# The Temple University Digital Pathology Corpus: The Breast Tissue Subset

*Z. Wevodau[1], B. Doshna[1], N. Jhala[2], I. Akhtar[2], I. Obeid[1] and J. Picone[1]*

1. The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
2. The Lewis Katz School of Medicine, Temple University, Philadelphia, Pennsylvania, USA
{zoe.wevodau, benjamin.doshna, iobeid, picone}@temple.edu, {nirag.jhala, israh.akhtar}@tuhs.temple.edu

The Neural Engineering Data Consortium (NEDC) is developing the Temple University Digital Pathology Corpus (TUDP), an open source database of high-resolution images from scanned pathology samples [1], as part of its National Science Foundation-funded Major Research Instrumentation grant titled "MRI: High Performance Digital Pathology Using Big Data and Machine Learning" [2]. The long-term goal of this project is to release one million images. We have currently scanned over 100,000 images and are in the process of annotating breast tissue data for our first official corpus release, v1.0.0. This release contains 3,505 annotated images of breast tissue including 74 patients with cancerous diagnoses (out of a total of 296 patients). In this poster, we will present an analysis of this corpus and discuss the challenges we have faced in efficiently producing high quality annotations of breast tissue.

It is well known that state of the art algorithms in machine learning require vast amounts of data. Fields such as speech recognition [3], image recognition [4] and text processing [5] are able to deliver impressive performance with complex deep learning models because they have developed large corpora to support training of extremely high-dimensional models (e.g., billions of parameters). Other fields that do not have access to such data resources must rely on techniques in which existing models can be adapted to new datasets [6]. A preliminary version of this breast corpus release was tested in a pilot study using a baseline machine learning system, ResNet18 [7], that leverages several open-source Python tools.

The pilot corpus was divided into three sets: train, development, and evaluation. Portions of these slides were manually annotated [1] using the nine labels in Table 1 [8] to identify five to ten examples of pathological features on each slide. Not every pathological feature is annotated, meaning excluded areas can include focuses particular to these labels that are not used for training. A summary of the number of patches within each label is given in Table 2. To maintain a balanced training set, 1,000 patches of each label were used to train the machine learning model. Throughout all sets, only annotated patches were involved in model development.

The performance of this model in identifying all the patches in the evaluation set can be seen in the confusion matrix of classification accuracy in Table 3. The highest performing labels were background, 97% correct identification, and artifact, 76% correct identification. A correlation exists between labels with more than 6,000 development

Table 1. A summary of the annotation labels used in the TUDP Corpus

| Label | Name | Description |
|---|---|---|
| artf | Artifact | Grease pen marks, stitches, and other non-histological features |
| bckg | Background | Stroma and other connective tissue |
| dcis | Ductal Carcinoma in Situ | Ductal carcinoma in situ and lobular carcinoma in situ |
| indc | Invasive Ductal Carcinoma | Invasive ductal carcinoma, invasive lobular carcinoma, and invasive mammary carcinoma |
| infl | Inflammation | Regions with high concentration of lymphocytes, indicating an immune response |
| nneo | Nonneoplastic | Abnormal growths that are not classified as cancerous, these include the subcategories of fibrosis, hyperplasia, sclerosing adenosis, calcifications, apocrine metaplasia, duct ectasia |
| norm | Normal | Normal ducts and lobules |
| null | Null | Indistinguishable tissue that arose from damage during tissue processing |
| susp | Suspicious | Regions of atypical ductal and lobular hyperplasia that are at risk for progressing to ductal and lobular carcinomas |

Table 2. An overview of the annotated pilot corpus

| Label | Train | Dev | Eval | Total |
|---|---|---|---|---|
| artf | 17,147 | 6,513 | 6,881 | 30,541 |
| bckg | 329,404 | 110,425 | 110,599 | 550,428 |
| dcis | 5,626 | 1,945 | 1,900 | 9,471 |
| indc | 6,574 | 2,528 | 2,599 | 11,701 |
| infl | 1,144 | 473 | 457 | 2,074 |
| nneo | 15,183 | 5,684 | 5,770 | 26,637 |
| norm | 4,524 | 1,755 | 1,745 | 8,024 |
| susp | 15,445 | 5,768 | 5,607 | 26,820 |

