Big Data Resources for Digital Pathology

S.S. Shalamzari[1], M. Bagritsevich[1], A. Melles[1], I. Obeid[1], J. Picone[1],
D. Connolly[2], C. Wu[2], B. Schultz[2], B. Brown[2], J. James[2], Y. Gong[2] and H. Wu[2]

1. The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
2. Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA
{somayeh.seifi.shalamzari, maria.bagritsevich, anne-mai.melles, iyad.obeid, picone}@temple.edu
{Denise.Connolly, Chao.Wu, Bryant.Schultz, Brittaney.Brown,
Johanna.James, Yulan.Gong, and Hong.Wu}@fccc.edu

The Neural Engineering Data Consortium (NEDC), known for its open source data resources [1], has previously released the Breast Tissue subset of the Temple University Digital Pathology Corpus (TUDP), which consists of 3,505 partially annotated images [2]. This corpus is part of a much larger repository of over 100,000 images that will be released as part of NEDC's digital pathology resources in 2023. In this abstract, we introduce our recently released corpus of 14,288 digital pathology images that were collected from Fox Chase Cancer Center's (FCCC) Biosample Repository [3], and describe some changes being made to TUDP to organize these corpora in a unified framework.

A. The Fox Chase Digital Pathology Corpus

Fox Chase Cancer Center's Biosample Repository Facility (BRF) coordinates the ethical collection and distribution of human biospecimens and associated clinical data under Institutional Review Board approved protocols [3]. Over the past two years we have digitized a large portion of the FCCC repository using the same process we have employed for TUDP [4]. This corpus will be referred to as the Fox Chase Digital Pathology (FCDP) Corpus. FCDP contains 14,288 high resolution images (typically 50K x 50K pixels) that are stored in Aperio's popular svs format [5]. The average file size is 244 MB. There are 13,230 subjects represented in the corpus, so most subjects appear once. The entire corpus requires 3.5 TB of storage. A svs file format is used because Aperio has done an excellent job using this format to compress the images. An open source viewer, ImageScope [6], is also available that supports this format. The baseline image contained in a svs file is stored using a lossless tiff compression format.

Given that the FCDP data originates from a National Cancer Institute designated Comprehensive Cancer Center research facility and hospital, the data includes a much higher percentage of cancerous slides in comparison to the TUDP data. The metadata available for each slide, which is summarized in Appendix A, includes a numeric slide and specimen identification number, as well as assorted information about the tissue, tumor, and stain type. The metadata is provided in an Excel spreadsheet format with a single line entry per image. A small percentage of slides (168) have multiple metadata entries reflecting multiple diagnoses. The first four columns (shaded in green) are derived values that we have added to the original spreadsheet provided by the BRF. The remaining columns reflect the original metadata.

There are a few columns that will be of special interest to machine learning researchers. The column labeled K with a heading of "ICDO Code (Tissue Site)" contains 192 distinct ICDO codes [7]. A brief explanation of these codes is provided in column L labeled "Tissue Site.1." The frequency of occurrence of these types is shown in Appendix B. A concise description of each tissue type is presented in [7]. The staining method employed for the slides in FCDP is Hematoxylin and Eosin (H&E). A complete list of staining methods can be found in Appendix C. These were adapted from [8]-[10].

To facilitate machine learning experiments, the data has been classified based on the grade calculated for tumors. According to a universal tumor grading [11], cancer grades are classified into four risk groups (well differentiated, moderately differentiated, poorly differentiated, and undifferentiated). These risk groups can be classified under three general classes: low grade (lg), intermediate grade (ig), and high grade (hg) as

shown in Table 1. We have also added an unknown label to each of these attributes to facilitate data management.

Table 1. Universal Tumor Grading System

Grade Group	Grade Risk	Class
I	well differentiated	low grade (lg)
II	moderately differentiated	intermediate grade (ig)
III	poorly differentiated	high grade (hg)
IV	undifferentiated	
X	unknown	unknown

Cancer is graded by comparing its appearance to that of normal cells under a microscope [11]. A tumor that contains well differentiated cells tends to appear as organized and closely resembles normal tissue. This is categorized as low grade (described in the metadata spreadsheet as G1, low grade, or grade group I). In contrast, tumors with poorly differentiated and undifferentiated cells that are disorganized and different from normal cells are classified as high grade (described in the metadata spreadsheet as G3, Grade III, grade group 3, poorly differentiated, G4, Grade IV, grade group 4 or undifferentiated). A moderately differentiated tumor (G2, Grade II, grade group 2, or moderately differentiated) has cells that are slightly organized under a microscope, but not to the same extent as a grade I tumor. Tumors of this type tend to grow and spread moderately. Based on the universal grading system, tumors with undetermined grading are considered as "GX" or " undetermined grade". Since there is no grade associated with them, They are not under the risk group category. These cases have been labeled using the unknown class.

The grading system used for cancer depends on the type of cancer. For example, prostate cancer uses a Gleason score [12][13][14], shown in Table 2. This type of data was classified into the same four categories shown in Table 1 using information found in columns AF ("Block Level Tissue Histology") and AG ("Gradeclin Desc"). In some cases the grade couldn't be determined from information in the metadata spreadsheet, so we categorized those as "unknown." In these cases, we analyzed information from other columns such as column AI ("General Stage") to supplement the available information. For example, whenever primary cancer turns into a distant metastasis [11], column AI ("General Stage") indicates the patient is facing a high grade cancer because the metastatic cancer can reach a secondary site, which means the cancer has spread to other parts of the body. Similarly, whenever column AF indicates a "Diffuse large B-cell lymphoma", we categorized the slide as hg. B-cell lymphoma is cancerous but its grade is not obvious. However, when a slide is classified as "Diffuse large B-cell lymphoma," it is definitely an indication of a high grade cancer. In some cases both columns AF and AG were empty, so these cases were categorized as "unknown."

Table 2. Gleason Score Grading System

Grade Group	Gleason Score	Risk Group	Class
G1	≤ 6	low/ very low	low grade (lg)
G2	7 (3+4)	intermediate	intermediate grade (ig)
G3	7 (4+3)		
G4	8	high/ very high	high grade (hg)
G5	9-10		

Following this type of decision-making process, and using as much information as possible from the metadata, we have classified all the slides in FCDP using the categories in Table 1. If a Gleason score was provided in column AG (e.g., "Grade II Moderately Diff / Mod Well Diff") it was used to make this determination. For some slides, though a Gleason score was not provided because the slide was not a prostate sample. In these cases, the entry in column AG indicates the grade (e.g., "Cell type not determined/stated/NA; unk pri; hi gr dysplasia," "HG: High-grade," "Well differentiated"). There are 58 unique values in column AG. Most of these (52) were fairly easy to interpret. However, for a few of these values (e.g., "T cell; T-precursor"), additional metadata was examined to make a determination. If this determination could not be made confidently from the metadata, a value of "unknown" was assigned.

