Original Article

Effects of Image Quantity and Image Source Variation on Machine Learning Histology Differential Diagnosis Models

Elham Vali-Betts[1], Kevin J. Krause[1], Alanna Dubrovsky[2], Kristin Olson[1], John Paul Graff[1], Anupam Mitra[1], Ananya Datta-Mitra[1], Kenneth Beck[1], Aristotelis Tsirigos[3], Cynthia Loomis[3], Antonio Galvao Neto[4], Esther Adler[3], Hooman H. Rashidi[1]

[1]Department of Pathology and Laboratory Medicine, University of California Davis School of Medicine, Sacramento, CA, USA, [2]Department of Psychiatry, Oregon Health and Science University, Portland, OR, USA, [3]Department of Psychiatry, School of Medicine, New York University, New York, NY, USA, [4]Department of Pathology, University of Colorado, Boulder, CO, USA

Submitted: 30-Aug-2020 Revised: 28-Sep-2020 Accepted: 28-Oct-2020 Published: 23-Jan-2021

Abstract

Aims: Histology, the microscopic study of normal tissues, is a crucial element of most medical curricula. Learning tools focused on histology are very important to learners who seek diagnostic competency within this important diagnostic arena. Recent developments in machine learning (ML) suggest that certain ML tools may be able to benefit this histology learning platform. Here, we aim to explore how one such tool based on a convolutional neural network, can be used to build a generalizable multi-classification model capable of classifying microscopic images of human tissue samples with the ultimate goal of providing a differential diagnosis (a list of look-alikes) for each entity. **Methods:** We obtained three institutional training datasets and one generalizability test dataset, each containing images of histologic tissues in 38 categories. Models were trained on data from single institutions, low quantity combinations of multiple institutions, and high quantity combinations of multiple institutions. Models were tested against withheld validation data, external institutional data, and generalizability test images obtained from Google image search. Performance was measured with macro and micro accuracy, sensitivity, specificity, and f1-score. **Results:** In this study, we were able to show that such a model's generalizability is dependent on both the training data source variety and the total number of training images used. Models which were trained on 760 images from only a single institution performed well on withheld internal data but poorly on external data (lower generalizability). Increasing data source diversity improved generalizability, even when decreasing data quantity: models trained on 684 images, but from three sources improved generalization accuracy between 4.05% and 18.59%. Maintaining this diversity and increasing the quantity of training images to 2280 further improved generalization accuracy between 16.51% and 32.79%. **Conclusions:** This pilot study highlights the significance of data diversity within such studies. As expected, optimal models are those that incorporate both diversity and quantity into their platforms.

Keywords: Convolutional neural network, differential diagnosis, generalization, histology, histopathology, image source variation, machine learning, multi-classification

Introduction

Histology is the foundation of microscopic tissue evaluation and pathology diagnoses.[1,2] This cornerstone of medicine is an integral part of medical school curricula and serves as a pillar for pathology education.[2] Understanding the normal histologic architecture is key in building a microscopy-based diagnostic competency, and subtle variations in tissue morphology are challenging to master for new learners. Unfortunately, teaching histology may require resources that are not always available in developing or underserved areas.

Many research groups are exploring new approaches to help make learning histology less challenging and more

Address for correspondence: Dr. Hooman H. Rashidi, 4400 V Street, Sacramento, CA 95817, USA.
E-mail: hrashidi@ucdavis.edu
Kevin J. Krause,
B.S. Degree, Biomedical Engineering,
4400 V Street, Sacramento, CA 95817, USA.
E-mail: kjkrause@ucdavis.edu

How to cite this article: Vali-Betts E, Krause KJ, Dubrovsky A, Olson K, Graff JP, Mitra A, *et al*. Effects of image quantity and image source variation on machine learning histology differential diagnosis models. J Pathol Inform 2021;12:5.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2021/12/1/5/307703

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_69_20

entertaining, including the University of New Jersey Medical School's use of an "audiovisual switching and projection system" to streamline the presentation of histology images in lectures;[3] the University of Granada's efforts to analyze factors impacting the motivation of various students to learn histology;[4] and Newcastle University's analysis of factors influencing the effectiveness of histology-oriented e-learning.[5]

Over the last decade, advancements in the field of information science and digital microscopy have started to reform the histology learning platform[6-9] and other medical disciplines.[10] However, these improvements may bring challenging new requirements, such as reliable internet access; authentic source-information; and easy accessibility. Hence, more advanced tools may be warranted to support the histology learning environment.

Fortunately, advancements in computational analysis, specifically machine learning (ML) and artificial intelligence (AI),[11] have recently enhanced the histopathology arena.[12-16] These advances are mostly credited to deep learning techniques using convolutional neural networks (CNNs) in various image analysis studies.[17-20]

Niazi *et al*. have shown that CNNs can be used to accurately assess the depth of bladder tumor penetration into the lamina propria, an important metric for treating and monitoring the progression of the disease.[19] Further, Coudray *et al*. used CNNs to predict adenocarcinoma and squamous cell carcinoma from normal lung tissue samples with an AUC of 0.98, matching the diagnostic performance of a trained pathologist.[20]

In this study, we explored the application of CNNs to the histologic learning platform, aiming to create an app capable of distinguishing tissue subtypes and recognizing their look-alikes. In addition, we studied the relationships between the number of images used for training, the number of different image sources used, and the ultimate generalizability of the resulting models. Ultimately, we identified the best performing model, based on generalizability, and deployed it to our histology ML app. Our app is now able to analyze an image of a histologic entity (tissue), able to identify it, and ultimately generate a differential diagnosis (list of look-alikes) [Figure 1].

Methods

Two institutional datasets were provided by the University of California, Davis (UCD) and New York University (NYU). Institutional Review Board (IRB) approval was obtained at the UCD (IRB ID: 1286225-1) and NYU (no IRB required) for the anonymized normal histology images used in this study. A third set of images was also obtained using several digital whole slide images from various public domain sites, hereafter referred to as external data (EXT). Histologic images in 38 categories of equal proportion [Figure 2] were obtained from each data source (UCD, NYU, EXT). In each category, 10 low power magnification (×4) and 10 high power magnification (×10) images were obtained yielding 20 images per category and a total of 760 images from each data source. We included both square and rectangular images, ranging from 100 to 1600 pixels wide and 100–900 pixels high. These images were collected in portable network graphics (PNG) format and then reviewed and verified by two board certified pathologists.

The above images were then used to create training and validation testing datasets for our ML studies. Eighty percent of each dataset were randomly selected to train a model, and the remaining 20% was withheld for internal validation testing. We also randomly resampled, retrained, and retested each of the datasets mentioned above 10 times to achieve a 10 k-fold cross-validation for the training-testing approach. Each model was trained through a transfer learning approach on the ResNet-50 CNN within Apple's Turicreate open source library. The Turicreate image classifier function performed automatic feature rescaling to resize our images to 224 pixels wide by 224 pixels high, per ResNet-50's input layer specifications.[21,22] We