Table 3. A confusion matrix for a baseline image classification system

| | artf | bckg | dcis | indc | infl | nneo | norm | susp |
|---|---|---|---|---|---|---|---|---|
| artf | 76% | 24% | 0% | 0% | 0% | 0% | 0% | 0% |
| bckg | 1% | 97% | 0% | 0% | 0% | 1% | 1% | 1% |
| dcis | 0% | 0% | 64% | 16% | 8% | 4% | 1% | 6% |
| indc | 0% | 0% | 3% | 41% | 55% | 0% | 0% | 1% |
| infl | 0% | 2% | 2% | 56% | 36% | 1% | 1% | 3% |
| nneo | 0% | 23% | 8% | 1% | 3% | 41% | 13% | 11% |
| norm | 6% | 25% | 4% | 4% | 4% | 41% | 18% | 4% |
| susp | 1% | 6% | 29% | 2% | 9% | 18% | 6% | 29% |

patches and accurate performance on the evaluation set. Additionally, these results indicated a need to further refine the annotation of invasive ductal carcinoma (“indc”), inflammation (“infl”), nonneoplastic features (“nneo”), normal (“norm”) and suspicious (“susp”).

This pilot experiment motivated changes to the corpus that will be discussed in detail in this poster presentation. To increase the accuracy of the machine learning model, we modified how we addressed underperforming labels. One common source of error arose with how non-background labels were converted into patches. Large areas of background within other labels were isolated within a patch resulting in connective tissue misrepresenting a non-background label. In response, the annotation overlay margins were revised to exclude benign connective tissue in non-background labels.

Corresponding patient reports and supporting immunohistochemical stains further guided annotation reviews. The microscopic diagnoses given by the primary pathologist in these reports detail the pathological findings within each tissue site, but not within each specific slide. The microscopic diagnoses informed revisions specifically targeting annotated regions classified as cancerous, ensuring that the labels “indc” and “dcis” were used only in situations where a micropathologist diagnosed it as such. Further differentiation of cancerous and precancerous labels, as well as the location of their focus on a slide, could be accomplished with supplemental immunohistochemically (IHC) stained slides. When distinguishing whether a focus is a nonneoplastic feature versus a cancerous growth, pathologists employ antigen targeting stains to the tissue in question to confirm the diagnosis. For example, a nonneoplastic feature of usual ductal hyperplasia will display diffuse staining for cytokeratin 5 (CK5) and no diffuse staining for estrogen receptor (ER), while a cancerous growth of ductal carcinoma in situ will have negative or focally positive staining for CK5 and diffuse staining for ER [9]. Many tissue samples contain cancerous and non-cancerous features with morphological overlaps that cause variability between annotators. The informative fields IHC slides provide could play an integral role in machine model pathology diagnostics.

Following the revisions made on all the annotations, a second experiment was run using ResNet18. Compared to the pilot study, an increase of model prediction accuracy was seen for the labels indc, infl, nneo, norm, and null. This increase is correlated with an increase in annotated area and annotation accuracy. Model performance in identifying the suspicious label decreased by 25% due to the decrease of 57% in the total annotated area described by this label. A summary of the model performance is given in Table 4, which shows the new prediction accuracy and the absolute change in error rate compared to Table 3.

The breast tissue subset we are developing includes 3,505 annotated breast pathology slides from 296 patients. The average size of a scanned SVS file is 363 MB. The annotations are stored in an XML format. A CSV version of the annotation file is also available which provides a flat, or simple, annotation that is easy for machine learning researchers to access and interface to their systems. Each patient is identified by

Table 4. A comparison matrix of the experiments done before and after data revision

