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 3,618 annotated images of breast tissue including 63 patients with cancerous diagnoses (out of a total of 299 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.

The breast tissue subset we are developing includes 3,618 annotated breast pathology slides from 299 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 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 9,582 total annotated regions with an average of 7.18 annotations per slide. Among those annotations, 3,497 are non-cancerous (normal, background, null, and artifact,) 3,376 are carcinogenic signs (inflammation, nonneoplastic and suspicious,) and 2,709 are cancerous labels (ductal carcinoma in situ and invasive ductal carcinoma in situ.)

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 41,000 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*.