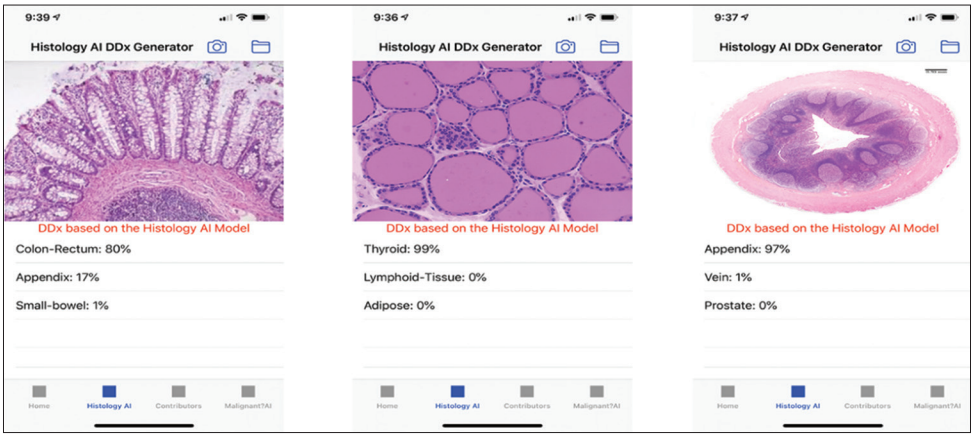


Figure 1: The above representative images are based on our best performing histology machine learning model that includes a combination of all sources and combines all images in each category. The top *n* (highest probability for the top 3 look a likes) are generated by this iOS app which highlights how such a histology differential diagnosis app can be used in practice

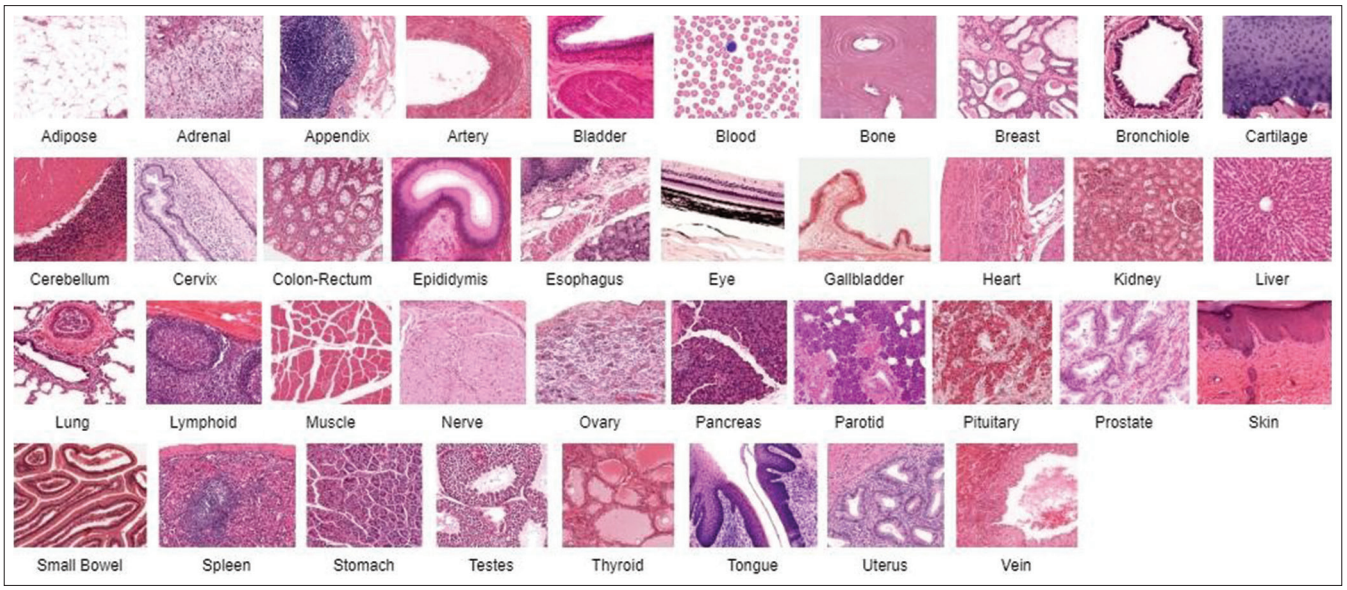


Figure 2: This figure alphabetically lists the 38 classes of histological tissue types used in this study

used the image classifier's default hyperparameters, as shown in the Turicreate documentation,[21] except for the maximum iterations parameter, which we set to 1000 iterations.

In addition to the above initial validation testing, we also performed an external validation step which tested each of the models generated from each data-source against the other data-sources' images. The external validation tests are depicted in Figure 3. For all validation tests we evaluated the top-*n* metrics (the top-*n* values of 1, 3, and 5) by selecting the 'n' highest probability score (s) from each prediction (the target label and it's top 1, top 3, and top 5 look-alikes) [Figure 3].

Finally, we combined the data from all three sources to explore the impact of data diversity in each model's true generalizability. To test the combination models' true generalizability, a fourth dataset was acquired using Google image search to collect 10 images from each of the above 38 categories from various online public domain sources. Notably, this "Google images" generalization dataset was not used in the training phase of any of the models tested and solely used for generalizability testing.

Two combination datasets were constructed: one with lower data quantity, and one with higher data quantity. To build the low quantity combination training set, 6 images were sampled from each tissue category from each data source, yielding 18 total images per tissue category, which ultimately yielded 684 total training images. Selecting 18 images per category in the combination study gives us the advantage of using fewer total data than in the individual study (684 training images vs. 760), so that we can explore the impact of data diversity without the confounding influence of increased data quantity. To further test the effect of both combined data diversity and data quantity, a high quantity combination study was also generated with the maximum data quantity from all three sources (UCD, NYU and EXT) using 20 images from each category from each data source which led to 60 images per category and ultimately yielded a total of 2280 training images. The "Google images" generalization dataset (described above) was then used to compare the performance (accuracies) of the low and high quantity combination models. Clopper-Pearson confidence limits were calculated to analyze the reliability of the results.[23] The null accuracy for this balanced multi-classification task was calculated as $\frac{1}{\text{number of classes}}$ to give context to the results.

Results

The null accuracy of these tests was calculated to be $\frac{1}{38}$ or 2.63%.

Individual data sources (noncombined) [for brevity, only top-5 results are shown here. Top-1 and top-3 results can be found in Appendix 2]

Per-label internal validation

For the EXT internal validation, the highest top-*n* of 5 per-label tissue (the top 5 look-alikes/differential diagnosis) sensitivities were adipose (1.00), eye (1.00), and heart (1.00), while the lowest were pituitary (0.96), appendix (0.96), and small-bowel (0.96), which were most frequently misclassified as liver, ovary, and kidney, respectively.

For the NYU internal validation, the highest top-*n* of 5 per-label tissue sensitivities were adipose (1.00), skin (1.00), and epididymis (1.00), while the lowest were adrenal (0.94), artery (0.96), and bronchiole (0.96), which were most frequently misclassified as uterus, adrenal, and breast, respectively.

For the UCD internal validation, the highest top-*n* of 5 per-label tissue sensitivities were kidney (1.00), lung (1.00), and adrenal (1.00), while the lowest were vein (0.90), appendix (0.94), and artery (0.95), which were most frequently misclassified as adipose, cervix, and appendix, respectively [Figure 4].