| | artf | bckg | dcis | indc | infl | nneo | norm | null | susp |
|---|---|---|---|---|---|---|---|---|---|
| **artf** | 95% (+19%) | 2% (-22%) | 0% (0%) | 0% (0%) | 0% (0%) | 1% (+1%) | 1% (+1%) | 1% (+1%) | 0% (0%) |
| **bckg** | 0% (-1%) | 91% (-6%) | 0% (0%) | 1% (+1%) | 0% (0%) | 2% (+1%) | 3% (+2%) | 2% (+2%) | 1% (0%) |
| **dcis** | 0% (0%) | 0% (0%) | 42% (-22%) | 7% (-9%) | 6% (+2%) | 24% (+20%) | 19% (+18%) | 1% (+1%) | 1% (-5%) |
| **indc** | 0% (0%) | 1% (+1%) | 3% (0%) | 65% (+24%) | 6% (-49%) | 8% (+8%) | 9% (+9%) | 5% (+5%) | 3% (+2%) |
| **infl** | 0% (0%) | 1% (-1%) | 1% (-1%) | 3% (-53%) | 63% (+27%) | 8% (+7%) | 19% (+18%) | 2% (+2%) | 3% (0%) |
| **nneo** | 0% (0%) | 2% (-21%) | 8% (0%) | 1% (0%) | 2% (-1%) | 42% (+1%) | 40% (+27%) | 2% (+2%) | 3% (-8%) |
| **norm** | 0% (-6%) | 3% (-22%) | 1% (-3%) | 0% (-4%) | 1% (-3%) | 11% (-30%) | 82% (+64%) | 1% (+1%) | 1% (-3%) |
| **null** | 0% (0%) | 14% (+14%) | 0% (0%) | 0% (0%) | 3% (+3%) | 6% (+6%) | 21% (+21%) | 54% (+54%) | 2% (+2%) |
| **susp** | 0% (-1%) | 1% (-5%) | 14% (-15%) | 3% (+1%) | 10% (+1%) | 23% (+5%) | 43% (+37%) | 2% (+2%) | 4% (-25%) |

an anonymized medical reference number. Within each patient's directory, one or more sessions are identified, also anonymized to the first of the month in which the sample was taken. These sessions are broken into groupings of tissue taken on that date (in this case, breast tissue). A deidentified patient report stored as a flat text file is also available. Within these slides there are a total of 16,971 total annotated regions with an average of 4.84 annotations per slide. Among those annotations, 8,035 are non-cancerous (normal, background, null, and artifact,) 6,222 are carcinogenic signs (inflammation, nonneoplastic and suspicious,) and 2,714 are cancerous labels (ductal carcinoma in situ and invasive ductal carcinoma in situ.)

The individual patients are split up into three sets: train, development, and evaluation. Of the 74 cancerous patients, 20 were allotted for both the development and evaluation sets, while the remain 34 were allotted for train. The remaining 222 patients were split up to preserve the overall distribution of labels within the corpus. This was done in hope of creating control sets for comparable studies. Overall, the development and evaluation sets each have 80 patients, while the training set has 136 patients.

In a related component of this project, slides from the Fox Chase Cancer Center (FCCC) Biosample Repository (*https://www.foxchase.org/research/facilities/genetic-research-facilities/biosample-repository-facility*) are being digitized in addition to slides provided by Temple University Hospital. This data includes 18 different types of tissue including approximately 38.5% urinary tissue and 16.5% gynecological tissue. These slides and the metadata provided with them are already anonymized and include diagnoses in a spreadsheet with sample and patient ID. We plan to release over 13,000 unannotated slides from the FCCC Corpus simultaneously with v1.0.0 of TUDP. Details of this release will also be discussed in this poster.

Few digitally annotated databases of pathology samples like TUDP exist due to the extensive data collection and processing required. The breast corpus subset should be released by November 2021. By December 2021 we should also release the unannotated FCCC data. We are currently annotating urinary tract data as well. We expect to release about 5,600 processed TUH slides in this subset. We have an additional 53,000 unprocessed TUH slides digitized. Corpora of this size will stimulate the development of a new generation of deep learning technology. In clinical settings where resources are limited, an assistive diagnoses model could support pathologists' workload and even help prioritize suspected cancerous cases.

ACKNOWLEDGMENTS

This material is supported by the National Science Foundation under grants nos. CNS-1726188 and 1925494. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] N. Shawki et al., "The Temple University Digital Pathology Corpus," in *Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications*, 1st ed., I. Obeid, I. Selesnick, and J. Picone, Eds. New York City, New York, USA: Springer, 2020, pp. 67-104. *https://www.springer.com/gp/book/9783030368432*.

[2] J. Picone, T. Farkas, I. Obeid, and Y. Persidsky, "MRI: High Performance Digital Pathology Using Big Data and Machine Learning." Major Research Instrumentation (MRI), Division of Computer and Network Systems, Award No. 1726188, January 1, 2018 – December 31, 2021. *https://www.isip.piconepress.com/projects/nsf_dpath/*.

[3] A. Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association* (INTERSPEECH), 2020, pp. 5036-5040. *https://doi.org/10.21437/interspeech.2020-3015*.

[4] C.-J. Wu *et al.*, "Machine Learning at Facebook: Understanding Inference at the Edge," in *Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2019, pp. 331–344. *https://ieeexplore.ieee.org/document/8675201*.