Table 3. Machine learning classes based on grade

Index	Class	Freq (%)
1	hg	7,762 (54.33%)
2	ig	3,932 (27.52%)
3	lg	912 (6.38%)
4	unknown	1682 (11.77%)
TOTAL		14,288 (100.00%)

In Table 3, we show the frequency of occurrence of each class. There is still an imbalance in the data, but given the volume of data available in FCDP, and given the image sizes, this should not pose any problems for machine learning research. Further, as we will explain shortly, we can augment FCDP with TUDP.

An important goal for all the corpora that we distribute is that every filename is unique. The directory and filenameing structures are shown in Table 4. There are three primary directories: "fcdp" for the database, "v1.0.0" for version-specific data, and "svs" for image type. Subdirectories are used to further segregate the data based on subject, specimen and tissue site identifiers. The last field in the directory pathname is particularly important since it encodes the tissue site (col. K) and tumor site (col. M) using ICCO codes [7]. A unique staining code has been assigned based on the table in Appendix C. There are two types of files in this release: (1) image data stored in files ending in ".svs" and (2) metadata stored in files ending in ".csv". This convention allows users to use standard Unix commands to sort and select subsets of the data. For example, a breast tissue study could be constructed by searching for all files that begin with "c50" for the ICDO code component of the filename (e.g. find . -name '*.svs' | grep '/c50.').

B. The Temple University Digital Pathology Corpus

We have previously released a substantial subset of our TUDP data known as the TUDP Breast Tissue Corpus [4]. This corpus consists of 3,505 partially annotated images [15]. This means that typically five to ten regions of interest have been labeled on each slide along with five to ten regions containing 'background' image data. This corpus can be used to develop cancer detection technology [2].

In this abstract, we discuss the release of the entire TUDP dataset, which contains 98,382 images. The imaging process used for TUDP is identical to that employed for the FCDP data and the previous TUDP Breast Tissue Corpus [4], and hence the average file size is 244 MB per image. The entire corpus requires 24 TB of storage. One notable advantage of the TUDP dataset is its diverse representation of tissue types and staining methods. The inclusion of various staining methods is particularly significant as it facilitates generalization in machine learning systems. Doing multiclass detection of cancerous slides across many types of tissues and stains is a much more difficult machine learning problem.

Table 4. The components of the directory structure and filename are described for a typical file (*fcdp/v1.0.0/svs/00026/000262564/001010285/c67.9_c67.2/000262564_001010285_st065_xt1_t000.svs*) in FCDP. For some cells, the column referenced in parentheses refers to the corresponding column in the metadata spreadsheet.

Component	Example	Description
directory name	fcdp/v1.0.0/svs/00026/000262564/001010285/c67.9_c67.2	a typical full directory specification
database	fcdp	an acronym for the corpus
version	v1.0.0	the version number
data type	svs	the root node for the image data and annotations
sequential ID	00026	a 5-digit number used to control directory complexity
subject ID	000262564	a 9-digit anonymized subject identifier (col. E)
specimen ID	001010285	a 9-digit specimen identifier (col. F)
ICDO codes	c67.9_c67.2	ICDO codes for tissue type (c67.9) and tissue site (c67.2) (cols. K and M respectively)
Component	**Example**	**Description**
filename	000368678_001013243_st065_xt01_t000.svs	a unique filename
identifier	000368678_001013243	subject ID (col. E) and specimen ID (col. F)
stain code	st065	a code assigned to the type of stain used in the slide (col. G)
slide number	xt1	a three-character code assigned to each slide (col. H)
token number	t000	a four-character number indicated the slide number in a series
extension	svs	indicates the type of file; ".svs" is used for an image data file

The Breast Tissue subset was released before we had obtained access to FCDP. Based on our experiences with the released data and the new FCCC data, we have decided to modify the filenaming conventions for TUDP so that the data parallels FCDP. Unfortunately, though we had access to pathology reports for the breast tissue subset, we were unable to access reports for the bulk of the TUDP data. Hence, slides cannot be classified with the same level of precision as FCDP. The revised renaming convention for TUDP is shown in Table 5. One minor difference between FCDP and TUDP is the way the slide number information is encoded. In TUDP, slides are typically organized in sets with 10 to 20 slides per set. The final identifier, refered to as the token number ("tNNN"), is used to indicate the slide number in the set. These are captured as part of the label image on the slide. The label images were manually transcribed and organized into a consistent set of codes as described in Table 5. Missing information was encoded using placeholders such as "cxx.x" for the ICDO codes. The slide label image was ultimately removed from the final image files as part of the anonymization process.

The slide number component of the filename also contains information about the way the sample was prepared. Users of this corpus can think of this field as simply a catalog number that helps us connect a digitally scanned image to the specific glass slide stored in Temple University's pathology slide archive. A typical value for this field is "a001_lvl000". This field begins with a letter followed by three digits (e.g., "a059", "j001") and refers to what is known as the block cut, which is the specific section of the tissue (e.g., upper left of the breast). The second component is the level, which is usually three characters "lvl" followed by a three digits. This code is used to distinguish different cuts from the same tissue site.

Though the TUDP Breast Tissue Corpus contains partial annotations [15], TUDP at large does not currently contain any new annotations. The TUDP Breast Tissue Corpus consists of 9 labels as shown in Table 6. The non-cancerous labels are "normal," "background," and "artifact." The labels "non-neoplastic," "inflammation," and "suspected" are classified as carcinogenic signs, indicating a high probability of being cancerous in the future. The labels "ductal carcinoma" and "invasive ductal carcinoma" represent cancerous conditions. The label "indistinguishable" is categorized as "none" since it commonly arises due to issues during the cutting or staining process, making it challenging to determine its classification conclusively.

Table 5. The components of the directory structure and filename are described for a typical file (*tudp/v1.0.0/svs/00036/ aaaaaaaaa/ bbbbbbbbbbbbbbb/c50.9_cxx.x/aaaaaaaaa_bbbbbbbbb_st065_a001_lvl000_t002.svs*) in TUDP.

Component	Example	Description
directory name	tudp/v1.0.0/svs/00036/aaaaaaaaa/ bbbbbbbbbbbbbbb/c50.9_cxx.x	a typical full directory specification
database	tudp	an acronym for the corpus
version	v1.0.0	the version number
data type	svs	the root node for the image data and annotations
sequential ID	00036	a 5-digit number used to control directory complexity
subject ID	aaaaaaaaa	a 9-digit anonymized subject identifier (ID)
specimen ID	bbbbbbbbbbbbbbb	a 15-character specimen identifier (ID)
ICDO codes	c50.9_cxx.x	ICDO codes for tissue type (c50.9) and tissue site (cxx.x)
Component	**Example**	**Description**
filename	aaaaaaaaa_bbbbbbbbb_st065_a001_lvl000_t 002.svs	a unique filename
identifier	aaaaaaaaa_bbbbbbbbbbbbbbb	subject ID (col. E) and specimen ID
stain code	st065	a code assigned to the type of stain used in the slide
slide number	a001_lvl000	a multi-character slide number code (4 + "_" + 6 characters)
token number	t002	a four-character number indicated the slide number in a series
extension	svs	indicates the type of file; ".svs" is used for an image data file

