Title

Impact of pre-analytical variables on deep learning accuracy in histopathology

Journal

Histopathology, 75(1)

ISSN

0309-0167

Authors

Jones, Andrew D

Graff, John Paul

Darrow, Morgan

et al.

Publication Date

2019-07-01

DOI

10.1111/his.13844

Peer reviewed

HHS Public Access

Author manuscript

Histopathology. Author manuscript; available in PMC 2020 July 01.

Published in final edited form as:

Histopathology. 2019 July ; 75(1): 39–53. doi:10.1111/his.13844.

Impact of pre-analytic variables on deep learning accuracy in histopathology

Andrew D. Jones,
Department of Pathology & Laboratory Medicine, University of California Davis Health

John Paul Graff,
Department of Pathology & Laboratory Medicine, University of California Davis Health

Morgan Darrow,
Department of Pathology & Laboratory Medicine, University of California Davis Health

Alexander Borowsky,
Department of Pathology & LaboratoryMedicine, University of California Davis Health

Kristin A. Olson,
Department of Pathology & LaboratoryMedicine, University of California Davis Health

Regina Gandour-Edwards,
Department of Pathology & Laboratory Medicine, University of California Davis Health

Ananya Datta Mitra,
Department of Pathology & Laboratory Medicine, University of California Davis Health

Dongguang Wei,
Department of Pathology & LaboratoryMedicine, University of California Davis Health

Guofeng Gao,
Department of Pathology & Laboratory Medicine, University of California Davis Health

Blythe Durbin-Johnson, and
University of California Davis, Division of Biostatistics

Hooman H. Rashidi*
Department of Pathology & Laboratory Medicine, University of California Davis Health

Abstract

*Corresponding Author: 4400 V Street, PATH BLDG, Sacramento, California, USA 95817, hrashidi@ucdavis.edu, +1.916.734.2525.

Author Contributions
Study design: ADJ, JPG, HHR
Preparation of manuscript and figures: ADJ, AB, HHR
Collection of slides and images: ADJ, ADM, DW, GG, HHR
Collection and tabulation of data: ADJ, JPG, HHR
Statistical analysis: BDJ
Board certified pathologist image review: JPG, MD, AB, KAO, RGE, HHR
Review of images, text, figures, and data: ADJ, JPG, MD, AB, KAO, RGE, ADM, DW, GG, HHR

Conflict of Interests
All authors have no competing or conflicting interests to declare.

Aims—Machine Learning (ML) binary classification in diagnostic histopathology is an area of intense investigation. Several assumptions, including training image quality/format and the number of training images required, appear similar in many studies irrespective of the paucity of supporting evidence.

Methods—We empirically compared training image file type, training set size, and two common Convolutional Neural Networks (CNN) using transfer learning (ResNet50, SqueezeNet). Thirty H&E slides with carcinoma or normal tissue from three tissue types (breast, colon, prostate) were photographed generating 3,000 partially overlapping images (1,000 per tissue type). These lossless PNGs were converted to lossy JPGs. Tissue type-specific binary classification ML models were developed using all PNG or JPG images, and repeated with a subset of 500, 200, 100, 50, 30, and 10 images. Eleven models were generated for each tissue type, at each quantity of training images, for each file type, and for each CNN, resulting in 924 models. Internal accuracies and generalization accuracies were compared.

Results—There was no meaningful significant difference between accuracies in PNG vs JPG models. Models trained with more images did not invariably perform better. ResNet50 typically outperformed SqueezeNet. Models were generalizable within a tissue type but not across tissue types.

Conclusions—Lossy JPG images were not inferior to lossless PNG images in our models. Large numbers of unique H&E slides were not required for training optimal ML models. This reinforces the need for an evidence-based approach to best practices for histopathologic ML.

Introduction

Many studies have been designed to answer specific analytic questions and test narrow analytic hypotheses regarding the tremendous potential of machine learning (ML) and artificial intelligence (AI) in histopathologic analysis.[1] While there have been calls for pathologists to become “information specialists” rather than diagnosticians,[2] little attention has been paid to understanding the best practices involved in creating ML models for pathologic image analysis—that is, the potential generalization of a given model. As a result, we know little about the way the analysis by the machine is influenced by the myriad choices that go into the pre-analytic phase of model generation.

In this study, we attempt to address the issue of pre-analytic variation in one specific domain. Convolutional Neural Network (CNN) platforms do not “see” histologic images in the same way that expert pathologists do,[3] raising the question of how the input file type and quality impact the efficacy of an ML model. Two common image file types, Portable Networks Graphics (or PNG) and Joint Photographic Experts Group (or JPG), were evaluated to determine whether input from one or the other would generate superior predictive models. More than an academic question, PNG files are “lossless” resulting in high quality at the expense of large files. JPG files, while grossly similar in appearance and quality to their PNG counterparts, are “lossy” and are typically significantly smaller in size. Smaller file size means the images are more frequently encountered, easier to transmit electronically, cheaper to store, and result in less training time when generating CNN models. While previous studies of other CNN architectures suggest that relatively high

quality compressed histopathologic JPGs perform well,[4] it is unclear whether the additional data in the PNG files would result in models with superior accuracy to those trained with JPG images with the CNNs and tissue types tested here.

We evaluated three different tissue types and trained two different CNNs using a variable number of training images to dichotomize histopathologic image data into benign or invasive carcinoma, using either PNG or JPG images as the training data set. In so doing, we created and tested 924 distinct models. We then tested the generalization of these models by inputting a vast array of images of variable quality obtained from sources distinct from the training images. Throughout the study, multiple board-certified pathologists vetted the images to ensure each image contained unambiguous diagnostic material that would permit a human (and, by extension, a machine) to readily sort the image into benign or invasive carcinoma. This rigorous pre-analytic control has allowed us to better understand how real-world application of machine learning in histopathology can be implemented. We hypothesize that leveraging this vigorous approach will provide statistical evidence that PNG and JPG training sets produce similar accuracy throughout various CNNs and pathology subspecialties.

Methods

IRB approval (IRB ID 1286225–1) was obtained and ten histopathologic slides were selected for each tissue type (breast, colon, and prostate). (See Figure 1) All slides were scanned at 20× magnification and were scanned using a variety of scanner manufacturers (e.g. Leica Biosystems Aperio XT, etc). Five slides for each tissue type showed unambiguous benign findings (excluding proliferative or intermediate lesions, and excluding benign mimics of carcinoma), and five slides for each tissue type showed unambiguous invasive carcinoma. Ten slides for each tissue type were found to produce valid models in a prior analysis. Whole slide images were obtained and 100 PNG images from each slide were generated, with 50 images taken at 4× magnification and 50 images taken at 10× magnification, resulting in 3,000 images total (1,000 images in 3 subspecialties). Images were obtained using operating system level screen capture tools. These 3,000 images were then evaluated by board-certified pathologists to ensure each image contained unambiguous benign or malignant findings. The amount of mesenchymal and epithelial elements varied from image to image, but each image was classifiable by a board-certified pathologist as showing unambiguous benign or malignant features. Histologic artifacts (e.g. tissue folds, microtome chatter, staining irregularities, etc) were not excluded, provided the resulting image could be unambiguously categorized by a pathologist.

These 3,000 high-quality PNG images were converted into a set of corresponding medium-quality (a setting of 2 on a scale from 1 to 4, with 4 being maximum quality) JPG images (see Figure 2). The total images from each tissue type (1,000 PNG and 1,000 JPG each for breast, colon, and prostate) were used to train 11 models each on two different CNNs, ResNet50 and SqueezeNet. The file set was then cut in half (resulting in 500 PNG and 500 JPG images each for breast, colon, and prostate), ensuring that equal proportions of 4× and 10× magnification images were in each set, and were used to train 11 models each on the

same two CNNs. This process was repeated for 200, 100, 50, 30, and 10 training images. This resulted in 924 unique models.

Transfer learning technique was used in our study. Transfer learning allowed us to retrain a well-established deep CNN (e.g. ResNet-50, SqueezeNet) to build our new deep CNN, specific to our classification task (e.g. a CNN model that is trained on breast cancer to be breast cancer specific or one that is trained on prostate cancer to be prostate cancer specific). The original CNN is retrained while retaining information on its core trained classes (greater than 1000 classes/entities). These core image classes in the original CNN (e.g. cars, trees, etc.) serve as the foundation for our new retrained CNN since many of the hidden layers within that CNN can be transferred and applied to our models. The idea is that these earlier hidden layers in the core CNN are representative of shared features within most images regardless of their content (e.g. delineated edge detection, overall shape, etc.).

This approach allows us to keep the aspects of the original CNN that are shared with our new classification categories while allowing us to fine-tune our new CNN to render it specific to the task at hand. During this retraining process, the training performance is measured by cross-entropy loss function to display the learning progress of our new model. The parameter that initially measures this task while in the training mode is the model's calculated "training accuracy" which calculates the percentage of accurately labeled images on the training batch using a random 5% subset of the images within the training set. The training steps used in the transfer learning process were done through the Turi Create python script library for retraining the ResNet50 and the SqueezeNet CNNs.