[5] I. Caswell and B. Liang, "Recent Advances in Google Translate," Google AI Blog: The latest from Google Research, 2020. [Online]. Available: *https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html*. [Accessed: 01-Aug-2021].

[6] V. Khalkhali, N. Shawki, V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, "Low Latency Real-Time Seizure Detection Using Transfer Deep Learning," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (SPMB), 2021, pp. 1-7. *https://www.isip.piconepress.com/publications/conference_proceedings/2021/ieee_spmb/eeg_transfer_learning/*.

[7] J. Picone, T. Farkas, I. Obeid, and Y. Persidsky, "MRI: High Performance Digital Pathology Using Big Data and Machine Learning," Philadelphia, Pennsylvania, USA, 2020. *https://www.isip.piconepress.com/publications/reports/2020/nsf/mri_dpath/*.

[8] I. Hunt, S. Husain, J. Simons, I. Obeid, and J. Picone, "Recent Advances in the Temple University Digital Pathology Corpus," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (SPMB), 2019, pp. 1–4. *https://ieeexplore.ieee.org/document/9037859*.

[9] A. P. Martinez, C. Cohen, K. Z. Hanley, and X. (Bill) Li, "Estrogen Receptor and Cytokeratin 5 Are Reliable Markers to Separate Usual Ductal Hyperplasia From Atypical Ductal Hyperplasia and Low-Grade Ductal Carcinoma In Situ," *Arch. Pathol. Lab. Med.*, vol. 140, no. 7, pp. 686–689, Apr. 2016. *https://doi.org/10.5858/arpa.2015-0238-OA*.

# The Temple University Digital Pathology Corpus: The Breast Tissue Subset

NEURAL ENGINEERING DATA CONSORTIUM
www.nedcdata.org

Z. Wevodau, B. Doshna, I. Obeid and J. Picone
The Neural Engineering Data Consortium, Temple University

N. Jhala ad I. Akhtar
The Lewis Katz School of Medicine, Temple University

College of Engineering
Temple University

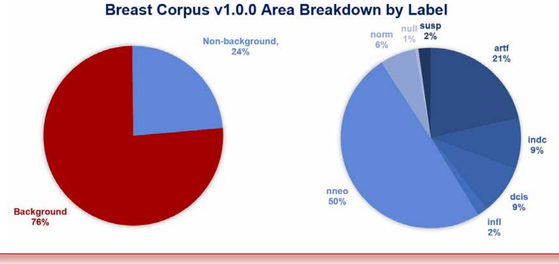
## Abstract


- Fields such as speech and image recognition have delivered impressive performance with complex deep learning models because they have developed large corpora to support training of extremely high-dimensional models (e.g., billions of parameters).
- Many bioengineering applications, such as digital pathology, lack these resources.
- The Breast Tissue subset of the Temple University Digital Pathology Corpus (DPATH) is our first official release and contains 3,505 slides from 296 patients.
- Portions of these slides have been manually annotated using nine labels and include an overall classification of cancerous vs. noncancerous.
- The annotations have been carefully reviewed by TUHS pathologists and a team of UG annotators.
- As part of this project, we will release a second corpus of 13,865 unannotated slides from the Biosample Repository at Fox Chase Cancer Center.


## Breast Tissue Corpus v1.0.0 Statistics

- The 3,505 slides belong to 296 patients with an average of 11.8 slides per patient.
- Of these 296 patients, 74 patients contain cancerous features (4.3% of the total annotated area): ductal carcinoma in situ or invasive ductal carcinoma.
- Slides are scanned at a 20x magnification (0.50 microns per pixel) and stored in a compressed tiff SVS format. The average file size is 363 MB.
- Each image includes an annotation file in XML and CSV formats.
- Pathology reports are also available for each set of slides. There are 316 reports, or an average of 11 slides per report.
- Reports are available as flat text files and contain sections such as “Clinical History,” “Microscopic Diagnosis” and “Gross Tissue Description.”
- Reports have been manually anonymized by our annotation team.
- There are over 54,000 words in these reports with over 13,000 unique words.


- Work is underway in a separate project to parse these documents into medical concepts.
- Of the total annotated area, 76% is background connective or adipose tissue. The remaining 24% is split into 8 feature labels.