Table 6. A unified classification scheme for FCDP and TUDP

Class	Label	Description / Features
Low Grade (lg)	Normal (norm)	normal ducts and lobules
	Background (bckg)	stroma, no ducts or lobules
	Artifact (artf)	grease pen marks, stitches, foreign bodies, etc.
Intermediate Grade (ig)	Non-Neoplastic (nneo)	fibrosis, hyperplasia, intraductal papilloma, adenosis, ectasia, etc.
	Inflammation (infl)	areas of inflammation
	Suspected (susp)	regions that are at risk of developing into cancerous regions
High Grade (hg)	Ductal Carcinoma in Situ (dcis)	ductal carcinoma in situ, and lobular carcinoma in situ
	Invasive Ductal Carcinoma (indc)	invasive ductal carcinoma, invasive lobular carcinoma, and invasive mammary carcinoma
unknown	Indistinguishable (null)	indistinguishable tissue, normally due to issues with the cut/stain

The labels “normal,” “background,” and “artifact” fall under the “lg” category. The labels “non-neoplastic,” “inflammation,” and “suspected” are grouped under the “ig” category. The labels “ductal carcinoma in situ” and “invasive ductal carcinoma in situ” are categorized as “hg”, while “indistinguishable” is placed under the "none" category. This rudimentary classification should help researchers select subsets of the data and conduct experiments using both corpora.

The TUDP Breast Tissue subset can be combined with similar FCDP slides to support experiments on cancer classification. The entire TUDP Corpus in its current form can be used to study unsupervised learning. We anticipate we will annotate portions of the new corpus as funding permits.

C. SUMMARY

This abstract introduces the release of two substantial resources pertinent to digital pathology, all of which are publicly available. These resources together contain over 116,000 high resolution images, making it one of the largest publicly available resources. Those interested in learning more about these resources, including how to gain access, can find additional information at the following URL: *www.isip.piconepress.com/projects/nsf_dpath/html/downloads.shtml*. At this URL, there are instructions for how to sign up for access to the corpus, instructions on how to download the data, and software that demonstrates how to build machine learning systems based on the data.

Data labeling is a costly task. The lack of detailed reports or metadata for the bulk of the TUDP Corpus poses challenges for researchers attempting to use this data for basic science research. The lack of annotations also poses challenges for machine learning researchers. A major topic of future research on this data is the use of unsupervised learning techniques to automatically label the data. We are exploring ways to automatically detect stain and tissue type. We are also annotating the breast tissue subset of the FCP data to complement the TUDP Breast Tissue subset. There are approximately 1,400 breast tissue slides in FCDP.

In addition, we are exploring ways to use this data in unsupervised learning so that we can improve the performance of our baseline classification system. Our long-term goal is to develop a single system that can handle all types of tissues, stains, and pathologies. Given the breadth of these resources, this is an extremely challenging task, since performance degrades rapidly as the number of potential classes is increased.

ACKNOWLEDGEMENTS

This material is based upon work supported in part by the National Science Foundation under grants under grants nos. CNS-1726188 and 1925494, by the Temple University Office of the Vice President for

Research, and by the Temple University College of Engineering Summer Research Experience for Undergraduates program. 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 or Temple University.

REFERENCES

[1] J. Picone, S. McKenzie, and M. Desai, "Unencumbered Electroencephalogram (EEG) Corpora: Advancing Technology Using Big Data Resources," in International Neuroinformatics Coordinating Facility Neuromatics Conference (INCF), Stockholm, Sweden, 2022, p. 11. url: *https://isip.piconepress.com/publications/presentations_invited/2022/incf/tuh_eeg/*.

[2] B. Doshna, Z. Wevodau, N. Jhala, I. Akhtar, I. Obeid, and J. Picone, "The Temple University Digital Pathology Corpus: The Breast Tissue Subset," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (SPMB), I. Obeid, I. Selesnick, and J. Picone, Eds., Philadelphia, Pennsylvania, USA: IEEE, 2021, pp. 1–3. doi: *10.1109/SPMB52430.2021.9672275*.

[3] D. Connolly, "Fox Chase Cancer Center Biosample Repository." 2023. Accessed: Dec. 02, 2023. [Online]. url: *https://www.foxchase.org/research/facilities/genetic-research-facilities/biosample-repository-facility*.

[4] N. Shawki *et al.*, "The Temple University Digital Pathology Corpus," in *Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications*, I. Obeid, I. Selesnick, and J. Picone, Eds., 1st ed., New York City, New York, USA: Springer, 2020, pp. 67–104. doi: *10.1007/978-3-030-36844-9*.

[5] "Aperio Format," OpenSlide. 2018. Accessed: Feb. 17, 2023. [Online]. url: *https://openslide.org/formats/aperio/*.

[6] L. Biosystems, "Aperio ImageScope - Pathology Slide Viewing Software," Leica Biosystems. 2018. [Online]. url: *https://www.leicabiosystems.com/digital-pathology/manage/aperio-image scope/*.

[7] "ICD-10-CM, Official Guidelines for Coding and Reporting", Centers for Medicare & Medicaid Services (CMS), January 01, 2020, url: *https://www.hhs.gov/guidance/sites/default/files/hhs-guidance-documents/ICD-10-CM_Guidelines-FY2020_final.pdf*.

[8] T. S. Gurina and L. Simms, "Histology, Staining.," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2023, pp. 1–5. [Online]. url: *https://www.ncbi.nlm.nih.gov/books/NBK557663/*.

[9] C. D. South, M. Yearsley, E. Martin, M. Arnold, W. Frankel, and H. Hampel, "Immunohistochemistry staining for the mismatch repair proteins in the clinical care of patients with colorectal cancer," *Genetics in Medicine*, vol. 11, pp. 812–817, Nov. 2009, doi: *10.1097/GIM.0b013e3181b99b75*.

[10] S. Renshaw, "Immunohistochemistry and Immunocytochemistry," in *Immunohistochemistry and Immunocytochemistry*, 2017, pp. 35–102. doi: *10.1002/9781118717769.ch3*.

[11] National Cancer Institute, "Tumor Grade." Aug. 01, 2022. url: *https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-grade#how-tumor-grade-is-determined*.

[12] W. Bulten et al., "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge," *Nature Medicine*, vol. 28, no. 1, pp. 154–163, Jan. 2022, doi: *10.1038/s41591-021-01620-2*.

[13] D. F. Gleason, "Histologic grading of prostate cancer: a perspective.," *Human Pathology*, vol. 23, no. 3, pp. 273–279, Mar. 1992. doi: *10.1016/0046-8177(92)90108-f*.

[14] D. F. Gleason, “The Veteran’s Administration Cooperative Urologic Research Group: histologic grading and clinical staging of prostatic carcinoma,” in *Urologic Pathology: The Prostate*, M. Tannenbaum, Ed., 1st ed., Philadelphia, Pennsylvania, USA: Lea and Febiger, 1977, pp. 171–198. url: *https://pubmed.ncbi.nlm.nih.gov/11905924/*.