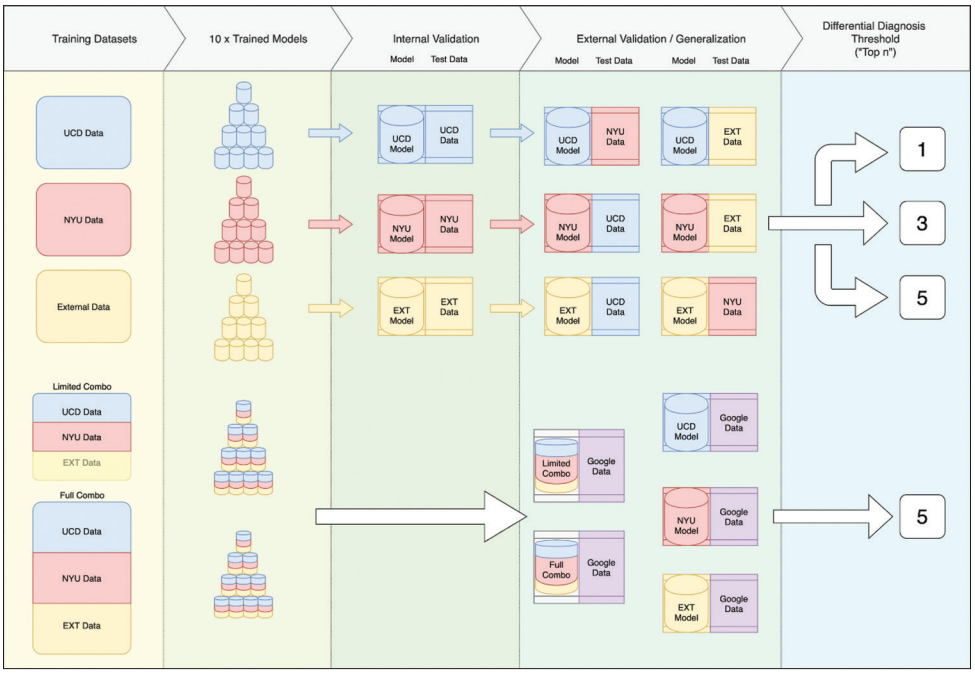


Figure 3: This chart depicts the overall study design. First, each of the three datasets are individually used to create the training sets. Second, each model is tested internally against the aforementioned withheld randomly selected test set to assess the models' internal validation accuracy with a 10 k-fold random sampling cross validation approach. Third, each model is tested externally against both of the other datasets to assess each model's performance, and the results are averaged across the ten models (another 10 k-fold cross validation). Then, each test is repeated with a "top *n*" correct criteria of one, three, and five which represents how each model performs in identifying the top 1, 3 or 5 differential diagnosis (top look-alikes) within each histologic category. Additionally, two combined datasets are generated from the three individual data sources (University of California, Davis, New York University, external data), one with restricted data quantity, and one with full data quantity. Once again, these datasets are resampled to train combination models along with 10 k-fold cross validation. Finally, all of the models, including both combination sets and all three individual datasets, were tested against a generalization test set (google images) obtained from online public domain images

Per label external validation

EXT was externally validated against NYU and UCD. For the EXT versus NYU test, the highest top-*n* of 5 per-label tissue sensitivities were adipose (1.00), thyroid (1.00), and bladder (0.98), while the lowest were blood (0.00), vein (0.02), and artery (0.04), which were most frequently misclassified as spleen, prostate, and nerve, respectively.

For the EXT vs UCD test, the highest top-*n* of 5 per-label tissue sensitivities were adipose (1.00), blood (1.00), and cerebellum (1.00), while the lowest were vein (0.08), lymphoid-tissue (0.08), and appendix (0.10), which were most frequently misclassified as liver, stomach, and stomach, respectively [Figure 4].

UCD was externally validated against EXT and NYU. For the UCD vs NYU test, the highest top-*n* of 5 per-label tissue sensitivities were adipose (1.00), thyroid (1.00), and spleen (1.00), while the lowest were blood (0.00), bronchiole (0.10), and vein (0.12), which were most frequently misclassified as vein, adipose, and esophagus, respectively.

For the UCD versus EXT test, the highest top-*n* of 5 per-label tissue sensitivities were adipose (1.00), blood (1.00), and heart (1.00), while the lowest were vein (0.08), tongue (0.16), and small-bowel (0.22), which were most frequently misclassified as eye, adipose, and stomach, respectively.

NYU was externally validated against EXT and UCD. For the NYU versus UCD test, the highest top-*n* of 5 per-label tissue sensitivities were bronchiole (1.00), bone (1.00), and muscle (1.00), while the lowest were prostate (0.02), liver (0.08), and cervix (0.12), which were most frequently misclassified as epididymis, pituitary, and artery, respectively.

For the NYU versus EXT test, the highest top-*n* of 5 per-label tissue sensitivities were bone (1.00), pituitary (1.00), and muscle (1.00), while the lowest were liver (0.00), ovary (0.16), and cervix (0.16), which were most frequently misclassified as pancreas, tongue, and pituitary, respectively.

Figure 5 shows the ranked (high to low) class sensitivities averaged across every top-5 external validation test. The highest sensitivity is observed for adipose (0.99), thyroid (0.95), and eye (0.89). Conversely, the lowest sensitivity is observed for vein (0.17), prostate (0.26), and artery (0.32), which were most frequently misidentified as eye, epididymis, and esophagus, respectively.

Figure 6 summarizes the internal and external per-label validation tests.

Cumulative internal validation on withheld 20%

The internal validation results were relatively the same for each data source: the top-*n* of 5 cumulative accuracy,

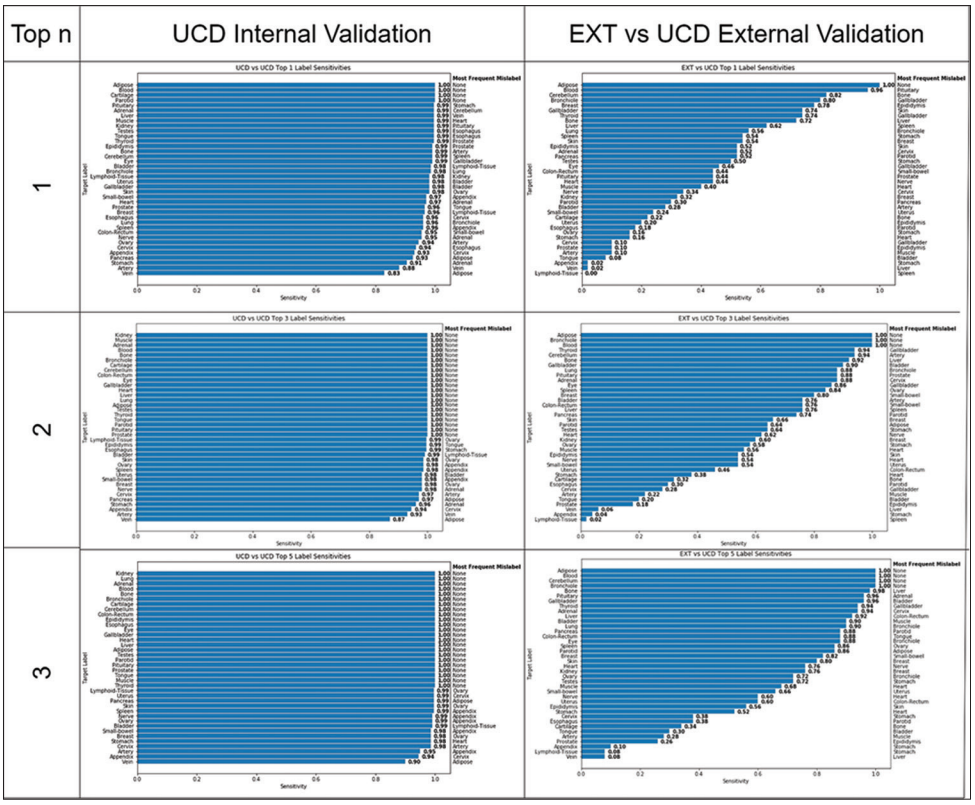


Figure 4: This image depicts the sensitivity graphs for each label in a given test and top-*n* value (top 1, 3, or 5 differential diagnosis predictions). In addition, the outside column indicates the most frequent incorrect label for a given target class. Note that the internal validation results appear similar amongst the different categories while the true discriminators are the model's external validation performances