The final weights and the number of optimized steps were saved based on the protocol buffer. All of the images were resized to 224×224 dimension (the required dimension for ResNet50 and SqueezeNet models within these neural networks). Following this initial validation, a "validation accuracy" was then calculated which involved testing another set of labeled images that were not used during the training batch. In our study, 80% of the images were used in the training batch while a random 20% of the images were set aside during training and tested for this validation accuracy step, hereafter known as the Internal Validation Accuracy. This approach allowed us to minimize the possibility of overfitting in our models which is thought to correlate with its generalization capability.

Subsequently, an external set of images that were completely unknown to our training set were used to test each model's true generalizability. This secondary external training set was obtained from public domain images selected via Google search and downloaded through a Fatkun batch process. These external test set images were available exclusively in JPG format.

These images were treated as unknowns and similarly evaluated by our board-certified pathologists to ensure they unambiguously showed either benign tissue or invasive carcinoma for each tissue type (breast, colon, and prostate). This resulted in 82 test images for breast models, 50 test images for colon models, and 70 test images for prostate models. Models were tested against their respective tissue subtypes and the accuracy (hereafter referred to as External Validation Accuracy) of each was recorded. Testing each of our fixed

optimized retrained CNN models against its respective external test set allowed us to assess the generalizability of these models and potential overfitting.

Results

Accuracy results for 924 individual model tests were recorded. Accuracy was compared between JPG and PNG files using ANOVA models. These models included effects for file type, number of images, ML model, tissue, and all two-, three-, and four-way interactions between file type, number of images, ML model, and tissue. Analyses were conducted using R, version 3.5.1 (R Core Team, 2018).

PNG vs JPG Overall

The overall combined internal validation mean accuracy of the ResNet50 models (including all tissue types and image training quantities) was 91.8% with PNG images and 91.0% with JPG images (difference in means: −0.8%; 95% CI: −3% to 1.5%; p=0.493); for SqueezeNet models, the mean accuracy was 92.5% with PNG images and 92.9% with JPG images (difference in means: 0.3%; 95% CI: −1.9% to 2.6%; p=0.774). (See Table 1 and Figure 3)

The overall combined external test set mean accuracy (i.e. External Validation Accuracy) of the ResNet50 models (including all tissue types and image training quantities) was 77.7% with PNG images and 78.7% with JPG images (difference in means: 1%; 95% CI: 0.1% to 1.9%; p=0.025); for SqueezeNet models, the mean accuracy was 72.8% with PNG images and 72.5% with JPG images (difference in means: −0.3%; 95% CI: −1.2% to 0.6%; p=0.528). (See Table 2 and Figure 3)

PNG vs JPG in Different Training Set Image Quantities

The mean accuracies of models trained with different numbers of images in the training sets were evaluated (combining all tissue types and CNNs); internal validation accuracy comparisons between PNG and JPG did not achieve significance. (See Table 3) The mean accuracies of these models were similarly compared when tested against the external test set. One comparison, training with 50 images, achieved statistical significance (p=0.047) with a mean accuracy of JPG superior to PNG (74.5% to 72.8%; difference in means 1.7%); however, the 95% confidence interval includes zero (95% CI: 0%, 3.4%). (See Table 4)

PNG vs JPG in Different CNNs and Training Set Image Quantities

The mean accuracies of models trained with different numbers of images in the training sets were evaluated by CNN (combining all tissue types). Internal validation mean accuracy comparisons between PNG and JPG showed a statistically significant difference in one group (SqueezeNet 10: PNG, 73.7%; JPG, 84.3%; p=<0.001). External test set mean accuracy comparisons between PNG and JPG showed a statistically significant difference in one group (ResNet50 50: PNG, 72.6%; JPG, 77.0%; p=<0.001). (See and Table 5 and Table 6) In both cases, JPG outperformed PNG.

PNG vs JPG in Different Tissues and CNNs

The mean accuracies of models trained with different tissue types were evaluated by CNN (combining all quantities of images in training sets). Internal validation mean accuracy comparisons between PNG and JPG showed no statistically significant difference. (See Table 7) External test set mean accuracy comparisons between PNG and JPG showed a statistically significant difference in colon models with the ResNet50 CNN (PNG: 80.6%; JPG: 82.3%; 95% CI: 0.2%, 3.3%; p=0.025). (See Table 8 and Figure 4)

PNG vs JPG in Different Tissues, CNNs, and Training Set Image Quantities

The mean accuracies of models trained with different tissue types, different CNNs, and with different quantities of images in the training sets were evaluated. Internal validation comparisons between PNG and JPG showed three categories that achieved statistically significant differences in mean accuracies, favoring JPG over PNG in 2 out of 3 cases (see Table 9 for details). Similarly, external test set comparisons between mean accuracies of PNG and JPG models showed four categories that achieved statistically significant differences, each favoring JPG over PNG (see Table 10, Figure 5, Figure 6, and Figure 7 for details). The highest accuracy for each CNN by tissue type and file type is shown in Figure 8.

Model Training Specificity by Pathology Subspecialty

Finally, the best and worst models for each tissue type on any CNN, trained with PNG or with JPG images were compared against other tissue type external test sets to assess the specificity of each model. Models performed substantially better when tested against tissue of the type with which they were trained than when compared with other tissue types (i.e. a model trained with breast images performed better when tested against the breast external image test set than when tested against either colon or prostate test sets). (See Figure 9)

Discussion

Many previous works have tried to answer highly specific analytic questions regarding image analysis and histopathology, including tagging images to train neural networks,[5] identifying and classifying mitoses[6,7] and other morphologic features,[8] distinguishing stroma from epithelium,[9] identifying lymph node and tissue metastases,[10–12] differentiating tumor grades,[13] whole slide image classification,[14–16] pre-screening slides for potential cancer,[17] interpreting IHC and other biomarkers,[18] and even predicting the genetic aberration underlying a cancer.[19] Many agree that transfer learning techniques provide the most useful path forward in CNN image analysis.[3,20–22] However, there are few instances in the pathology literature that attempt to explain which pre-analytic variables should be controlled when creating ML/AI models to evaluate histopathologic image data. While it is enticing to generate a single model with impressive internal validation accuracy, each model created (even if the same training image set is utilized) generates unique neuron weights and testing outcomes (in part due to the randomized generation of the internal validation image set from the original training image set). This means that any single model is not likely to be exactly reproduceable by researchers using the same CNN and training images. Indeed, we found variability in internal validation accuracy scores, which typically increased as the

number of images in the training set decreased. We believe that creating 11 models with each CNN at each quantity of training images (i.e. 1,000, 500, 200, 100, 50, 30, and 10 images) helped normalize this random variability.

Further, while internal validation accuracy is useful in determining whether one's training image set is sufficiently distinct (i.e. in this case, the benign images are sufficiently different from the invasive carcinoma images), it does not necessarily correlate with accuracy when images separate from the training set are evaluated.[23] That is, the internal validation accuracy does not give a measure of generalization of the model. We attempted to address that concern by utilizing completely separate image sets for training and testing, the latter coming from various internet sources and of variable magnification, color profiles, and quality. To ensure this hodgepodge of images could accurately evaluate the generalization of our models, all images were verified by board-certified pathologists as containing unambiguous diagnostic features. This allowed us to assess the generalization of our models; i.e. how they would perform against a "real world" test. Overall, models showed a drop in accuracy, but many subsets continued to perform exceptionally well and approached their respective internal validation accuracies.

Finally, to show that our models trained with one tissue type were generalizable only to the extent that they could evaluate similar types of tissue, we tested the best and worst performing models for each tissue type against test sets from different tissues. The results show that models trained with colon histopathologic images, for instance, work well when evaluating other colonic histopathology, but not when evaluating breast or prostate histopathology. This result held regardless of whether the models were trained with PNG or JPG images. The same trend noted in the colon models was also seen in the prostate and breast models.

We have attempted to show, through a rigorous approach that resulted in 924 unique ML models, whether a difference exists in the accuracy of image analysis based on the file type of image used in training. While several of our models' mean accuracies achieved a statistically significant difference when comparing lossless PNG-trained CNNs to lossy JPG-trained CNNs, these differences were sporadic, small in magnitude, and often without a meaningful clinical difference in actual performance. Additionally, almost all instances of statistical significance favored models trained with JPG rather than PNG images, suggesting that there is little to no benefit to training models with larger PNG images. This is especially beneficial, as the smaller size and increased portability of JPG images makes model generation faster and reduces storage requirements. In our image set, the average difference in PNG vs JPG file size was an order of magnitude, a significant amount of storage when considering the thousands of images involved in training CNNs. This also potentially opens doors for more input modalities, as images captured with portable cameras and digital microscopy equipment are most frequently JPG.