[15] J. Simons, Z. Wevodau, B. Doshna, I. Obeid, and J. Picone, “The Temple University Hospital DPATH Corpus: Annotation Guidelines,” Temple University, Philadelphia, Pennsylvania, USA, Jan. 2021. url: *https://isip.piconepress.com/publications/reports/2021/tuh_dpath/annotations/*.

APPENDIX A: THE FCCC CORPUS METADATA FIELDS

The fields below are used in the metadata spreadsheet distributed with the data. The spreadsheet contains the metadata on a tab titled "fccc," and additional tabs showing all possible values of these fields.

Index	Col	Name	Description	Example
1	A	Formatted File ID	the basename of the filenames that contain the images and annotations.	000000197_001003366_st065_xt1_t000
2	B	Tissue Description	a concatenation of columns K and M	c15.9_c15.5
3	C	Stain Code	a code assigned using a list of standard stain types included with the corpus (e.g., "st065" = "H&E")	st065
4	D	Derived Grade	a classification of a tumor based on four levels (hg, ig, lg, unknown)	hg
5	E	Participant ID	an anonymized subject identifier.	697920
6	F	Specimen	a numeric identifier used to track the tissue sample	1024138
7	G	Stain	a 4-character code used to describe the type of stain used	0hne
8	H	Slide#	the slide number in a series of slides from the same specimen	T1
9	I	Tissue Collection Age	the age of the subject in calendar years at the time of collection.	77
10	J	Year Collected	the calendar year when the specimen was collected	2021
11	K	ICDO Code (Tissue Site)	the International Classification of Diseases for Oncology (ICDO) code for the tissue site	C15.9
12	L	Tissue Site.1	the anatomical site of the tissue site corresponding to the ICDO code for the tissue site	Esophagus, NOS
13	M	ICDO Code (Tumor Site)	the ICDO code for the tumor site	C15.5
14	N	Tumor Site	the anatomical site of the tumor corresponding to the ICDO code for the tumor site	Esophagus, lower third
15	O	NCDB Site Desc	a descriptive label or name given to the tumor site based on the information provided by the National Cancer Database (NCDB)	Esophagus
16	P	NCDB Site Category Desc	the category or group to which the tumor site belongs, based on the information provided by the National Cancer Database (NCDB)	Digestive System
17	Q	Laterality	the side (left or right) where the tumor originated	Not A Paired Organ
18	R	Year of Initial Diagnosis	the year when the patient was first diagnosed with the disease or condition under consideration (e.g., cancer)	2020
19	S	Age at Diagnosis	the age of the patient in calendar years at the time of their initial diagnosis	76
20	T	Meta To Site ICDO	the ICDO code for the metastatic site, which is the secondary location where cancer has spread from its primary site	C74.9

Index	Col	Name	Description	Example
21	U	Meta To Site	the specific anatomical site where the cancer has metastasized or spread to from its primary location	Adrenal gland, NOS
22	V	Year of Birth	year of birth for the subject	1944
23	W	Year of Death	year of death for the subject	2021
24	X	Gender	gender of the subject (male/female/other/transexual)	Male
25	Y	Race	race (e.g., Asian, Black, White)	White
26	Z	NCI Race	National Cancer Institute (NCI) classification for race (Asian, White, Unknown)	Unknown
27	AA	Ethnicity	ethnicity (e.g., Spanish, Puerto Rican, Hispanic, Unknown)	Unknown Whether Span
28	AB	NCI Ethnicity	NCI classification of ethnicity (Cuban, Hispanic, Non Hispanic, Unknown)	Unknown
29	AC	Histology	description of the cellular composition and characteristics of the tissue	Ductal Carcinoma NOS
30	AD	Behavior	tumor behavior (e.g., CA IN SITU)	MALIG-PRIMARY
31	AE	ICDO (Histology)	ICDO code for histology	8500/3
32	AF	Block Level Tissue Histology	description of the block level tissue histology based on an examination and analysis of the tissue sample	Ductal Carcinoma NOS
33	AG	Gradeclin Desc	classification of the tumor grade based on clinical evaluation; indicates the degree of abnormality or aggressiveness of cancer cells compared to normal cells	Grade III Poorly Differentiated
34	AH	Gradepath Desc	description or classification of the tumor grade based on the degree of differentiation of the tumor cells and is a measure of how abnormal the cancer cells look under the microscope	Grade cannot be assessed (GX); Unknown
35	AI	General Stage	overall or general classification or staging of cancer based on the extent of the disease and its spread	Regional To Lymph Nodes
36	AJ	Clinical T Stage	classification of the size and extent of the primary tumor at the time of diagnosis, as determined by clinical examination and imaging tests	cX
37	AK	Clinical N Stage	the number of nearby lymph nodes that have cancer	cX
38	AL	Clinical M Stage	refers to whether the cancer has metastasized, meaning that the cancer has spread from the primary tumor to other parts of the body	cX
39	AM	Clinical S Stage	classification of the elevation of the serum tumor Markers (alpha-fetoprotein (AFP), beta human chorionic gonadotropin (beta-hCG), and lactate dehydrogenase (LDH))	99
40	AN	Pathologic T Stage	classification or staging of cancer based on the size and extent of the primary tumor, as determined by examination of the tumor tissue during pathological analysis	p2

Index	Col	Name	Description	Example
41	AO	Pathologic N Stage	classification or staging of cancer based on the involvement or spread of cancer to the regional lymph nodes, as determined by examination of the lymph nodes during pathological analysis	p1
42	AP	Pathologic M Stage	classification or staging of cancer based on the presence or absence of distant metastasis (spread of cancer to other organs or distant sites), as determined by examination of the tumor tissue during pathological analysis	p0
43	AQ	Pathologic Stage Group	classification or staging of cancer based on the elevation of the serum tumor Markers	2B
44	AR	Months First Recurrence	the duration of time in months between the initial diagnosis of cancer and the detection or occurrence of the first recurrence of cancer after the completion of treatment	022
45	AS	Cancer Status	the current condition or status of the individual with regards to cancer (it may indicate whether the individual is currently undergoing treatment, in remission, or experiencing disease progression)	Evidence of this cancer
46	AT	Tobacco History	Patient history of smoking	Unknown
47	AU	Alcohol History	Patient history of alcohol use	Alcohol Usage Unknown
48	AV	Family History	is there evidence of cancer in the family history	Unknown