## Annotation Labels

- Using Aperio ImageScope, nine labels were used to identify five to ten examples of pathological features on each slide.
- Certain labels have subsets of more specific features such as nonneoplastic features which covers apocrine metaplasia, fibroadenomas, sclerosing adenosis, calcifications, fibrocystic changes, and ductal hyperplasia.
- Not every pathological feature is annotated, meaning excluded areas can include focuses particular to these labels that are not used for model training.

| Label | Name | Description |
|---|---|---|
| artf | Artifact | Grease pen marks, stitches, and other non-histological features |
| bckg | Background | Stroma and other connective tissue |
| null | Null | Indistinguishable tissue caused by tissue processing damage |
| norm | Normal | Normal ducts and lobules |
| infl | Inflammation | Regions with high concentration of lymphocytes, indicating an immune response |
| nneo | Nonneoplastic | Abnormal growths that are not classified as cancerous, including the subcategories of fibrosis, hyperplasia, sclerosing adenosis, calcifications, apocrine metaplasia, and duct ectasia |
| susp | Suspicious | Regions of atypical ductal and lobular hyperplasia that are at risk for progressing to ductal and lobular carcinomas |
| dcis | Ductal Carcinoma in Situ | Ductal carcinoma in situ and lobular carcinoma in situ |
| indc | Invasive Ductal Carcinoma | Invasive ductal carcinoma, invasive lobular carcinoma, and invasive mammary carcinoma |

## Label Confusion in Pilot Study

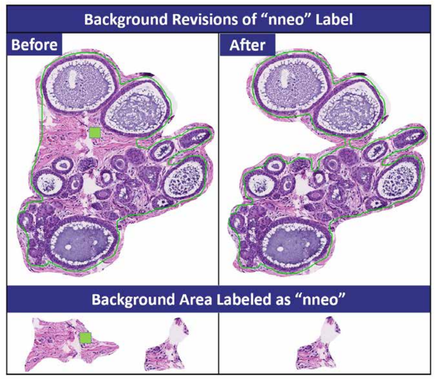
- A preliminary version of the breast corpus was tested in a pilot study using a baseline machine learning system, ResNet18, that leverages open-source Python tools.
- The highest performing labels in the development set were background (97% correct identification) and artifact (76% correct identification).
- A correlation existed between labels with more than 6,000 development patches and accurate performance on the evaluation set.
- Background was identified as the largest source of error in the identification of other labels.
- Model confusion between invasive ductal carcinoma (“indc”) and inflammation (“infl”) indicated annotator error.
- Labels with a correct identification ratio less than 0.75, dcis, indc, infl, nneo, norm, and susp, required further revisions.

Model's Prediction

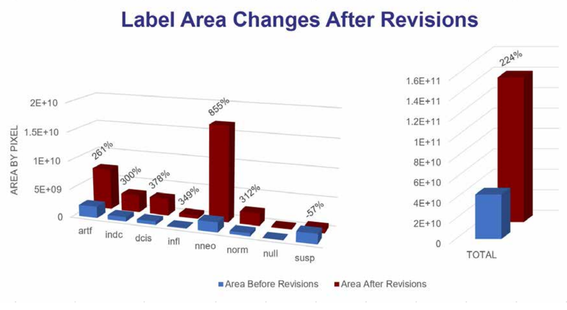
| Label | artf | bckg | dcis | indc | infl | nneo | norm | null | susp |
|---|---|---|---|---|---|---|---|---|---|
| artf | 0.76 | 0.24 | 0 | 0 | 0 | 0 | 0 | - | 0 |
| bckg | 0.01 | 0.97 | 0 | 0 | 0 | 0.01 | 0.01 | - | 0.01 |
| dcis | 0 | 0 | 0.64 | 0.16 | 0.08 | 0.04 | 0.01 | - | 0.06 |
| indc | 0 | 0 | 0.03 | 0.41 | 0.55 | 0 | 0 | - | 0.01 |
| infl | 0 | 0.02 | 0.02 | 0.56 | 0.36 | 0.01 | 0.01 | - | 0.03 |
| nneo | 0 | 0.23 | 0.08 | 0.01 | 0.03 | 0.41 | 0.13 | - | 0.11 |
| norm | 0 | 0.25 | 0.04 | 0.04 | 0.04 | 0.41 | 0.18 | - | 0.04 |
| null | - | - | - | - | - | - | - | - | - |
| susp | 0.01 | 0.06 | 0.29 | 0.02 | 0.09 | 0.18 | 0.06 | - | 0.29 |