Limitations of the study include the categorization schema which separated lesions into only two categories (i.e. benign or malignant). It is unclear whether categorization into greater than two categories could achieve similar levels of accuracy and generalization with the approach utilized here. Additionally, there may be data within the lossless PNG files that is

not necessary for pathologists or machines to make a benign vs malignant categorization, but which may aid in other predictive algorithms (e.g. predicting molecular phenotype, invasive potential, categorizing intermediate and proliferative lesions, etc). This study suggests that a limited number of slides (i.e. 10) can suffice and can produce a valid model through transfer learning for prediction; it is unclear whether including more slides in the training set would produce better (or more generalizable) results. It is also unclear whether utilization of alternate whole slide scanning hardware and software would produce variant results. Finally, it is unclear if similar findings could be obtained using higher magnification (i.e. 40×). Further research is necessary to address these and other pre-analytic conditions.

It is our hope that other researchers will continue to evaluate the pre-analytic variables that should be considered prior to embarking on a study of the enormous potential ML/AI tools can offer pathologists and our patients. When the trust in our diagnostic tools and the safety of our patients is at stake, we must ensure that we maintain the high standard of analytic rigor we demand in other aspects of our diagnostics toolkit.

Acknowledgements

Parts of this project described were supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1 TR001860. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Djuric U, Zadeh G, Aldape K, Diamandis P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ Precis Oncol. 2017;1(1):22. [PubMed: 29872706]
2. Jha S, Topol EJ. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. JAMA. 2016;316(22):2353–2354. [PubMed: 27898975]
3. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J. 2018;16:34–42. [PubMed: 30275936]
4. Samuel Dodge LK. Understanding How Image Quality Affects Deep Neural Networks. International Conference on Quality of Multimedia Experience (QoMEX) 2016.
5. Qu J, Hiruta N, Terai K, Nosato H, Murakawa M, Sakanashi H. Gastric Pathology Image Classification Using Stepwise Fine-Tuning for Deep Neural Networks. J Healthc Eng. 2018;2018:8961781. [PubMed: 30034677]
6. Wahab N, Khan A, Lee YS. Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. Comput Biol Med. 2017;85:86–97. [PubMed: 28477446]
7. Pan X, Li L, Yang H, et al. Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks. Neurocomputing. 2017;229:88–99.
8. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform. 2016;7:29. [PubMed: 27563488]
9. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing. 2016;191:214–223. [PubMed: 28154470]
10. David F Steiner M, Robert MacDonald, Yun Liu, Peter Truszkowski, Hipp Jason D., Christopher Gammage, Florence Thng, Lily Peng, and Stumpe Martin C.. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. American Journal of Surgical Pathology 2018;00(00).

11. Liu Y, Kohlberger T, Norouzi M, et al. Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection. Arch Pathol Lab Med. 2018.
12. Mermel MSaC. Applying Deep Learning to Metastatic Breast Cancer Detection Google AI Blog. Vol 2018: Google AI; 2018.
13. Mina Khoshdeli AB, Bahram Parvin. Deep Learning Models Differentiate Tumor Grades from H&E Stained Histology Sections. Paper presented at: Conf Proc IEEE Eng Med Biol Soc; 29 October 2018, 2018; Honolulu, HI.
14. Sharma H, Zerbe N, Klempert I, Hellwich O, Hufnagl P. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. Comput Med Imaging Graph. 2017;61:2–13. [PubMed: 28676295]
15. Kamyar Nazeri AA, and Mehran Ebrahimi. Two-Stage Convolutional Neural Network for Breast Cancer Histology Image Classification. International Conference on Image Analysis and Recognition; 2018; Portugal
16. Le Hou DS, Tahsin M. Kurc Yi Gao, Davis James E., and Saltz Joel H.. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016:2424–2433.
17. Kayser K, Gortler J, Bogovac M, et al. AI (artificial intelligence) in histopathology--from image analysis to automated diagnosis. Folia Histochem Cytobiol. 2009;47(3):355–361. [PubMed: 20164018]
18. Sheikhzadeh F, Ward RK, van Niekerk D, Guillaud M. Automatic labeling of molecular biomarkers of immunohistochemistry images using fully convolutional networks. PLoS One. 2018;13(1):e0190783. [PubMed: 29351281]
19. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559–1567. [PubMed: 30224757]
20. Robertson S, Azizpour H, Smith K, Hartman J. Digital image analysis in breast pathology-from image processing techniques to artificial intelligence. Transl Res. 2018;194:19–35. [PubMed: 29175265]
21. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine. 2018;27:317–328. [PubMed: 29292031]
22. Brady Kieffer MB, Kalra Shivam, Tizhoosh HR. Convolutional Neural Networks for Histopathology Image Classification: Training vs. Using Pre-Trained Networks. International Conference on Image Processing Theory, Tools and Applications; 2017; Montreal, CA.
23. Andrew H Beck ARS, Samuel Leung, Marinelli Robert J., Nielsen Torsten O., van de Vijver Marc J., West Robert B., van de Rijn Matt, Koller Daphne. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. Science Translational Medicine. 2011;3(108).

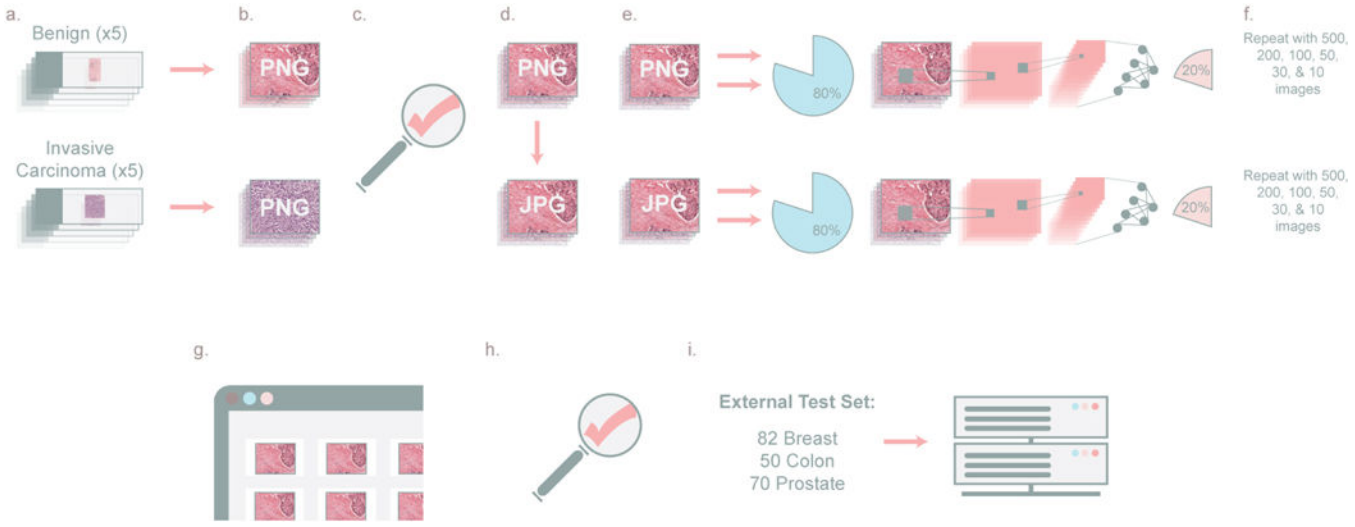

Figure 1:

Overview of Methods. **a.** Ten slides, five benign and five invasive carcinoma, were chosen for each tissue type (breast, colon, and prostate). **b.** Lossless PNG images were obtained from each slide (100 images per slide, 50 at 4× and 50 at 10×). **c.** Multiple board certified pathologists reviewed the 3,000 captured images to ensure unambiguous diagnostic material on each slide. **d.** Lossless PNG files were converted to medium quality lossy JPG files. **e.** Eleven models using ResNet50 and 11 models using SqueezeNet were generated for the 1,000 PNG images for each tissue type, and again for 1,000 JPG images for each tissue type. Eighty percent of the images were used to train the model, while a random 20% were kept in reserve to determine the internal validation accuracy of each model generated. **f.** This process is repeated using 500 PNG and 500 JPG images, with 200 images, 100 images, 50 images, 30 images, and 10 images. **g.** The external test set of benign and carcinoma images for each tissue type was obtained from an internet search. **h.** Multiple board certified pathologists reviewed the external test set to ensure unambiguous diagnostic material was present in each image. **i.** In all, 82 breast images, 50 colon images, and 70 prostate images were used to test 924 independent models.