APPENDIX B: A SUMMARY OF THE TISSUE SITES IN FCDP

Index	ICDO Code (Tissue Site)	Tissue Site.1	Freq	%
1	C00.9	Lip, NOS	7	0.05%
2	C01.9	Base of Tongue, NOS	3	0.02%
3	C02.1	Tongue, border	5	0.04%
4	C02.2	Tongue, ventral surface, NOS	2	0.01%
5	C02.3	Tongue, ant 2/3, NOS	7	0.05%
6	C02.9	Tongue, NOS	74	0.52%
7	C03.0	Gum, upper	4	0.03%
8	C03.1	Gum, lower	8	0.06%
9	C03.9	Gum, NOS	3	0.02%
10	C04.0	Floor of mouth, anterior	3	0.02%
11	C04.9	Floor of mouth, NOS	5	0.04%
12	C05.0	Palate, hard	1	0.01%
13	C05.1	Palate, soft, NOS	3	0.02%
14	C05.9	Palate, NOS	1	0.01%
15	C06.0	Cheek mucosa	2	0.01%
16	C06.2	Retromolar area	10	0.07%
17	C06.9	Malignant Neoplasm of Mouth	10	0.07%
18	C07.9	Parotid gland	69	0.48%
19	C08.0	Submandibular gland	10	0.07%
20	C08.1	Sublingual gland	2	0.01%
21	C08.9	Salivary gland, major, NOS	5	0.04%
22	C09.8	Tonsil, overlapping lesion	1	0.01%
23	C09.9	Tonsil, NOS	33	0.23%
24	C10.9	Oropharynx, NOS	1	0.01%
25	C11.1	Nasopharynx, posterior wall	1	0.01%
26	C11.9	Nasopharynx, NOS	1	0.01%
27	C12.9	Pyriform sinus	6	0.04%
28	C13.0	Postcricoid region	1	0.01%
29	C14.0	Pharynx, NOS	5	0.04%
30	C15.2	Esophagus, abdominal	1	0.01%
31	C15.5	Esophagus, lower third	6	0.04%
32	C15.9	Esophagus, NOS	64	0.45%
33	C16.0	Cardia, NOS	8	0.06%
34	C16.1	Stomach, fundus	1	0.01%
35	C16.2	Stomach, body	6	0.04%
36	C16.3	Gastric antrum	15	0.11%
37	C16.4	Pylorus	3	0.02%
38	C16.5	Stomach, lesser curvature, NOS	3	0.02%
39	C16.6	Stomach, greater curvature, NOS	3	0.02%
40	C16.8	Stomach, overlapping lesion	5	0.04%
41	C16.9	Stomach, NOS	185	1.30%
42	C17.0	Duodenum	16	0.11%
43	C17.1	Jejunum	3	0.02%

Index	ICDO Code (Tissue Site)	Tissue Site.1	Freq	%
44	C17.2	Ileum	8	0.06%
45	C17.9	Small intestine, NOS	91	0.64%
46	C18.0	Cecum	76	0.53%
47	C18.1	Appendix	10	0.07%
48	C18.2	Colon, ascending	41	0.29%
49	C18.3	Colon, hepatic flexure	11	0.08%
50	C18.4	Colon, transverse	28	0.20%
51	C18.5	Colon, splenic flexure	6	0.04%
52	C18.6	Colon, descending	18	0.13%
53	C18.7	Colon, sigmoid	134	0.94%
54	C18.8	Colon, overlapping lesion	2	0.01%
55	C18.9	Colon, NOS	889	6.22%
56	C19.9	Rectosigmoid junction	52	0.36%
57	C20.9	Rectum, NOS	146	1.02%
58	C21.0	Anus, NOS	5	0.04%
59	C21.8	Rectum, anus and anal canal, overlapping lesion	1	0.01%
60	C22.0	Liver	141	0.99%
61	C22.1	Intrahepatic bile duct	4	0.03%
62	C23.9	Gallbladder	12	0.08%
63	C24.0	Extrahepatic bile duct	2	0.01%
64	C24.1	Ampulla of vater	5	0.04%
65	C24.9	Biliary tract, NOS	1	0.01%
66	C25.0	Pancreas, head	23	0.16%
67	C25.1	Pancreas, body	1	0.01%
68	C25.2	Pancreas, tail	6	0.04%
69	C25.3	Pancreatic duct	1	0.01%
70	C25.4	Islets of langerhans	2	0.01%
71	C25.8	Pancreas, overlapping lesion	1	0.01%
72	C25.9	Pancreas, NOS	219	1.53%
73	C26.9	Gastrointestinal tract, NOS	10	0.07%
74	C30.0	Nasal cavity	3	0.02%
75	C31.0	Maxillary, sinus	15	0.11%
76	C31.1	Sinus, ethmoid	1	0.01%
77	C31.2	Sinus, frontal	1	0.01%
78	C31.9	Accessory sinus, NOS	2	0.01%
79	C32.0	Glottis	7	0.05%
80	C32.1	Supraglottis	13	0.09%
81	C32.8	Larynx, overlapping lesion	2	0.01%
82	C32.9	Larynx, NOS	24	0.17%
83	C33.9	Trachea	3	0.02%
84	C34.0	Main bronchus	4	0.03%
85	C34.1	Lung, upper lobe	496	3.47%
86	C34.2	Lung, middle lobe	42	0.29%
87	C34.3	Lung, lower lobe	259	1.81%
88	C34.8	Lung, overlapping lesion	9	0.06%

Index	ICDO Code (Tissue Site)	Tissue Site.1	Freq	%
89	C34.9	Lung, NOS	683	4.78%
90	C37.9	Thymus	26	0.18%
91	C38.1	Mediastinum, anterior	3	0.02%
92	C38.2	Mediastinum, posterior	1	0.01%
93	C38.3	Mediastinum, NOS	8	0.06%
94	C38.4	Pleura, NOS	24	0.17%
95	C40.0	Upper limb, scapula & associated joints, long bones	5	0.04%
96	C40.2	Lower limb and associated joints, long bones	8	0.06%
97	C40.3	Lower limb and associated joints, short bones	1	0.01%
98	C40.8	Overlapping lesion of bones, joints and articular cartilage of limbs	1	0.01%
99	C40.9	Bone of limb, NOS	5	0.04%
100	C41.1	Mandible	27	0.19%
101	C41.3	Rib, Sternum, Clavicle and Associated Joints	5	0.04%
102	C41.4	Pelvic bones, sacrum, coccyx and associated joints	8	0.06%
103	C41.9	Bone, NOS	5	0.04%
104	C42.1	Bone marrow	2	0.01%
105	C42.2	Spleen	34	0.24%
106	C44.2	Skin, Unspecified malignant Neoplasm of skin of ear and external auricular canal	4	0.03%
107	C44.3	Skin of other and unspecified parts of face	28	0.20%
108	C44.4	Skin of scalp and neck	41	0.29%
109	C44.5	Skin of trunk	35	0.25%
110	C44.6	Skin of upper limb and shoulder	37	0.26%
111	C44.7	Skin of lower limb and hip	36	0.25%
112	C44.9	Skin, NOS	126	0.88%
113	C48.0	Retroperitoneum	94	0.66%
114	C48.1	Peritoneum, specified parts	88	0.62%
115	C48.2	Peritoneum, NOS	17	0.12%
116	C48.8	Retroperitoneum and peritoneum, overlapping lesion	2	0.01%
117	C49.0	Connective, subq, other soft tiss; head, face, neck	194	1.36%
118	C49.1	Connective, subq, other soft tiss; upper limb, shoulder	33	0.23%
119	C49.2	Connective, subq, other soft tiss; lower limb, hip	115	0.81%
120	C49.3	Connective, subq, other soft tiss; thorax	36	0.25%
121	C49.4	Connective, subq, other soft tiss; abdomen	66	0.46%
122	C49.5	Connective, subq, other soft tiss; pelvis	42	0.29%
123	C49.6	Connective, subq, other soft tiss; trunk	5	0.04%
124	C49.8	Connective, subq, other soft tiss; overlapping lesion	1	0.01%
125	C49.9	Connective, subq, other soft tiss; NOS	65	0.46%
126	C50.0	Nipple	3	0.02%
127	C50.1	Breast, central portion	27	0.19%
128	C50.2	Breast, upper-inner quadrant	56	0.39%
129	C50.3	Breast, lower-inner quadrant	27	0.19%
130	C50.4	Breast, upper-outer quadrant	156	1.09%
131	C50.5	Breast, lower-outer quadrant	37	0.26%
132	C50.6	Breast, axillary tail	1	0.01%
133	C50.8	Breast, overlapping lesion	95	0.67%