## Annotation Revisions

- To increase the accuracy of the machine learning model, the annotations of underperforming labels were adjusted.
- Large areas of background within other labels were isolated within a patch resulting in connective tissue misrepresenting a non-background label.
- The annotation overlay margins were revised to exclude benign connective tissue in non-background labels:




- Daily meetings with a microscopic pathologist guided diagnoses of areas not specifically mentioned in patient reports.
- Usage of cancerous labels, dcis and indc, only occurred in instances where patient reports’ microscopic diagnosis indicated.
- Under annotated features such as inflammation, null, or normal tissue were identified to balance the area of each label.
- Immunohistochemical staining indicated reference points for the location of cancerous foci on slides containing both cancerous and precancerous features (e.g., atypical ductal hyperplasia vs low grade ductal carcinoma in situ using CK5 and ER).




- Revisions resulted in 34,544,211 64x64 pixel patches, a 224% increase in comparison to the area originally annotated in our pilot corpus release.
- All labels at least doubled in area except susp which represents a diagnosis between precancerous and cancerous. The decrease in suspicious area annotated is likely indicative of increased histological understanding and correction to either nneo or dcis/indc.

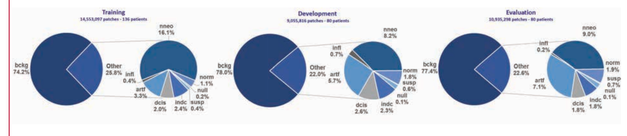
## Towards Improving Performance

- We compared performance of the baseline ResNet18 system on our preliminary release to performance on the expanded version of the corpus:

Model's Prediction

| Label | artf | bckg | dcis | indc | infl | nneo | norm | null | susp |
|---|---|---|---|---|---|---|---|---|---|
| artf | 0.185 | -0.22 | 0 | 0 | 0 | 0.005 | 0.015 | 0.015 | 0 |
| bckg | -0.01 | -0.06 | 0 | 0.005 | 0 | 0.01 | 0.02 | 0.015 | -0.005 |
| dcis | 0 | 0.005 | -0.22 | -0.09 | -0.025 | 0.195 | 0.175 | 0.01 | -0.045 |
| indc | 0 | 0.015 | -0.005 | 0.24 | -0.49 | 0.08 | 0.09 | 0.045 | 0.02 |
| infl | 0 | -0.01 | -0.01 | -0.53 | 0.265 | 0.075 | 0.18 | 0.015 | 0.005 |
| nneo | 0 | -0.215 | 0.005 | 0.005 | -0.01 | 0.005 | 0.27 | 0.02 | -0.08 |
| norm | 0 | -0.22 | -0.035 | -0.04 | -0.035 | -0.3 | 0.64 | 0.015 | -0.03 |
| null | 0 | 0.14 | 0 | 0.005 | 0.03 | 0.06 | 0.21 | 0.535 | 0.015 |
| susp | -0.01 | -0.055 | -0.15 | 0.015 | 0.005 | 0.05 | 0.375 | 0.02 | -0.25 |

- An increase in model prediction accuracy was seen for labels artf, indc, infl, nneo, norm, and null.
- The increase in accuracy is correlated with an increase in annotated area and annotation accuracy.
- Inversely, the model performance identifying susp labels decreased by 25% due to a decrease of 57% n the annotated area described by this label.
- The decrease in the model’s ability to identify dcis by 22% could be attributed to the physical similarities dcis shares with nneo’s ductal hyperplasia.
- Training, development, and evaluation sets have been partitioned within this release. Of the 74 cancerous patients, 20 patients each were assigned to the development and evaluation sets, and the remaining 34 to the training set. This ensured both dev and eval sets had a similar distribution of indc and dcis labels. The remaining 222 patients were split up to preserve the overall distribution of labels withing the entire breast corpus.




## Summary and Future Work

- We will release 13,865 slides captured from the Biosample Repository at Fox Chase Cancer Center (FCCC). These slides contain 18 types of tissue (38.5% prostate, 16.5% gynecological, 45% other).
- We expect to release an additional 5,600 TUH annotated slides of urinary tissue (mainly bladder and prostate tissue).
- We will also release open-source software to analyze and classify images in 1Q’2022.

## Acknowledgements


- This material is supported by the National Science Foundation under grants nos. CNS-1726188 and 1925494. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
- We are grateful for the support of Denise Connolly and Chao Wu for making the FCCC data available and facilitating the transfer process. The FCCC data is an extremely important historical archive.