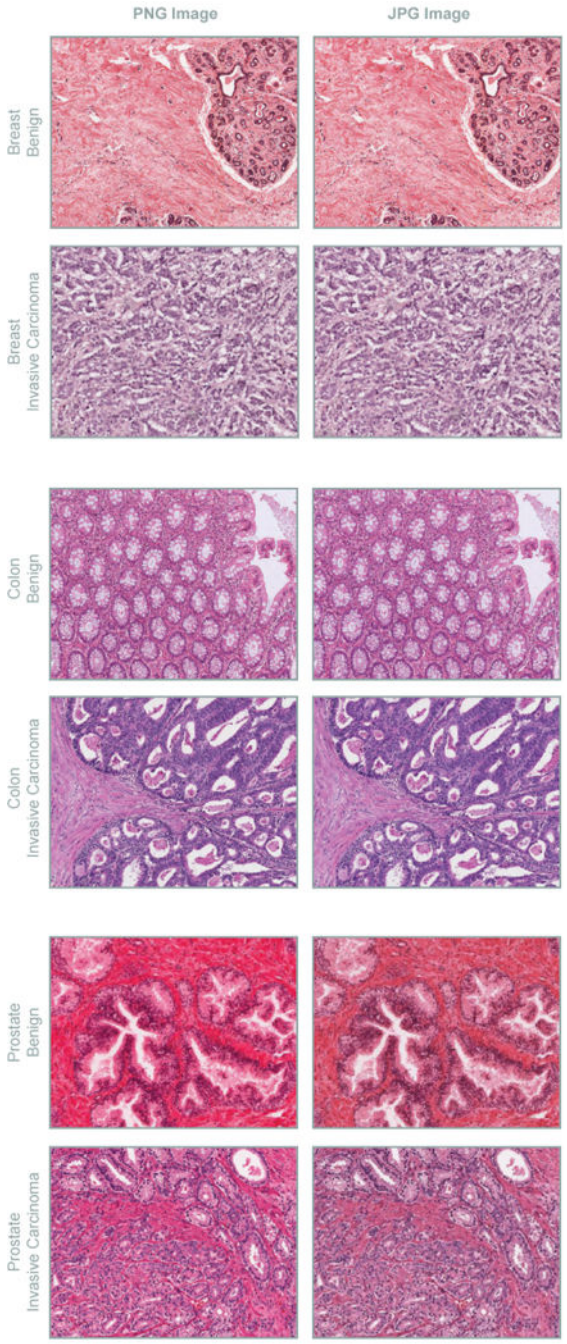

Figure 2:
Representative lossless PNG and medium-quality lossy JPG images from each tissue type. While the color profiles differ slightly between the PNG and JPG images, the overall quality is not grossly different.

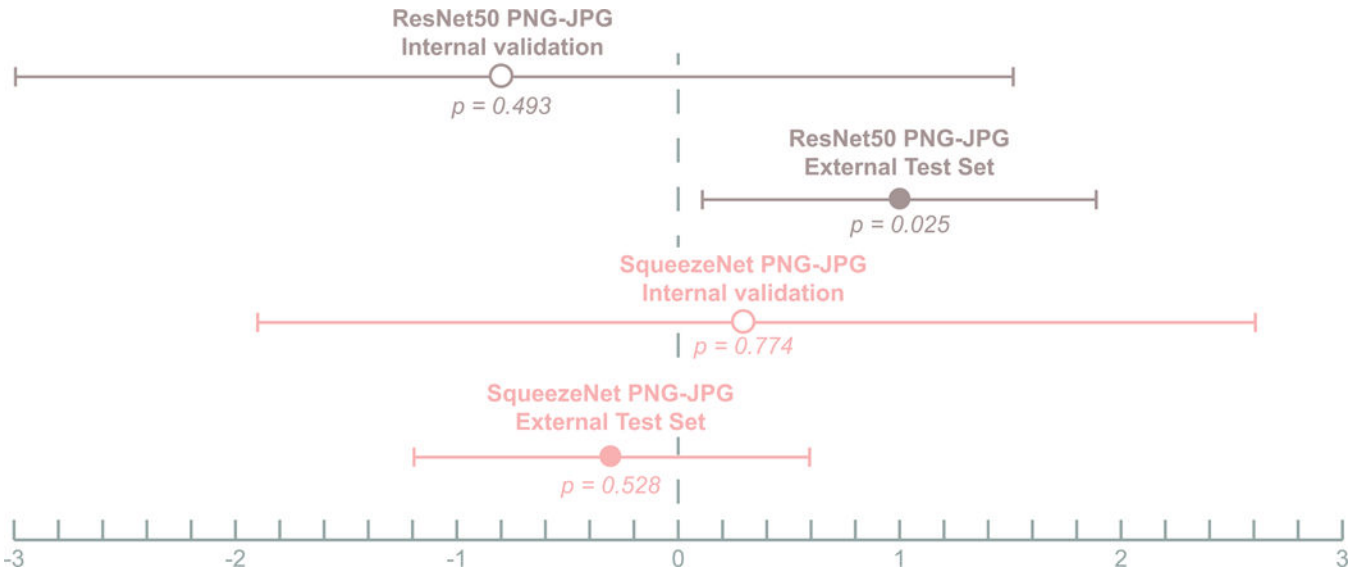

Figure 3:

Comparison between difference in mean accuracies with 95% confidence intervals of ResNet50 (combined breast, colon, and prostate with all image quantity training sets) and SqueezeNet when comparing PNG vs JPG trained models. The difference in mean accuracy between PNG and JPG models was significant only for ResNet50 when evaluated against the external test images.

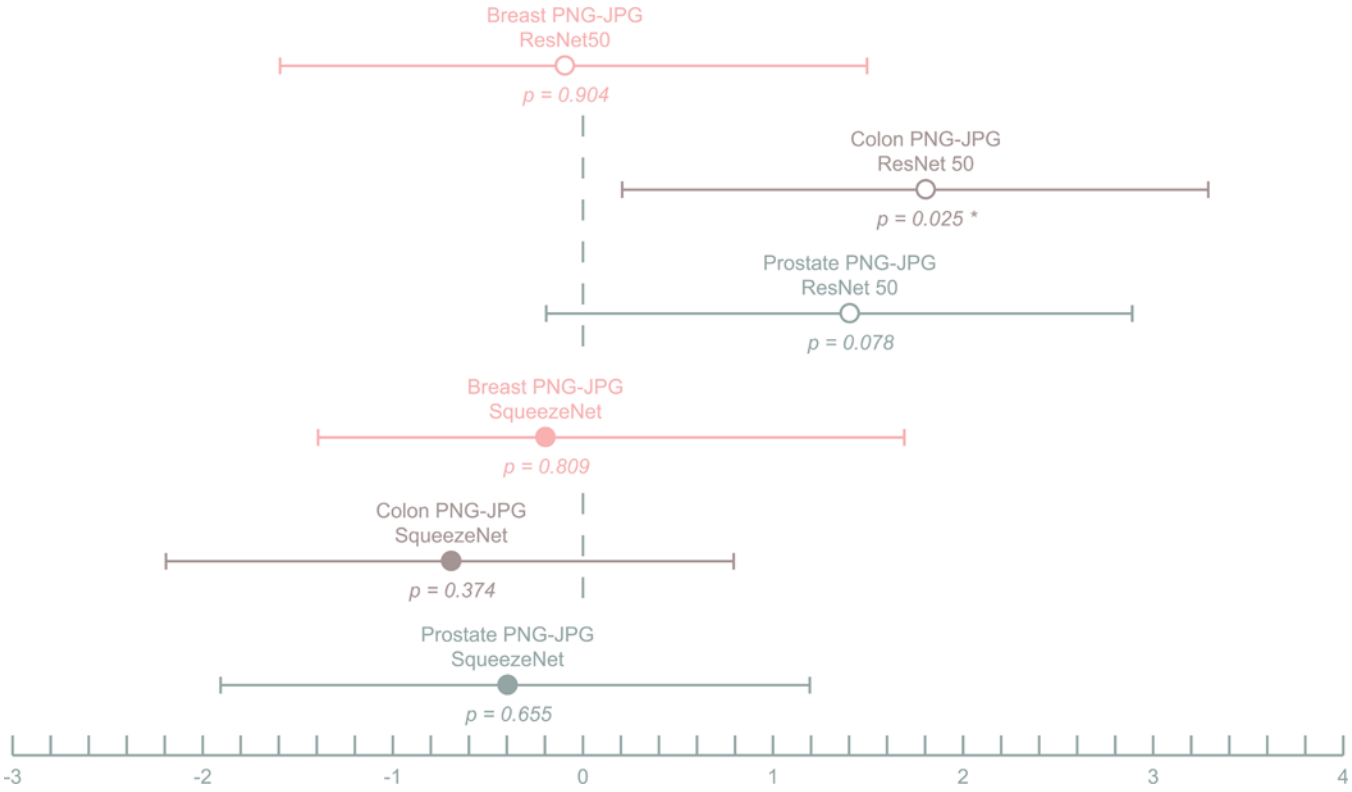

Figure 4:

Comparison between difference in mean accuracies with 95% confidence intervals of tissue-specific models (combined 10 – 1,000 image training sets) when tested against external text sets comparing PNG vs JPG trained models. Open circles represent ResNet50 CNN; closed circles represent SqueezeNet CNN. The difference in mean accuracy between PNG and JPG models for Colon ResNet50 with PNGs was 80.6%, while the mean accuracy with JPGs was 82.3%. This represents a statistically significant but clinically insubstantial difference.