Index	ICDO Code (Tissue Site)	Tissue Site.1	Freq	%
134	C50.9	Breast, NOS	1063	7.44%
135	C51.0	Labium majus	1	0.01%
136	C51.9	Vulva, NOS	34	0.24%
137	C52.9	Vagina, NOS	9	0.06%
138	C53.0	Endocervix	10	0.07%
139	C53.1	Exocervix	2	0.01%
140	C53.9	Cervix uteri	40	0.28%
141	C54.1	Endometrium	725	5.07%
142	C54.2	Myometrium	3	0.02%
143	C54.9	Corpus uteri	1	0.01%
144	C55.9	Uterus, NOS	244	1.71%
145	C56.9	Ovary	780	5.46%
146	C57.0	Fallopian tube	20	0.14%
147	C57.4	Uterine adnexa	1	0.01%
148	C60.0	Prepuce	1	0.01%
149	C60.1	Glans penis	1	0.01%
150	C60.2	Penis, body	1	0.01%
151	C60.9	Penis, NOS	18	0.13%
152	C61.9	Prostate gland	1819	12.73%
153	C62.0	Testis, undescended	1	0.01%
154	C62.1	Testis, descended	6	0.04%
155	C62.9	Testis, NOS	122	0.85%
156	C63.0	Epididymis	3	0.02%
157	C63.1	Spermatic cord	2	0.01%
158	C63.2	Scrotum, NOS	3	0.02%
159	C63.7	Genital organs, male, other specified parts	4	0.03%
160	C64.9	Kidney, NOS	2285	15.99%
161	C65.9	Renal pelvis	19	0.13%
162	C66.9	Ureter	69	0.48%
163	C67.0	Trigone of bladder	4	0.03%
164	C67.1	Dome of bladder	2	0.01%
165	C67.2	Bladder, lateral wall	3	0.02%
166	C67.3	Bladder, anterior wall	1	0.01%
167	C67.4	Bladder, posterior wall	4	0.03%
168	C67.5	Bladder neck	1	0.01%
169	C67.8	Bladder, overlapping lesion	9	0.06%
170	C67.9	Bladder, NOS	375	2.62%
171	C68.0	Urethra	10	0.07%
172	C69.9	Eye, NOS	1	0.01%
173	C73.9	Thyroid gland	321	2.25%
174	C74.9	Adrenal gland, NOS	22	0.15%
175	C75.0	Parathyroid gland	1	0.01%
176	C76.0	Head, face, or neck, NOS	44	0.31%
177	C76.1	Thorax, NOS	9	0.06%
178	C76.2	Abdomen, NOS	5	0.04%

Index	ICDO Code (Tissue Site)	Tissue Site.1	Freq	%
179	C76.3	Pelvis, NOS	8	0.06%
180	C76.4	Upper limb, NOS	4	0.03%
181	C76.5	Lower limb, NOS	3	0.02%
182	C76.7	Ill-defined sites, other	1	0.01%
183	C77.0	Lymph nodes of head, face and neck	66	0.46%
184	C77.1	Lymph nodes, intrathoracic	5	0.04%
185	C77.2	Lymph nodes, intra-abdominal	5	0.04%
186	C77.3	Lymph nodes of axilla or arm	45	0.32%
187	C77.4	Lymph nodes of inguinal region or leg	63	0.44%
188	C77.5	Lymph nodes, pelvic	7	0.05%
189	C77.8	Lymph nodes of multiple regions	24	0.17%
190	C77.9	Lymph node, NOS	58	0.41%
191	C80.9	Unknown primary site	23	0.16%
192	Cxx.x	unknown	2	0.01%

APPENDIX C: A Summary of the Stain Types Used in Our Digital Pathology Corpora

The names for these stains came from these sources:

- *National Society for Histology*: https://www.nsh.org/blogs/natalie-paskoski/2021/05/04/the-alcian-blue-stain-for-histology
- *NoRDx MainHealth*: https://nordx.testcatalog.org/show/Histology-Stains
- *National Society of Health*: https://www.nih.gov/
- *National Library of Medicine*: https://www.nlm.nih.gov/
- *Pathology Outlines*: Stains & CD Markers: https://www.pathologyoutlines.com/stains.html
- *Wikipedia*: https://en.wikipedia.org/wiki/Hematein

The stain label NEDC refers to codes assigned by the authors. The label "Unknown" refers to a label assigned when the slide documentation had no stain information.

The codes "stxxx" are numeric codes assigned by the authors to facilitate data management.

Index	Code	Stain	Source
1	st001	Actin-MSA	NoRDx MainHealth
2	st002	AFB (acid-fast bacillus) stain	National Library of Medicine
3	st003	AFB, Aur-Rhod. (mycobacterium)	National Library of Medicine
4	st004	AFB, Fites (m leprae)	National Library of Medicine
5	st005	AFB, Putts (mycobacterium)	National Library of Medicine
6	st006	AFP	NoRDx MainHealth
7	st007	Alcian blue	National Society for Histology
8	st008	Alcian Blue, ph 2.5	National Library of Medicine
9	st009	Alk-1	NoRDx MainHealth
10	st010	Arginase-1	NoRDx MainHealth
11	st011	B72.3*	NoRDx MainHealth
12	st012	Bcl-2*	NoRDx MainHealth
13	st013	Bcl-6*	NoRDx MainHealth
14	st014	Ber-EP4*	NoRDx MainHealth
15	st015	Beta-Catenin	NoRDx MainHealth
16	st016	BRAF V600E	NoRDx MainHealth
17	st017	BREAST Triple	NoRDx MainHealth
18	st018	Brown and Hopps	National Library of Medicine
19	st019	C-MYC	NoRDx MainHealth
20	st020	CAIX	NoRDx MainHealth
21	st021	Calcitonin	NoRDx MainHealth
22	st022	Caldesmon	NoRDx MainHealth
23	st023	Calponin	NoRDx MainHealth
24	st024	Calretinin	NoRDx MainHealth
25	st025	Carmine	National Library of Medicine
26	st026	CD 10*	NoRDx MainHealth
27	st027	CD 117 (c-kit)	NoRDx MainHealth
28	st028	CD 138	NoRDx MainHealth
29	st029	CD 15 (LeuM1)	NoRDx MainHealth