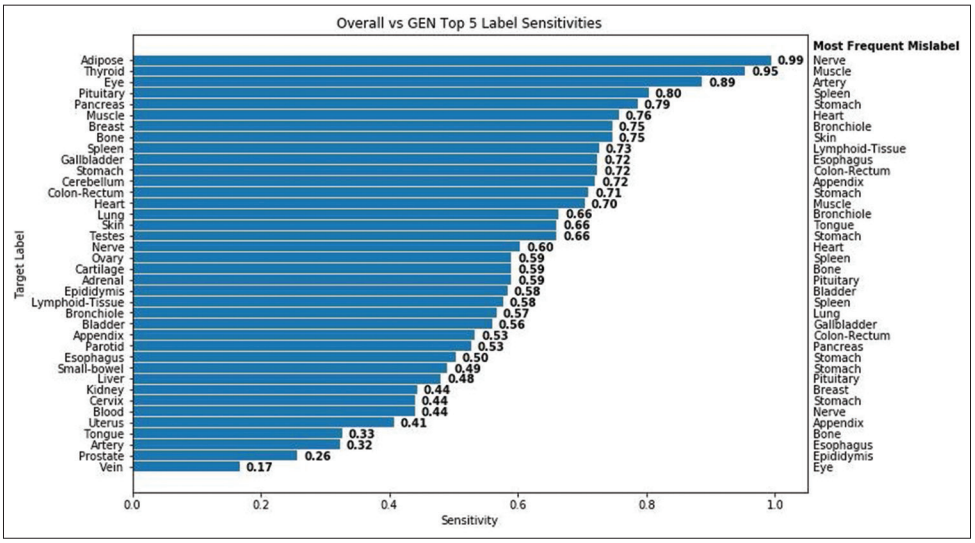


Figure 5: This chart depicts the average sensitivities across every external validation test, showing trends in the overall performance of each label (far left side, "Target Label"). Highest sensitivities were noted in adipose, thyroid and eye while the lowest sensitivities were noted in vein, prostate and artery. The most frequent mislabel for each entity is also listed on the far right side of each histologic entity

cumulative sensitivity, cumulative positive predictive value, cumulative sensitivity, and cumulative f1-score were all 0.99. For each top-*n* of 3 global metric UCD and EXT both scored 0.99, while NYU scored 0.98. For the top-*n* of 1 global metrics UCD and EXT both scored 0.97, while NYU scored 0.95 [Figure 7].

Cumulative external validation (generalization results)

Figure 7 shows the results of the external validation tests, for top-*n* of 1, 3, and 5. For top-*n* of 5, the EXT versus UCD was the highest performing test. This test showed accuracy of 0.69, F1-score of 0.66, and sensitivity of 0.69. The remaining tests can be found in Figure 7.

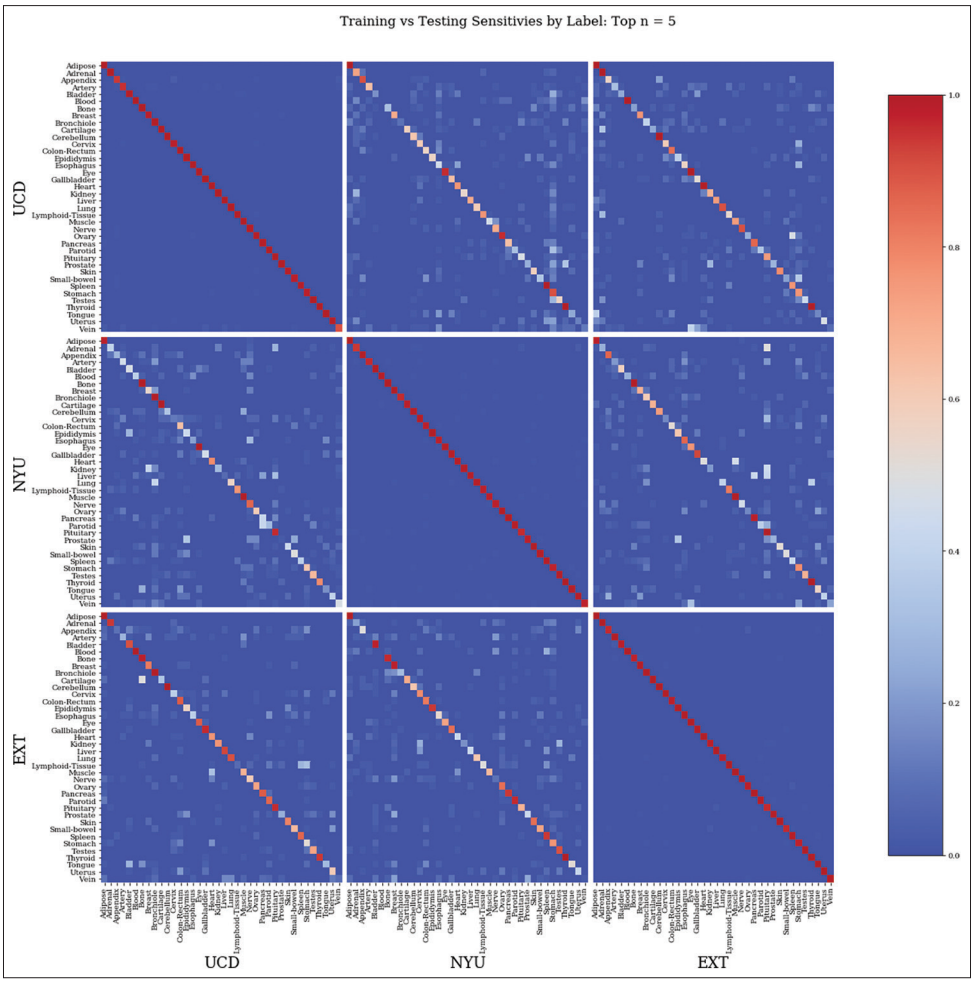


Figure 6: This chart depicts the correlation for each label and each validation pairing. The outer left Y axis is the training dataset and the outer bottom X axis is the testing dataset. True positives appear along the diagonal of each chart, and false positives appear outside of the diagonal. The strongest correlations are depicted in red (as expected each individual entity from a given training source when tested against its own individual entity's testing source (e.g., University of California, Davis Adipose tested against University of California, Davis adipose) will show the highest correlation (i.e., depicted as red). Thus, a stronger collection of positives along the diagonal indicates higher sensitivity (depicted as red) while the lower correlation for each entity will be less red (highest correlation = red, lowest correlation = blue). Additionally, the large number of light blue dots present off the diagonal are indicative of each individual entity's mislabeled correlate with their respective look-alike histologic entity

Combination model generalizability

The individual data sources, UCD, NYU, and EXT, accurately classified 58.77%, 51.57%, and 58.40% of public domain Google images, respectively. The per-label results from these tests are provided in appendix 1. The low quantity combination dataset of 684 images accurately classified 61.16% of images, achieving a 4.05% improvement over UCD, 18.59% improvement over NYU, and 4.73% improvement over EXT. The high quantity combination dataset of 2280 images accurately classified 68.48% of public domain images, achieving a 16.51% improvement over UCD, 51.57% improvement over NYU, and 17.27% improvement over EXT [Table 1].

Discussion

Our combination analysis demonstrated that training with a more diverse dataset could outperform a less diverse dataset in a generalization test, even when the more diverse dataset had fewer total images. Furthermore, we demonstrated that a dataset which is more diverse and has higher quantity could outperform both datasets: high diversity with low quantity, and low diversity with low quantity. Most importantly, in addition to having increased quantity, these results highlight the importance of data diversity in training a generalizable ML model. Further, the results of our tests are high relative to the null accuracy of a naïve 38-class multiclassifier, though improvements should be explored in future studies.