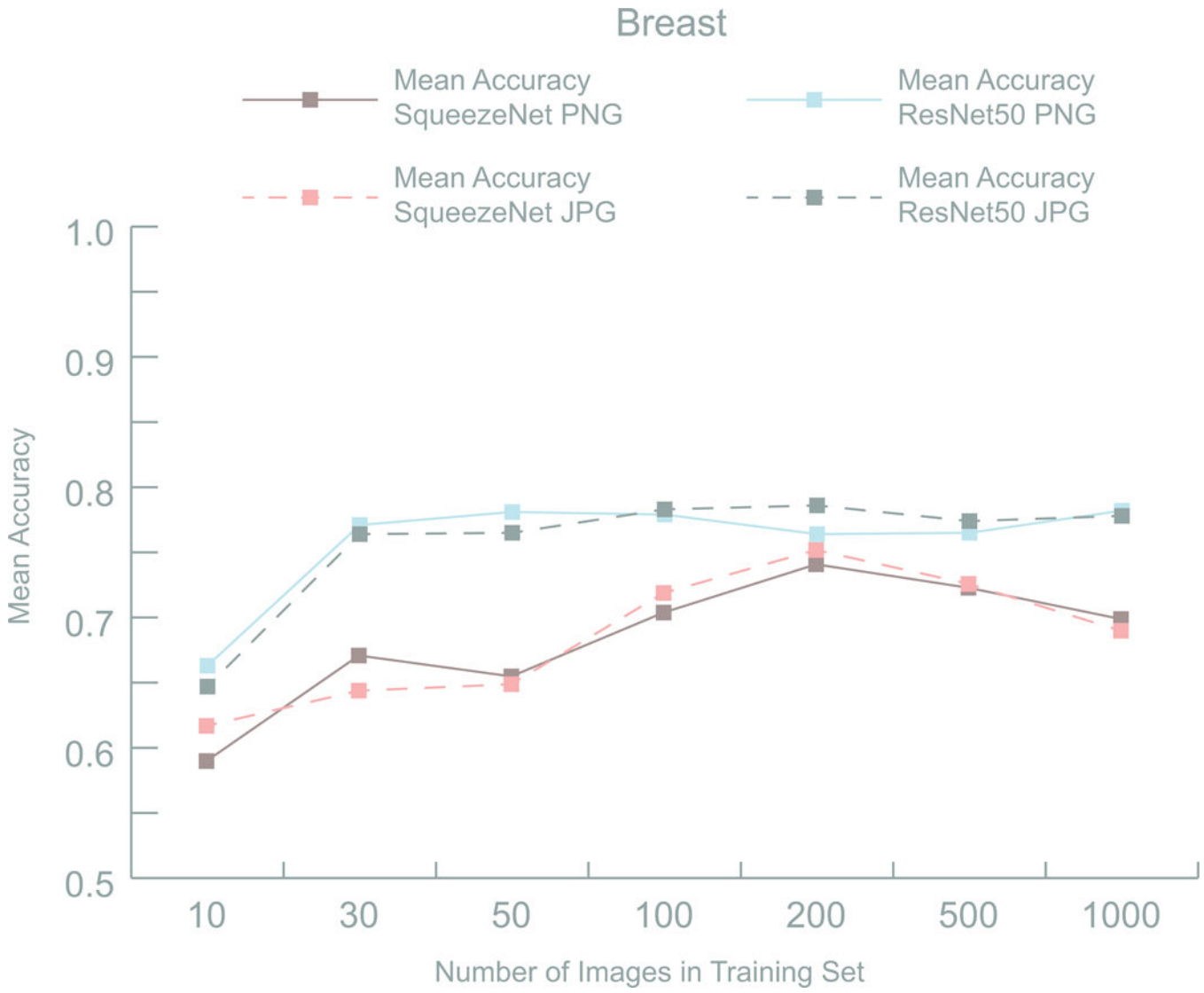

Figure 5:
Comparison of mean external test set accuracies for breast models. No significant difference in means of PNG and JPG ML models was detected at any quantity of training images.

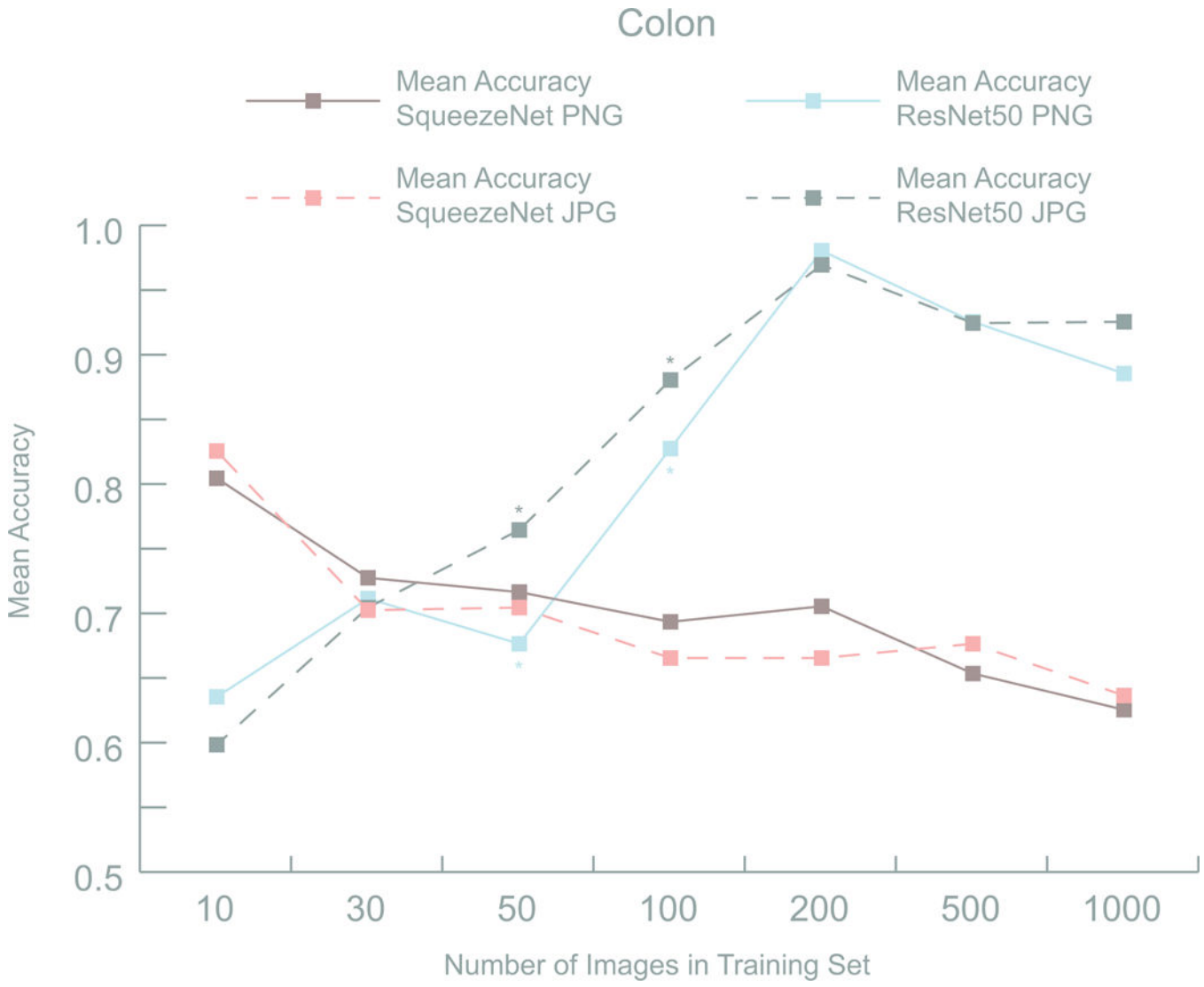

Figure 6:

Comparison of mean external test set accuracies for colon models. A significant difference in mean accuracy between PNG and JPG trained models was detected in the ResNet50 models using 50 and 100 training images.

