NEURAL ENGINEERING
DATA CONSORTIUM
www.nedcdata.org

The Temple University Digital Pathology Corpus: The Breast Tissue Subset

Z. Wevodau, B. Doshna, I. Obeid and J. Picone — N. Jhala ad I. Akhtar

The Neural Engineering Data Consortium, Temple University — The Lewis Katz School of Medicine, Temple University

College of Engineering
Temple University

Abstract

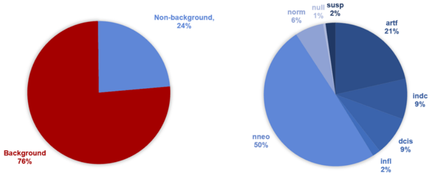
- Fields such as speech and image recognition have delivered impressive performance with complex deep learning models because they have developed large corpora to support training of extremely high-dimensional models (e.g., billions of parameters).
- Many bioengineering applications, such as digital 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.

Breast Corpus v1.0.0 Area Breakdown by Label



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

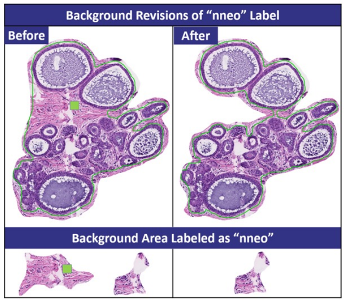
- A preliminary version of the breast corpus was tested in a pilot study using a baseline machine learning system, ResNet18, that leverages open-source Python tools.
- The highest performing labels in the development set were background (97% correct identification) and artifact (76% correct identification).
- 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

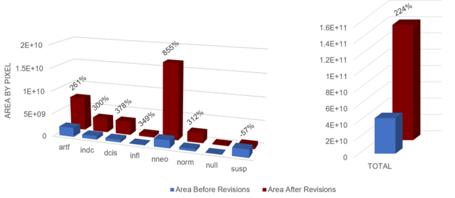
Annotation Revisions

- 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:


Background Revisions of “nneo” Label
Before
After
Background Area Labeled as “nneo”

- 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).

Label Area Changes After Revisions



- 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.

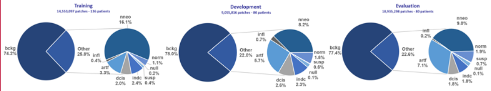
Towards Improving Performance

- 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.