Our analysis also showed a positive association between the performance (accuracy, sensitivity) and the level of top-*n* differential diagnosis being used. This suggests that the differential diagnoses are picking up on architectural similarities in tissues. This feature is useful for teaching new histology learners to recognize similarities and common look-alikes among different tissues. This look-alike clustering

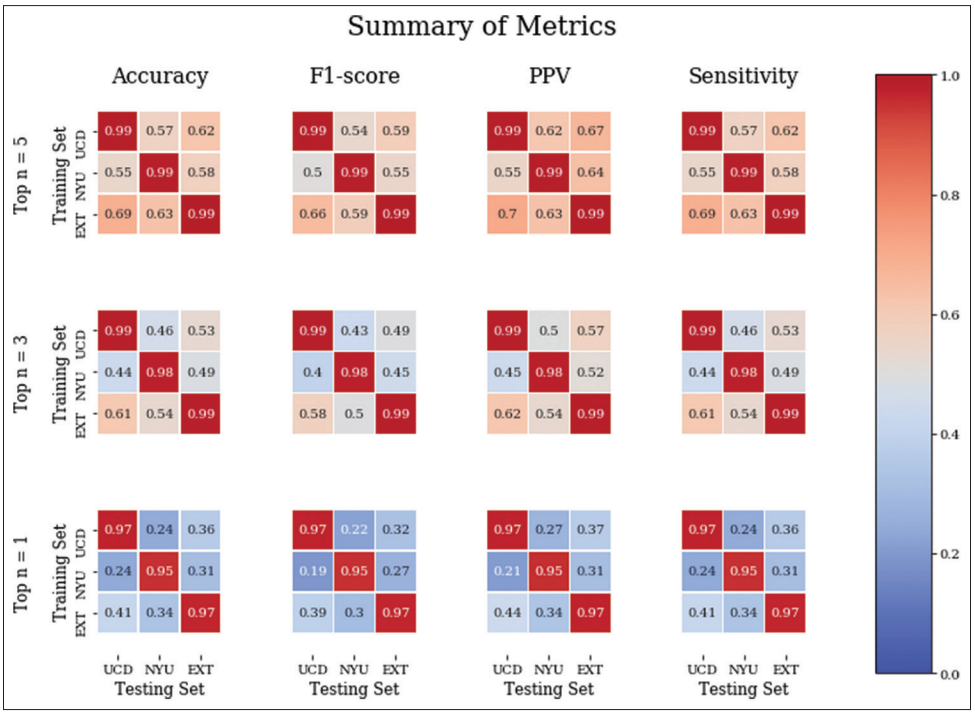


Figure 7: Correlation of the accuracy, f1 score, positive predictive value, and sensitivity of internal and external test results for each top-*n* correct value are shown. As expected the $n = 5$ (top 5 look-alikes) has the best performance parameters (compared to $n = 1$ or $n = 3$)

Table 1: Generalization accuracy comparisons (single data source vs. combined sources)

Data source	Single image source (UCD or NYU or EXT) Accuracy: 760 images	Combined image source (UCD + NYU + EXT) Accuracy: 684 images	Percentage improvement of combined image source
UCD	0.5877 (0.5693-0.6330)	0.6116 (0.5649-0.6584)	+4.05% (−0.78-4.00)
NYU	0.5157 (0.4899-0.5549)	0.6116 (0.5649-0.6584)	+18.59% (15.31-18.66)
EXT	0.5840 (0.5413-0.6057)	0.6116 (0.5649-0.6584)	+4.73% (4.35-8.70)
Data source	**Single image source (UCD or NYU or EXT) Accuracy: 760 images**	**Combined image source (UCD + NYU + EXT) Accuracy: 2280 images**	**Percentage improvement of combined image source**
UCD	0.5877 (0.5693-0.6330)	0.6848 (0.6554-0.7123)	+16.51% (12.53-15.11
NYU	0.5157 (0.4899-0.5549)	0.6848 (0.6554-0.7123)	+32.79% (28.38-33.78)
EXT	0.5840 (0.5413-0.6057)	0.6848 (0.6554-0.7123)	+17.27% (17.61-21.07)

"Single accuracy" depicts the mean accuracy (with 95% CI) of a single data source (e.g., UCD). "Combined accuracy" depicts the corresponding mean accuracy and interval of the respective combination dataset (684 or 2280). Percent improvement indicates by what percentage accuracy was improved by the combination dataset, over the individual data source. The single source models were generated on datasets that contained 760 images while the combined dataset noted above (UCD + NYU + EXT) includes 684 Images (18 images/category). The full quantity combined models contained - 2280 images (60 images/category). UCD: University of California Davis; NYU: New York University; EXT: External dataset; CI: Confidence interval

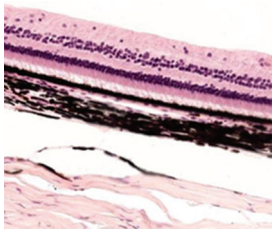
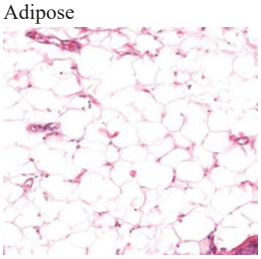
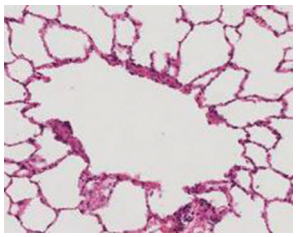
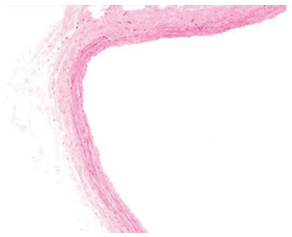
may be an appropriate complement to other histology learning modalities – lectures, textbooks, videos, etc.

In addition, our combination study tested models against images obtained from online Google public domain images, which ultimately were the most difficult to classify across every dataset. Reviewing these images showed that they are highly irregular, inconsistent, and often contaminated with text and graphics. Because the models were trained on clean images, they may struggle to classify the less polished images in the Google search dataset. A study by Jones *et al*., demonstrated that JPEG images and PNG images can be used to train similarly accurate ML models.[14] However, because these ML models were trained on relatively "lossless" PNG images, they may struggle to classify the comparatively "lossy" JPEG images in the Google search dataset.[24,25] Future studies may be useful to explore employing the less polished data and a variety of image file formats into the training data.

In our study, the highest performance predictions were on adipose and thyroid tissue types. The simplicity of their architectures, and the lack of other background tissues, compared to other tissue images, may make these tissue types easy to distinguish. Despite adipose tissue's high accuracy, it was occasionally misidentified as bronchiole tissue. Adipose-bronchiole confusion may be caused by the presence of lung tissue in the background of bronchiole, which resembles adipose tissue [Table 2].

One frequently misidentified tissue was artery, which was most misidentified as nerve. This could be explained by the circular cross section of the nerve with neural fibers appearing like elements in the arteries such as red blood cells. Table 3 illustrates the similarities between arterial and nervous tissues across institutions. Moreover, the striking similarities between arterial and neural tissue, and the incidences of confusion with one another, are evidence that the model is learning tissue architectures to a level where it can make intelligent mistakes, or mistakes that a human would be likely to encounter. Incorporating more examples of these tissues into training may prove beneficial in distinguishing them from one another [Table 3].