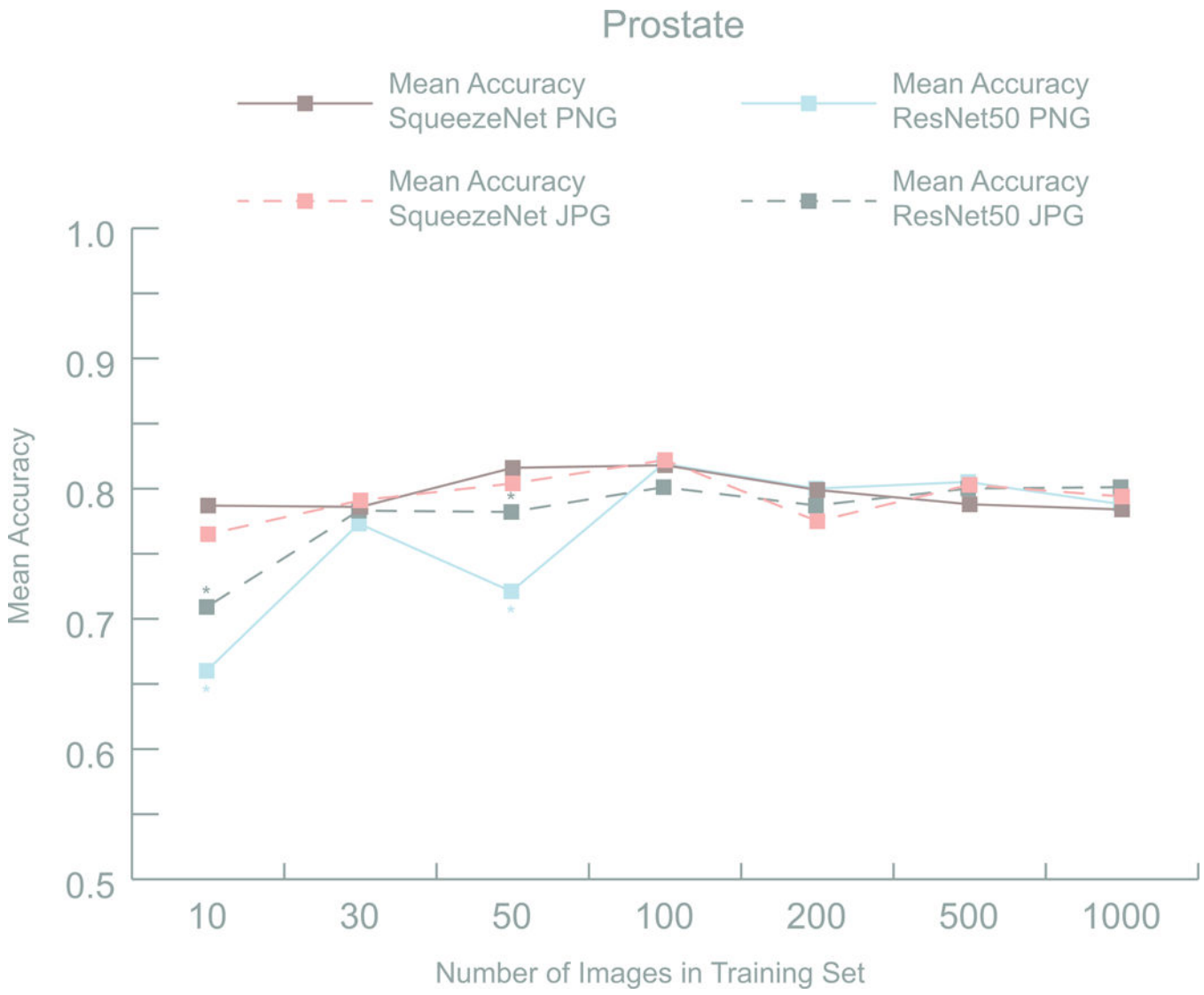

Figure 7:

Comparison of mean external test set accuracies for prostate models. A significant difference in mean accuracy between PNG and JPG trained models was detected in the ResNet50 models using 10 and 50 training images.

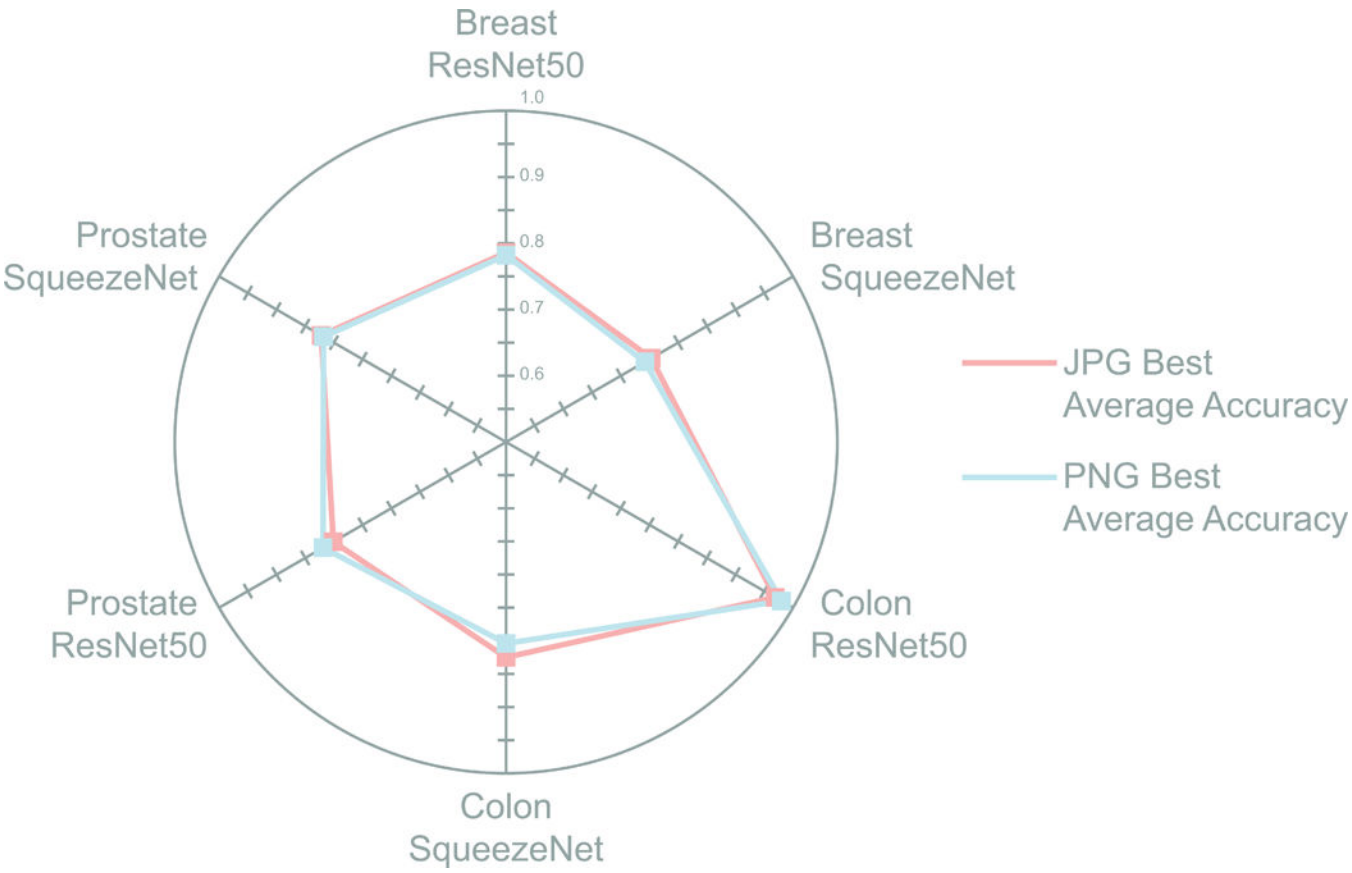

Figure 8:
Comparison of best mean accuracy for each CNN, tissue type, and file type. No significant difference between PNG and JPG trained models was detected.

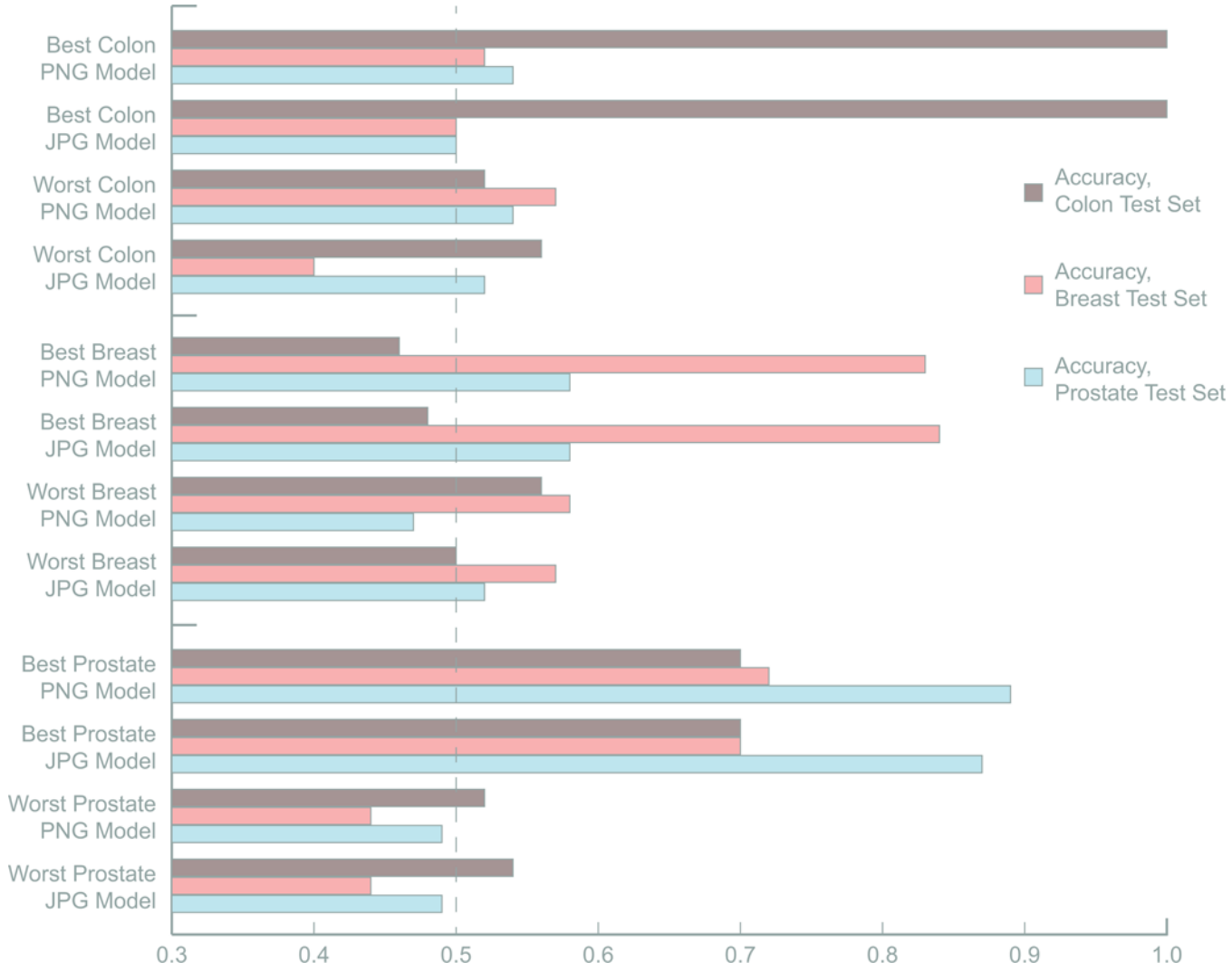

Figure 9:

Cross-discipline comparison of models showing specificty to tissue type. The best performing (genarilizable) models in each tissue show clear specificity for tissue type when trained with either PNG or JPG images.

Table 1:

Comparison of accuracy between filetypes by ML Model, Internal Validation Test Set. Includes all tissue types and quantities of image training sets.

ML Model	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Resnet50	91%	91.8%	−0.8% (−3%, 1.5%)	0.493
SqueezeNet	92.9%	92.5%	0.3% (−1.9%, 2.6%)	0.774

Table 2:

Comparison of accuracy between filetypes by ML Model, External (Google images) Test Set. Includes all tissue types and quantities of image training sets.

ML Model	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Resnet50	78.7%	77.7%	1% (0.1%, 1.9%)	0.025 *
SqueezeNet	72.5%	72.8%	−0.3% (−1.2%, 0.6%)	0.528

Table 3:

Comparison of accuracy between filetypes by Number of Images, Internal Test Set. Includes all tissue types and all neural networks.

Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
10	72%	68.4%	3.5% (−0.7%, 7.7%)	0.098 .
30	89.6%	93.8%	−4.2% (−8.4%, 0%)	0.051 .
50	94.3%	94.7%	−0.4% (−4.6%, 3.8%)	0.843
100	95.8%	95.7%	0.1% (−4.1%, 4.3%)	0.954
200	96.4%	96.6%	−0.1% (−4.3%, 4%)	0.947
500	97.6%	98.1%	−0.5% (−4.7%, 3.7%)	0.818
1000	97.8%	97.9%	0% (−4.2%, 4.2%)	0.990

Table 4:

Comparison of accuracy between filetypes by Number of Images, External Test Set. Includes all tissue types and all neural networks.

Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
10	69.4%	69%	0.4% (−1.3%, 2.1%)	0.641
30	73.1%	74%	−0.8% (−2.5%, 0.8%)	0.324
50	74.5%	72.8%	1.7% (0%, 3.4%)	0.047 *
100	77.9%	77.4%	0.5% (−1.2%, 2.2%)	0.573
200	78.9%	79.8%	−0.9% (−2.6%, 0.8%)	0.291
500	78.4%	77.7%	0.7% (−1%, 2.4%)	0.399
1000	77.1%	76.1%	1% (−0.7%, 2.7%)	0.234

Table 5:

Comparison of accuracy between filetypes by ML Model and Number of Images, Internal Test Set. Includes all tissue types.

ML Model	Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Resnet50	10	59.6%	63.1%	−3.5% (−9.5%, 2.4%)	0.242
Resnet50	30	90.8%	94.8%	−4.1% (−10%, 1.9%)	0.180
Resnet50	50	96.1%	93.6%	2.5% (−3.5%, 8.4%)	0.416
Resnet50	100	95.9%	96.6%	−0.7% (−6.6%, 5.2%)	0.818
Resnet50	200	97.6%	97.2%	0.5% (−5.5%, 6.4%)	0.881
Resnet50	500	98.6%	98.4%	0.2% (−5.7%, 6.2%)	0.935
Resnet50	1000	98.5%	98.8%	−0.3% (−6.3%, 5.6%)	0.909
SqueezeNet	10	84.3%	73.7%	10.6% (4.7%, 16.5%)	<0.001 ***
SqueezeNet	30	88.5%	92.8%	−4.3% (−10.2%, 1.6%)	0.156
SqueezeNet	50	92.6%	95.9%	−3.3% (−9.2%, 2.6%)	0.274
SqueezeNet	100	95.8%	94.8%	0.9% (−5%, 6.9%)	0.755
SqueezeNet	200	95.2%	95.9%	−0.7% (−6.7%, 5.2%)	0.808
SqueezeNet	500	96.5%	97.8%	−1.2% (−7.1%, 4.7%)	0.684
SqueezeNet	1000	97.2%	96.9%	0.3% (−5.6%, 6.2%)	0.923

Table 6:

Comparison of accuracy between filetypes by ML Model and Number of Images, External Test Set. Includes all tissue types.

ML Model	Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Resnet50	10	65.2%	65.2%	−0.1% (−2.4%, 2.3%)	0.944
Resnet50	30	75%	75.1%	−0.1% (−2.5%, 2.2%)	0.922
Resnet50	50	77%	72.6%	4.4% (2.1%, 6.8%)	<0.001 ***
Resnet50	100	82.1%	80.9%	1.3% (−1.1%, 3.6%)	0.295
Resnet50	200	84.7%	84.8%	−0.1% (−2.4%, 2.3%)	0.962
Resnet50	500	83.2%	83.2%	0.1% (−2.3%, 2.4%)	0.959
Resnet50	1000	83.5%	81.8%	1.7% (−0.7%, 4%)	0.169
SqueezeNet	10	73.6%	72.7%	0.9% (−1.5%, 3.2%)	0.466
SqueezeNet	30	71.3%	72.8%	−1.6% (−3.9%, 0.8%)	0.195
SqueezeNet	50	71.9%	72.9%	−1% (−3.4%, 1.3%)	0.390
SqueezeNet	100	73.6%	73.9%	−0.3% (−2.7%, 2.1%)	0.803
SqueezeNet	200	73.1%	74.9%	−1.7% (−4.1%, 0.6%)	0.148
SqueezeNet	500	73.5%	72.2%	1.4% (−1%, 3.7%)	0.254
SqueezeNet	1000	70.7%	70.3%	0.4% (−2%, 2.7%)	0.757

Table 7:

Comparison of accuracy between filetypes by ML model and subspecialty, internal test set. Includes all quantities of images in training sets.

Tissue	ML Model	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Breast	Resnet50	94%	94.3%	−0.3% (−4.2%, 3.6%)	0.888
Breast	SqueezeNet	90.7%	89.7%	1.1% (−2.8%, 5%)	0.584
Colon	Resnet50	83.1%	84.3%	−1.2% (−5.1%, 2.7%)	0.540
Colon	SqueezeNet	91%	89.1%	1.9% (−2%, 5.7%)	0.346
Prostate	Resnet50	96%	96.8%	−0.9% (−4.7%, 3%)	0.666
Prostate	SqueezeNet	96.9%	98.8%	−2% (−5.8%, 1.9%)	0.322

Table 8:

Comparison of accuracy between filetypes by ML model and subspecialty, external test set. Includes all quantities of images in training sets.

Tissue	ML Model	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Breast	Resnet50	75.7%	75.8%	−0.1% (−1.6%, 1.5%)	0.904
Breast	SqueezeNet	68.6%	68.4%	0.2% (−1.4%, 1.7%)	0.809
Colon	Resnet50	82.3%	80.6%	1.8% (0.2%, 3.3%)	0.025 *
Colon	SqueezeNet	69.6%	70.3%	−0.7% (−2.2%, 0.8%)	0.374
Prostate	Resnet50	78.1%	76.7%	1.4% (−0.2%, 2.9%)	0.078 .
Prostate	SqueezeNet	79.3%	79.7%	−0.4% (−1.9%, 1.2%)	0.655

Table 9:

Comparison of accuracy between filetypes by ML model, number of images, and subspecialty, internal test set.