30	st030	CD 163	NoRDx MainHealth
31	st031	CD 20 (L26)*	NoRDx MainHealth
32	st032	CD 23*	NoRDx MainHealth
33	st033	Cd 3 (Poly)*	NoRDx MainHealth
34	st034	CD 30(Ki-1)	NoRDx MainHealth
35	st035	CD 31	NoRDx MainHealth
36	st036	CD 34 (QBEnd)	NoRDx MainHealth
37	st037	CD 4	NoRDx MainHealth
38	st038	CD 43 (MT-1)	NoRDx MainHealth
39	st039	CD 44*	NoRDx MainHealth
40	st040	CD 45 (LCA)*	NoRDx MainHealth
41	st041	CD 5*	NoRDx MainHealth
42	st042	CD 56	NoRDx MainHealth
43	st043	CD 61	NoRDx MainHealth
44	st044	CD 68 (Marcophage)	NoRDx MainHealth
45	st045	CD 79a*	NoRDx MainHealth
46	st046	CD 8	NoRDx MainHealth
47	st047	CD 99 ("HBA.71")/MIC2	National Library of Medicine
48	st048	CDX2	NoRDx MainHealth
49	st049	CEA*	NoRDx MainHealth
50	st050	Chromogranin*	NoRDx MainHealth
51	st051	CMV	NoRDx MainHealth
52	st052	Colloidal iron	National Library of Medicine
53	st053	Congo Red	National Library of Medicine
54	st054	Cresyl Echt Violet (neuro, glia)	National Library of Medicine
55	st055	Crystal Violet, (amyloid)	National Library of Medicine
56	st056	Cyclin D1	NoRDx MainHealth
57	st057	CytoKeratin 19*	NoRDx MainHealth
58	st058	CytoKeratin 20*	NoRDx MainHealth
59	st059	CytoKeratin 5/6*	NoRDx MainHealth
60	st060	CytoKeratin 7*	NoRDx MainHealth
61	st061	CytoKeratin 8&18(CAM5.2)*	NoRDx MainHealth
62	st062	CytoKeratin 903*	NoRDx MainHealth
63	st063	CytoKeratin AE1/AE3*	NoRDx MainHealth
64	st064	D2-40	NoRDx MainHealth
65	st065	Desmin	NoRDx MainHealth
66	st066	DOG-1	NoRDx MainHealth
67	st067	E-cadherin	NoRDx MainHealth
68	st068	EMA*	NoRDx MainHealth
69	st069	ER*	NoRDx MainHealth
70	st070	ERG	NoRDx MainHealth
71	st071	Factor 13a	NoRDx MainHealth
72	st072	Fontana-Masson (Melanin)	National Library of Medicine
73	st073	GATA3	NoRDx MainHealth
74	st074	GCDFP-15/BRST-2	National Library of Medicine
75	st075	GFAP*	NoRDx MainHealth
76	st076	Giemsa (bacteria)	National Library of Medicine

77	st077	Giemsa (mast cells)	National Library of Medicine
78	st078	Giemsa Stain/Romanowsky	National Library of Medicine
79	st079	Glypican-3	NoRDx MainHealth
80	st080	Gomori methenamine silver stain (GMS)	National Library of Medicine
81	st081	Gram (bacteria)	National Library of Medicine
82	st082	Gram Stain	National Library of Medicine
83	st083	Grocott Meth Silver (fungus)	National Library of Medicine
84	st084	Grocott Meth Silver (pneumocystis)	National Library of Medicine
85	st085	Grocott Meth Silver (urate crystals)	National Library of Medicine
86	st086	H Pylori	NoRDx MainHealth
87	st087	H&E	National Library of Medicine
88	st088	HBME-1*	NoRDx MainHealth
89	st089	Hematin and Hematoxylin	Wikipedia
90	st090	HepPar-1	NoRDx MainHealth
91	st091	Her 2 Neu	NoRDx MainHealth
92	st092	HMB-45*	NoRDx MainHealth
93	st093	HSV 1&2	NoRDx MainHealth
94	st094	IgG	NoRDx MainHealth
95	st095	IgG4	NoRDx MainHealth
96	st096	Immunohistochemistry staining (IHC)	National Library of Medicine
97	st097	Inhibin	NoRDx MainHealth
98	st098	Iron Stain	National Library of Medicine
99	st099	Gomori's Iron	National Library of Medicine
100	st100	Jones (basement membrane)	National Library of Medicine
101	st101	Kappa*	NoRDx MainHealth
102	st102	Ki-67 / Mib-1*	NoRDx MainHealth
103	st103	Ki67 / Mart1 Dual	NoRDx MainHealth
104	st104	Lambda	NoRDx MainHealth
105	st105	Luxol Fast Blue (myelin)	National Library of Medicine
106	st106	Mammaglobin	NoRDx MainHealth
107	st107	MART-1*	NoRDx MainHealth
108	st108	Masson’s Trichrome	National Library of Medicine
109	st109	Melanin Bleach	National Library of Medicine
110	st110	MLH-1	NoRDx MainHealth
111	st111	MOC-31	NoRDx MainHealth
112	st112	MSH-2	NoRDx MainHealth
113	st113	MSH-6	NoRDx MainHealth
114	st114	Mucicarmine	National Library of Medicine
115	st115	Mucicarmine, (crypto)	National Library of Medicine
116	st116	Mucicarmine, Mayer	National Library of Medicine
117	st117	Mum 1	NoRDx MainHealth
118	st118	Myeloperoxidase (mpo)	NoRDx MainHealth
119	st119	MyoD1	NoRDx MainHealth
120	st120	Myogenin (myf4)	NoRDx MainHealth
121	st121	Myoglobin	NoRDx MainHealth
122	st122	Napsin-A*	NoRDx MainHealth
123	st123	NEG	NoRDx MainHealth