Table 2: Comparison of selected tissue types

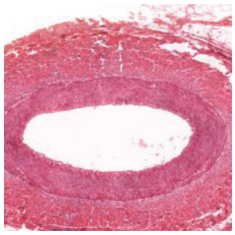
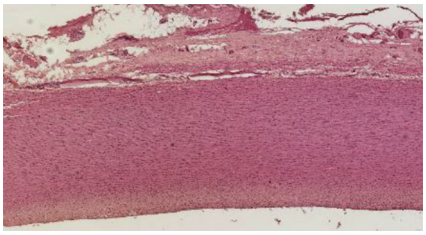
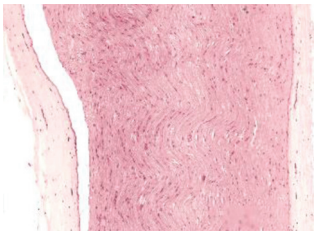
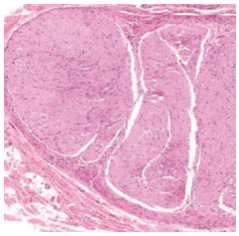
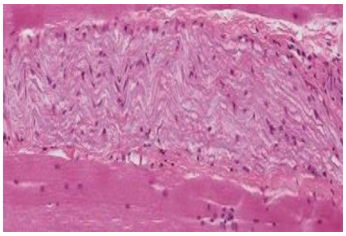
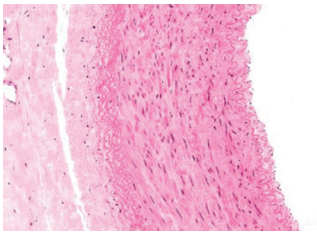
Eye	Adipose
	
Bronchiole	**Vein**
	

The differences and similarities between four tissue types with some overlapping features (e.g., noted similarities between adipose and bronchiole or similarities between the eye and vein histology)

Incorporating multiple data sources may also be beneficial for improving model flexibility. In our study, we found that the UCD and EXT datasets used a blood-smear technique,[26] while the NYU dataset used a cross-sectional technique to gather blood images. Not surprisingly, UCD and EXT struggled to classify blood images from NYU, and vice versa. Interestingly, both combination studies showed improvement in classifying blood images, suggesting that incorporating both techniques improved model flexibility and generalizability [Table 4].

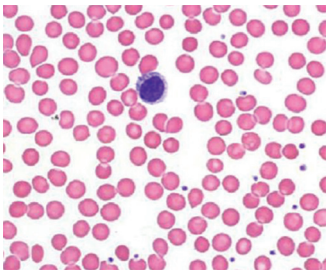
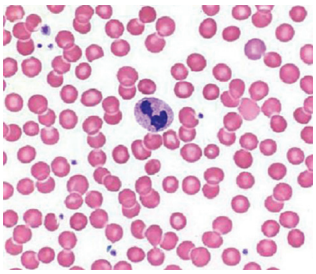
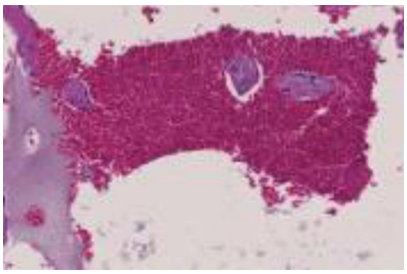
A limitation of our study is the relatively small number of images available (760 images per dataset) compared to traditional CNNs, which include thousands to millions of images.[27] In order to compensate for the small data size, this study employed a transfer learning technique. In this technique, a large CNN is pretrained on millions of images. Next, the model's layers are frozen, and a small number of new layers are added. Finally, the new model is trained on a smaller dataset, only adjusting the new layers. This technique can produce highly generalizable, large CNNs, with relatively small training sets.[12,13,27,28] Many examples of this strategy exist in various CNN classification tasks in which low quantity data are a challenge.[29,30] This study utilizes the ResNet-50 transfer learning architecture,[14,31] though many other architectures exist, such as AlexNet, VGG, Inception, and DenseNet.[14,27] Since some studies suggest that Inception V3 may slightly outperform ResNet-50 for some classification tasks,[32] it may

Table 3: Comparison of various artery and nerve tissues

Artery		
UCD	**NYU**	**EXT**
		
Nerve		
UCD	**NYU**	**EXT**
		

Nerve and arterial tissues bear striking resemblances, which may explain their classification confusion. Further, the confusion between these tissues shows evidence that the learning algorithm is intelligent enough to formulate smart, or insightful, mistakes. UCD: University of California Davis; NYU: New York University; EXT: External dataset

Table 4: Comparison of blood slides between training datasets

UCD blood	EXT	NYU blood
		

UCD (left) and external (center) utilize a smearing technique for blood slides, while New York University (right) utilizes a vessel cross-sectional technique. These differences may account for errors in blood slide identification between data sources. UCD: University of California Davis; NYU: New York University; EXT: External dataset

be worthwhile for a future study to repeat this on the Inception V3 transfer learning architecture.

Overall, this study has illuminated the pathway toward a fully functional histopathology AI learning tool. Moreover, this study has yielded some valuable insights which will aid our understanding of histological multi-classification tasks, though future larger studies are required to support our findings and further enhance our understanding within this exciting new field.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

References

1. Bloom W, Fawcett DW. A Textbook of Histology. 10th ed. Philadelphia, PA: Saunders; 1975.
2. Garcia M, Victory N, Navarro-Sempere A, Segovia Y. Students' views on difficulties in learning histology. Anat Sci Educ 2019;12:541-9.
3. Gona AG, Berendsen PB, Alger EA. New approach to teaching histology. J Int Assoc Med Sci Educ 2005;15.
4. Campos-Sanchez A, Lopez-Nunez JA, Scionti G, Garzon I, González-Andrades M, Alaminos M, *et al*. Developing an audiovisual notebook as a self-learning tool in histology: Perceptions of teachers and students. Anat Sci Educ 2014;7:209-18.
5. Backhouse M, Fitzpatrick M, Hutchinson J, Thandi CS, Keenan ID. Improvements in anatomy knowledge when utilizing a novel cyclical "observe-reflect-draw-edit-repeat" learning process. Anat Sci Educ 2017;10:7-22.
6. Helle L, Nivala M, Kronqvist P. More technology, better learning resources, better learning? Lessons from adopting virtual microscopy in undergraduate medical education. Anat Sci Educ 2013;6:73-80.
7. Lee LM, Goldman HM, Hortsch M. The virtual microscopy database-sharing digital microscope images for research and education. Anat Sci Educ 2018;11:510-5.
8. Coleman R. Can histology and pathology be taught without microscopes? The advantages and disadvantages of virtual histology. Acta Histochem 2009;111:1-4.
9. Silva-Lopes VW, Monteiro-Leal LH. Creating a histology-embryology free digital image database using high-end microscopy and computer techniques for on-line biomedical education. Anat Rec B New Anat 2003;273:126-31.
10. Schwamborn K. The importance of histology and pathology in mass spectrometry imaging. Adv Cancer Res 2017;134:1-26.
11. Gupta A, Harrison PJ, Wieslander H, Pielawski N, Kartasalo K, Partel G, *et al*. Deep Learning in image cytometry: A review. Cytometry A 2019;95:366-80.
12. Gibney E. Google AI algorithm masters ancient game of Go. Nature 2016;529:445-6.
13. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: The present landscape of supervised methods. Acad Pathol 2019;6:2374289519873088.
14. Jones AD, Graff JP, Darrow M, Borowsky A, Olson KA, Gandour-Edwards R, *et al*. Impact of pre-analytical variables on deep learning accuracy in histopathology. Histopathology 2019;75:39-53.
15. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J 2018;16:34-42.
16. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, *et al*. Artificial intelligence in pathology. J Pathol Transl Med 2019;53:1-12.
17. Litjens G, Sanchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
18. Li C, Wang X, Liu W, Latecki LJ. DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. Med Image Anal 2018;45:121-33.
19. Niazi MK, Tavolara T, Arole V, Parwani A, Lee C, Gurcan M. MP58-06 automated staging of t1 bladder cancer using digital pathologic H and E images: A deep learning approach. J Urol 2018;199(4S):e775.
20. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, *et al*. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-67.
21. Turi Create API 6.4.1 Documentation Turicreate. Image Classifier Create Apple. Available from: https://www.apple.github.io/turicreate/docs/api/generated/turicreate.image_classifier.create.html. [Last accessed on 2020 Oct 04].
22. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition; 2015. Available from: https://www.arxiv.org/abs/1512.03385. [Last accessed on 2020 Oct 04].
23. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statist Sci 2001;16:101-33.
24. Tan L. Image file formats. Biomed Imaging Interv J 2006;2:e6.
25. Dodge S, Karam L, editors. Understanding how Image Quality affects Deep Neural Networks. 2016 8th International Conference on Quality of Multimedia Experience (QoMEX); 2016.
26. Gulati G, Song J, Florea AD, Gong J. Purpose and criteria for blood smear scan, blood smear examination, and blood smear review. Ann Lab Med 2013;33:1-7.
27. Yamashita R, Nishio M, Do RK, Togashi K. Convolutional neural networks: An overview and application in radiology. Insights Imaging 2018;9:611-29.
28. Oquab M, Bottou L, Laptev I, Sivic J, editors. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern

Recognition; 2014.
29. Paul R, Hawkins SH, Balagurunathan Y, Schabath MB, Gillies RJ, Hall LO, *et al*. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography 2016;2:388-95.
30. Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, Rim TH. Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. PLoS One 2017;12:e0187336.
31. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual deep convolutional neural network predicts MGMT methylation status. J Digit Imaging 2017;30:622-8.
32. Xiao T, Liu L, Li K, Qin W, Yu S, Li Z. Comparison of transferred deep neural networks in ultrasonic breast masses discrimination. Biomed Res Int 2018;2018:4605191.

Appendix 1: Classification per-label accuracy against public domain images

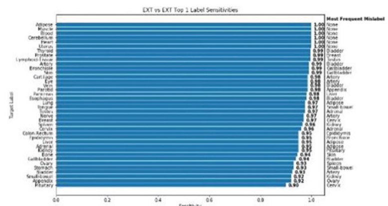
Label	Accuracy (mean [95% CI])				
	Combo (684)	Combo (2280)	UCD	NYU	EXT
Adipose	0.52 (0.34-0.69)	0.68 (0.51-0.81)	0.33 (0.23-0.44)	0.50 (0.36-0.64)	0.43 (0.31-0.56)
Adrenal	0.33 (0.10-0.65)	0.52 (0.37-0.68)	0.41 (0.26-0.58)	0.60 (0.36-0.81)	0.59 (0.36-0.79)
Appendix	0.86 (0.57-0.98)	0.79 (0.59-0.92)	0.61 (0.42-0.77)	0.73 (0.52-0.88)	0.58 (0.37-0.77)
Artery	0.70 (0.35-0.93)	0.50 (0.35-0.65)	0.31 (0.18-0.45)	0.62 (0.42-0.79)	0.38 (0.21-0.58)
Bladder	0.45 (0.17-0.77)	0.48 (0.26-0.70)	0.60 (0.26-0.88)	0.14 (0.03-0.35)	0.35 (0.22-0.51)
Blood	1.00 (0.54-1.00)	0.91 (0.76-0.98)	1.00 (0.85-1.00)	0.50 (0.12-0.88)	0.88 (0.69-0.97)
Bone	0.80 (0.28-0.99)	0.67 (0.52-0.80)	0.76 (0.50-0.93)	0.57 (0.41-0.72)	0.60 (0.44-0.75)
Breast	0.36 (0.13-0.65)	0.95 (0.74-1.00)	0.73 (0.45-0.92)	0.65 (0.41-0.85)	0.29 (0.18-0.43)
Bronchiole	0.58 (0.28-0.85)	0.42 (0.26-0.59)	0.54 (0.25-0.81)	0.49 (0.35-0.63)	0.65 (0.38-0.86)
Cartilage	0.82 (0.48-0.98)	0.74 (0.58-0.86)	1.00 (0.81-1.00)	0.33 (0.22-0.46)	0.75 (0.51-0.91)
Cerebellum	0.73 (0.45-0.92)	1.00 (0.79-1.00)	0.61 (0.36-0.83)	0.85 (0.62-0.97)	1.00 (0.72-1.00)
Cervix	0.25 (0.05-0.57)	0.62 (0.42-0.79)	0.38 (0.25-0.53)	0.58 (0.33-0.80)	0.62 (0.38-0.82)
Colon-rectum	0.76 (0.50-0.93)	0.66 (0.47-0.81)	0.54 (0.37-0.70)	0.42 (0.28-0.57)	0.72 (0.51-0.88)
Epididymis	1.00 (0.48-1.00)	0.82 (0.63-0.94)	0.61 (0.39-0.80)	0.29 (0.08-0.58)	0.39 (0.22-0.58)
Esophagus	0.27 (0.06-0.61)	0.58 (0.33-0.80)	0.49 (0.32-0.65)	0.36 (0.11-0.69)	0.50 (0.28-0.72)
Eye	1.00 (0.48-1.00)	0.77 (0.59-0.90)	0.75 (0.53-0.90)	0.56 (0.35-0.76)	0.94 (0.71-1.00)
Gallbladder	0.50 (0.16-0.84)	0.56 (0.30-0.80)	0.82 (0.48-0.98)	0.45 (0.23-0.68)	0.39 (0.23-0.58)
Heart	0.00 (0.00-0.97)	0.38 (0.09-0.76)	0.50 (0.19-0.81)	0.54 (0.25-0.81)	0.57 (0.29-0.82)
Kidney	0.50 (0.01-0.99)	0.64 (0.43-0.82)	1.00 (0.79-1.00)	0.67 (0.22-0.96)	0.57 (0.18-0.90)
Liver	0.90 (0.55-1.00)	0.83 (0.63-0.95)	0.73 (0.45-0.92)	0.00 (0.00-0.71)	0.43 (0.24-0.63)
Lung	0.67 (0.38-0.88)	0.70 (0.51-0.85)	0.80 (0.56-0.94)	0.72 (0.51-0.88)	0.75 (0.59-0.87)
Lymphoid	0.64 (0.31-0.89)	0.89 (0.72-0.98)	1.00 (0.72-1.00)	0.59 (0.36-0.79)	0.93 (0.66-1.00)
Muscle	0.72 (0.47-0.90)	0.54 (0.37-0.71)	0.86 (0.65-0.97)	0.72 (0.53-0.86)	0.93 (0.76-0.99)
Nerve	0.41 (0.18-0.67)	0.68 (0.51-0.81)	0.57 (0.39-0.74)	0.30 (0.19-0.43)	0.52 (0.31-0.72)
Ovary	0.50 (0.23-0.77)	0.53 (0.35-0.71)	0.71 (0.51-0.87)	0.80 (0.52-0.96)	0.54 (0.33-0.74)
Pancreas	0.67 (0.09-0.99)	0.75 (0.35-0.97)	1.00 (0.54-1.00)	0.19 (0.06-0.38)	0.71 (0.42-0.92)
Parotid	0.89 (0.52-1.00)	1.00 (0.80-1.00)	0.70 (0.47-0.87)	0.88 (0.47-1.00)	0.92 (0.62-1.00)
Pituitary	0.75 (0.19-0.99)	0.67 (0.35-0.90)	0.88 (0.64-0.99)	0.56 (0.31-0.78)	1.00 (0.16-1.00)
Prostate	0.79 (0.54-0.94)	0.73 (0.54-0.88)	0.83 (0.59-0.96)	0.67 (0.38-0.88)	0.31 (0.09-0.61)
Skin	0.31 (0.09-0.61)	0.90 (0.68-0.99)	0.54 (0.37-0.71)	0.68 (0.45-0.86)	0.76 (0.55-0.91)
Small-bowel	0.71 (0.44-0.90)	0.68 (0.48-0.84)	0.74 (0.49-0.91)	0.64 (0.43-0.82)	0.67 (0.45-0.84)
Spleen	0.75 (0.35-0.97)	0.88 (0.72-0.97)	0.74 (0.49-0.91)	0.69 (0.39-0.91)	0.75 (0.53-0.90)
Stomach	0.62 (0.24-0.91)	0.33 (0.13-0.59)	0.60 (0.41-0.77)	0.64 (0.35-0.87)	0.29 (0.16-0.45)
Testes	0.57 (0.18-0.90)	0.55 (0.32-0.77)	1.00 (0.75-1.00)	0.30 (0.15-0.49)	0.45 (0.17-0.77)
Thyroid	1.00 (0.72-1.00)	0.91 (0.71-0.99)	0.68 (0.48-0.84)	1.00 (0.77-1.00)	0.83 (0.63-0.95)
Tongue	0.54 (0.25-0.81)	0.76 (0.60-0.89)	0.54 (0.33-0.73)	0.59 (0.41-0.75)	0.67 (0.43-0.85)
Uterus	0.47 (0.21-0.73)	0.59 (0.36-0.79)	1.00 (0.63-1.00)	0.50 (0.21-0.79)	0.54 (0.33-0.74)
Vein	0.45 (0.23-0.68)	0.62 (0.44-0.78)	0.19 (0.07-0.37)	0.53 (0.34-0.72)	0.64 (0.31-0.89)

