10(20):4876. https://doi.org/10.7150/jca.28769
18. Kezlarian B, Lin O. Artificial intelligence in thyroid fine needle aspiration biopsies. *Acta Cytol*. 2021;65(4):324–329. https://doi.org/10.1159/000512097
19. Skandarajah A, Reber CD, Switz NA, Fletcher DA. Quantitative imaging with a mobile phone microscope. *PLoS One*. 2014;9(5):e96906. https://doi.org/10.1371/journal.pone.0096906
20. Rivenson Y, Ceylan Koydemir H, Wang H, et al. Deep learning enhanced mobile-phone microscopy. *ACS Photonics*. 2018;5(6):2354–2364. https://doi.org/10.1021/acsphotonics.8b00146
21. de Haan K, Ceylan Koydemir H, Rivenson Y, et al. Automated screening of sickle cells using a smartphone-based microscope and deep learning. *NPJ Digit Med*. 2020;3(1):1–9. https://doi.org/10.1038/s41746-020-0282-y
22. 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(3):181–187. https://doi.org/10.1002/cncy.21631
23. 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. https://doi.org/10.1002/cncy.22238
24. 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. https://doi.org/10.1089/thy.2015.0020