Tissue	ML Model	Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Breast	Resnet50	10	84.8%	81.1%	3.8% (−6.5%, 14%)	0.469
Breast	Resnet50	30	89.9%	98.5%	−8.6% (−18.8%, 1.7%)	0.101
Breast	Resnet50	50	95.9%	90.2%	5.7% (−4.6%, 16%)	0.276
Breast	Resnet50	100	92.6%	94.9%	−2.3% (−12.6%, 8%)	0.660
Breast	Resnet50	200	98%	98.6%	−0.7% (−10.9%, 9.6%)	0.896
Breast	Resnet50	500	98.7%	98.3%	0.4% (−9.8%, 10.7%)	0.937
Breast	Resnet50	1000	98.1%	98.4%	−0.3% (−10.5%, 10%)	0.957
Breast	SqueezeNet	10	90.9%	66.7%	24.2% (14%, 34.5%)	<0.001 ***
Breast	SqueezeNet	30	81.8%	93.6%	−11.8% (−22.1%, −1.6%)	0.024 *
Breast	SqueezeNet	50	86.6%	91.7%	−5.1% (−15.4%, 5.1%)	0.327
Breast	SqueezeNet	100	90.8%	88.3%	2.5% (−7.8%, 12.7%)	0.635
Breast	SqueezeNet	200	93.7%	94.5%	−0.8% (−11.1%, 9.4%)	0.872
Breast	SqueezeNet	500	95.3%	97.2%	−1.9% (−12.2%, 8.3%)	0.715
Breast	SqueezeNet	1000	96.2%	95.6%	0.6% (−9.7%, 10.8%)	0.915
Colon	Resnet50	10	18.2%	26.5%	−8.3% (−18.6%, 1.9%)	0.111
Colon	Resnet50	30	82.5%	86%	−3.6% (−13.8%, 6.7%)	0.494
Colon	Resnet50	50	92.3%	90.7%	1.7% (−8.6%, 11.9%)	0.749
Colon	Resnet50	100	95%	94.8%	0.2% (−10%, 10.5%)	0.967
Colon	Resnet50	200	97.1%	95.2%	1.9% (−8.4%, 12.2%)	0.717
Colon	Resnet50	500	97.8%	97.9%	−0.1% (−10.3%, 10.2%)	0.990
Colon	Resnet50	1000	98.6%	98.8%	−0.3% (−10.6%, 10%)	0.955
Colon	SqueezeNet	10	72.7%	59.1%	13.6% (3.4%, 23.9%)	0.009 **
Colon	SqueezeNet	30	84.8%	84.6%	0.1% (−10.1%, 10.4%)	0.980
Colon	SqueezeNet	50	94.4%	95.9%	−1.4% (−11.7%, 8.8%)	0.782
Colon	SqueezeNet	100	97.1%	96.2%	0.8% (−9.4%, 11.1%)	0.871
Colon	SqueezeNet	200	95.6%	95.1%	0.5% (−9.7%, 10.8%)	0.920
Colon	SqueezeNet	500	95.7%	97%	−1.3% (−11.5%, 9%)	0.806
Colon	SqueezeNet	1000	96.6%	96%	0.6% (−9.6%, 10.9%)	0.907
Prostate	Resnet50	10	75.8%	81.8%	−6.1% (−16.3%, 4.2%)	0.246
Prostate	Resnet50	30	100%	100%	0% (−10.3%, 10.3%)	1.000
Prostate	Resnet50	50	100%	100%	0% (−10.3%, 10.3%)	1.000
Prostate	Resnet50	100	100%	100%	0% (−10.3%, 10.3%)	1.000
Prostate	Resnet50	200	97.9%	97.7%	0.1% (−10.1%, 10.4%)	0.977
Prostate	Resnet50	500	99.3%	99%	0.4% (−9.9%, 10.6%)	0.940
Prostate	Resnet50	1000	98.8%	99.3%	−0.5% (−10.7%, 9.8%)	0.930
Prostate	SqueezeNet	10	89.4%	95.5%	−6.1% (−16.3%, 4.2%)	0.246
Prostate	SqueezeNet	30	98.9%	100%	−1.1% (−11.4%, 9.1%)	0.828
Prostate	SqueezeNet	50	96.7%	100%	−3.3% (−13.6%, 6.9%)	0.524

Tissue	ML Model	Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Prostate	SqueezeNet	100	99.5%	100%	−0.5% (−10.8%, 9.8%)	0.923
Prostate	SqueezeNet	200	96.3%	98.2%	−1.9% (−12.1%, 8.4%)	0.718
Prostate	SqueezeNet	500	98.6%	99.1%	−0.5% (−10.7%, 9.8%)	0.926
Prostate	SqueezeNet	1000	98.8%	99.1%	−0.3% (−10.5%, 10%)	0.955

Table 10:

Comparison of accuracy between filetypes by ML model, number of images, and subspecialty, external test set.

Tissue	ML Model	Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Breast	Resnet50	10	64.7%	66.3%	−1.6% (−5.6%, 2.5%)	0.457
Breast	Resnet50	30	76.4%	77.1%	−0.7% (−4.8%, 3.4%)	0.750
Breast	Resnet50	50	76.5%	78.1%	−1.6% (−5.6%, 2.5%)	0.457
Breast	Resnet50	100	78.3%	77.9%	0.3% (−3.8%, 4.4%)	0.873
Breast	Resnet50	200	78.6%	76.4%	2.2% (−1.9%, 6.3%)	0.288
Breast	Resnet50	500	77.4%	76.5%	0.9% (−3.2%, 5%)	0.671
Breast	Resnet50	1000	77.8%	78.2%	−0.3% (−4.4%, 3.8%)	0.873
Breast	SqueezeNet	10	61.8%	59.1%	2.7% (−1.4%, 6.8%)	0.202
Breast	SqueezeNet	30	64.5%	67.2%	−2.7% (−6.8%, 1.4%)	0.202
Breast	SqueezeNet	50	65%	65.6%	−0.7% (−4.8%, 3.4%)	0.751
Breast	SqueezeNet	100	72%	70.5%	1.4% (−2.7%, 5.5%)	0.490
Breast	SqueezeNet	200	75.3%	74.2%	1.1% (−3%, 5.2%)	0.595
Breast	SqueezeNet	500	72.7%	72.4%	0.3% (−3.8%, 4.4%)	0.873
Breast	SqueezeNet	1000	69.1%	70%	−0.9% (−5%, 3.2%)	0.671
Colon	Resnet50	10	59.8%	63.5%	−3.6% (−7.7%, 0.5%)	0.082 .
Colon	Resnet50	30	70.4%	71.1%	−0.7% (−4.8%, 3.4%)	0.727
Colon	Resnet50	50	76.4%	67.6%	8.7% (4.6%, 12.8%)	<0.001 ***
Colon	Resnet50	100	88%	82.7%	5.3% (1.2%, 9.4%)	0.012 *
Colon	Resnet50	200	96.9%	98%	−1.1% (−5.2%, 3%)	0.601
Colon	Resnet50	500	92.4%	92.5%	−0.2% (−4.3%, 3.9%)	0.931
Colon	Resnet50	1000	92.5%	88.5%	4% (−0.1%, 8.1%)	0.055 .
Colon	SqueezeNet	10	82.5%	80.4%	2.2% (−1.9%, 6.3%)	0.296
Colon	SqueezeNet	30	70.2%	72.7%	−2.5% (−6.6%, 1.5%)	0.223
Colon	SqueezeNet	50	70.4%	71.6%	−1.3% (−5.4%, 2.8%)	0.542
Colon	SqueezeNet	100	66.5%	69.3%	−2.7% (−6.8%, 1.4%)	0.191
Colon	SqueezeNet	200	66.5%	70.5%	−4% (−8.1%, 0.1%)	0.055 .
Colon	SqueezeNet	500	67.6%	65.3%	2.4% (−1.7%, 6.5%)	0.257
Colon	SqueezeNet	1000	63.6%	62.5%	1.1% (−3%, 5.2%)	0.601
Prostate	Resnet50	10	70.9%	66%	4.9% (0.8%, 9%)	0.018 *
Prostate	Resnet50	30	78.3%	77.3%	1% (−3.1%, 5.1%)	0.618
Prostate	Resnet50	50	78.2%	72.1%	6.1% (2%, 10.2%)	0.004 **
Prostate	Resnet50	100	80.1%	81.9%	−1.8% (−5.9%, 2.3%)	0.383
Prostate	Resnet50	200	78.7%	80%	−1.3% (−5.4%, 2.8%)	0.533
Prostate	Resnet50	500	80%	80.5%	−0.5% (−4.6%, 3.6%)	0.804
Prostate	Resnet50	1000	80.1%	78.8%	1.3% (−2.8%, 5.4%)	0.533
Prostate	SqueezeNet	10	76.5%	78.7%	−2.2% (−6.3%, 1.9%)	0.290
Prostate	SqueezeNet	30	79.1%	78.6%	0.5% (−3.6%, 4.6%)	0.803

Tissue	ML Model	Image Count	Mean with JPG	Mean with PNG	Difference in Means (95% CI)	P-Value
Prostate	SqueezeNet	50	80.4%	81.6%	−1.2% (−5.3%, 2.9%)	0.575
Prostate	SqueezeNet	100	82.2%	81.8%	0.4% (−3.7%, 4.5%)	0.854
Prostate	SqueezeNet	200	77.5%	79.9%	−2.3% (−6.4%, 1.8%)	0.263
Prostate	SqueezeNet	500	80.3%	78.8%	1.4% (−2.7%, 5.5%)	0.493
Prostate	SqueezeNet	1000	79.4%	78.4%	0.9% (−3.2%, 5%)	0.661