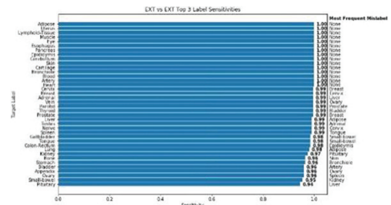
The accuracy (with 95% CI) for both combination models (684 and 2280) and each of the three individual institutions when tested against public domain images. UCD: University of California Davis; NYU: New York University; EXT: External dataset; CI: Confidence interval

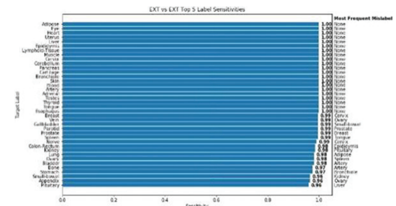
Appendix 2: Per-label sensitivity graphs for each UCD, NYU, and EXT testing permutation. The right column of each graph indicates the most frequent mislabel for the target label indicated by the left y-axis

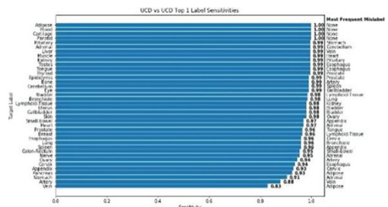
Differential criteria

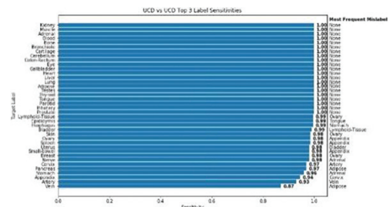
Top 1 | **Top 3** | **Top 5**

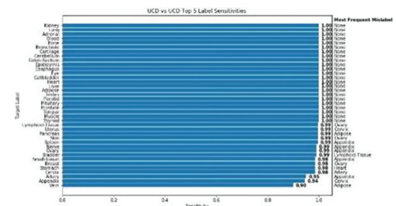


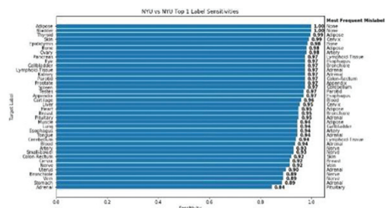


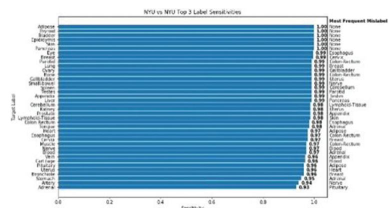


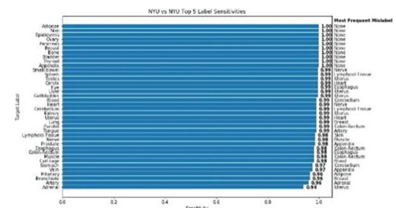


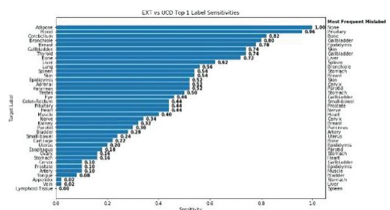


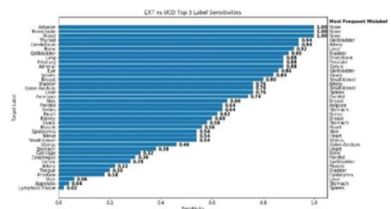


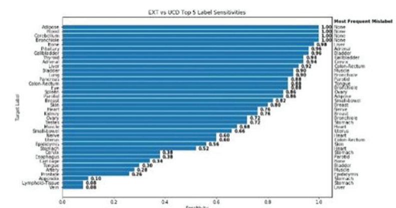


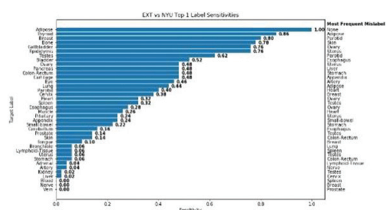


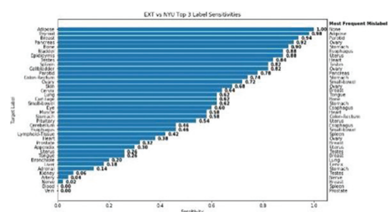


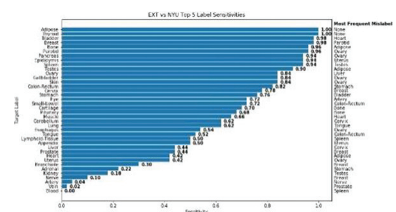


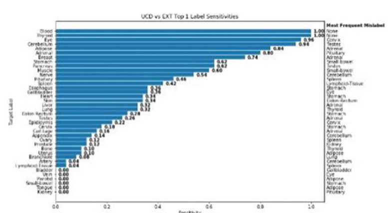


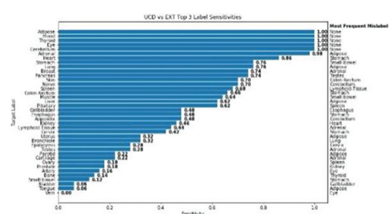


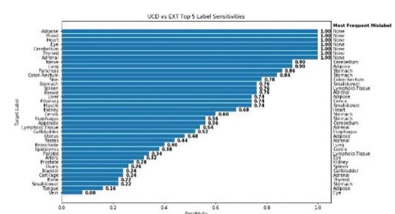












Contd...

Appendix 2: Contd...

Differential criteria