124	st124	Nissl Stain	National Library of Medicine
125	st125	NKX3.1	NoRDx MainHealth
126	st126	Oct-4	NoRDx MainHealth
127	st127	Oil Red O	National Library of Medicine
128	st128	Oil Red O (fat)	National Library of Medicine
129	st129	P 16*	NoRDx MainHealth
130	st130	P 40	NoRDx MainHealth
131	st131	P 53*	NoRDx MainHealth
132	st132	P 57	NoRDx MainHealth
133	st133	P 63*	NoRDx MainHealth
134	st134	P120	NoRDx MainHealth
135	st135	P504's (AMACAR)	NoRDx MainHealth
136	st136	Papanicolaou Stain	National Library of Medicine
137	st137	PAS (basement membrane)	National Library of Medicine
138	st138	PAS (diastase digestion)	National Library of Medicine
139	st139	PAS (periodic acid-Schiff)	National Library of Medicine
140	st140	PAS, McManus (fungus)	National Library of Medicine
141	st141	PAX-5	NoRDx MainHealth
142	st142	PAX-8	NoRDx MainHealth
143	st143	PHH3	NoRDx MainHealth
144	st144	PIN4	NoRDx MainHealth
145	st145	PMS-2	NoRDx MainHealth
146	st146	PR*	NoRDx MainHealth
147	st147	Prussian Blue	National Library of Medicine
148	st148	PSA*	NoRDx MainHealth
149	st149	Reticulin Stain	National Library of Medicine
150	st150	Reticulin, Gomori	National Library of Medicine
151	st151	S-100*	NoRDx MainHealth
152	st152	SALL4*	NoRDx MainHealth
153	st153	SATB2	NoRDx MainHealth
154	st154	Silver	National Library of Medicine
155	st155	SMMHC	NoRDx MainHealth
156	st156	SOX10	NoRDx MainHealth
157	st157	STAT6	NoRDx MainHealth
158	st158	Sudan Black	National Library of Medicine
159	st159	Synaptophysin	NoRDx MainHealth
160	st160	TDT	NoRDx MainHealth
161	st161	Thyroglobulin*	NoRDx MainHealth
162	st162	TTF-1*	NoRDx MainHealth
163	st163	Verhoeff-Van Gieson	National Library of Medicine
164	st164	Verhoff's Elastic Stain	National Library of Medicine
165	st165	Vimentin	NoRDx MainHealth
166	st166	Von Kossa (calcium)	National Library of Medicine
167	st167	Warthin-Starry (legionella)	National Library of Medicine
168	st168	WT-1	NoRDx MainHealth
169	st169	Copper	Pathology Outlines - Stains & CD Markers
170	st170	BCOR	Pathology Outlines - Stains & CD Markers

171	st171	Brachyury	Pathology Outlines - Stains & CD Markers
172	st172	CD1a	Pathology Outlines - Stains & CD Markers
173	st173	CIC-DUX4	Pathology Outlines - Stains & CD Markers
174	st174	FLI1	Pathology Outlines - Stains & CD Markers
175	st175	GRM1	Pathology Outlines - Stains & CD Markers
176	st176	H3K36M	Pathology Outlines - Stains & CD Markers
177	st177	H3G34W	Pathology Outlines - Stains & CD Markers
178	st178	MDM2	Pathology Outlines - Stains & CD Markers
179	st179	NKX2.2	Pathology Outlines - Stains & CD Markers
180	st180	SOX9	Pathology Outlines - Stains & CD Markers
181	st181	CAMTA1	Pathology Outlines - Stains & CD Markers
182	st182	CDK4	Pathology Outlines - Stains & CD Markers
183	st183	DDIT3	Pathology Outlines - Stains & CD Markers
184	st184	EBER	Pathology Outlines - Stains & CD Markers
185	st185	Factor XIIIa	Pathology Outlines - Stains & CD Markers
186	st186	FOSB	Pathology Outlines - Stains & CD Markers
187	st187	GLUT1	Pathology Outlines - Stains & CD Markers
188	st188	H3K27me3	Pathology Outlines - Stains & CD Markers
189	st189	HHV8/LANA1	Pathology Outlines - Stains & CD Markers
190	st190	MelanA	Pathology Outlines - Stains & CD Markers
191	st191	MUC-4	Pathology Outlines - Stains & CD Markers
192	st192	NR4A-3	Pathology Outlines - Stains & CD Markers
193	st193	pan-TRK	Pathology Outlines - Stains & CD Markers
194	st194	PROX1	Pathology Outlines - Stains & CD Markers
195	st195	RB-1	Pathology Outlines - Stains & CD Markers
196	st196	SDHB	Pathology Outlines - Stains & CD Markers
197	st197	SMARCA4/BRG1	Pathology Outlines - Stains & CD Markers
198	st198	SMARCB1/IN1	Pathology Outlines - Stains & CD Markers
199	st199	SMA	Pathology Outlines - Stains & CD Markers
200	st200	SYT/SSX	Pathology Outlines - Stains & CD Markers
201	st201	TFE-3	Pathology Outlines - Stains & CD Markers
202	st202	TLE-1	Pathology Outlines - Stains & CD Markers
203	st203	androgen receptor	Pathology Outlines - Stains & CD Markers
204	st204	PHLDA1 (trichoepithelioma versus BBC)	Pathology Outlines - Stains & CD Markers
205	st205	BAP-1	Pathology Outlines - Stains & CD Markers
206	st206	MLH-3	Pathology Outlines - Stains & CD Markers
207	st207	treponema IHC	Pathology Outlines - Stains & CD Markers
208	st208	DPC-4/SMADA-4	Pathology Outlines - Stains & CD Markers
209	st209	PDL-1	Pathology Outlines - Stains & CD Markers
210	st210	ARID1A	Pathology Outlines - Stains & CD Markers
211	st211	FOXL2	Pathology Outlines - Stains & CD Markers
212	st212	hCG	Pathology Outlines - Stains & CD Markers
213	st213	HNF-1B	Pathology Outlines - Stains & CD Markers
214	st214	OCT 3/4	Pathology Outlines - Stains & CD Markers
215	st215	SF-1	Pathology Outlines - Stains & CD Markers
216	st216	PLAP	Pathology Outlines - Stains & CD Markers
217	st217	smooth muscle actin	Pathology Outlines - Stains & CD Markers

218	st218	RCC (Renal cell carcinoma)	Pathology Outlines - Stains & CD Markers
219	st219	AMACR	Pathology Outlines - Stains & CD Markers
220	st220	Prostein/ P501S	Pathology Outlines - Stains & CD Markers
221	st221	HIK1086	Pathology Outlines - Stains & CD Markers
222	st222	NUT/NUTM-1	Pathology Outlines - Stains & CD Markers
223	st223	PTEN	Pathology Outlines - Stains & CD Markers
224	st224	PLAG-1	Pathology Outlines - Stains & CD Markers
225	st225	Galectin-3	Pathology Outlines - Stains & CD Markers
226	st226	IgG4	Pathology Outlines - Stains & CD Markers
227	st227	Prealbumin	Pathology Outlines - Stains & CD Markers
228	st228	adenovirus	Pathology Outlines - Stains & CD Markers
229	st229	C4d	Pathology Outlines - Stains & CD Markers
230	st230	EGFR	Pathology Outlines - Stains & CD Markers
231	st231	DNAJB9 (FN1)	Pathology Outlines - Stains & CD Markers
232	st232	TROP2	Pathology Outlines - Stains & CD Markers
233	st233	2-succino-cysteine	Pathology Outlines - Stains & CD Markers
234	st234	ACTH	Pathology Outlines - Stains & CD Markers
235	st235	ATDX (ATRX)	Pathology Outlines - Stains & CD Markers
236	st236	NeuN	Pathology Outlines - Stains & CD Markers
237	st237	NFP / neurofilament	Pathology Outlines - Stains & CD Markers
238	st238	Olig-2	Pathology Outlines - Stains & CD Markers
239	st239	Pit-1	Pathology Outlines - Stains & CD Markers
240	st240	SSTR2A	Pathology Outlines - Stains & CD Markers
241	st241	Tpit	Pathology Outlines - Stains & CD Markers
242	st242	Alkaline phosphatase	Pathology Outlines - Stains & CD Markers
243	st243	COX (Cytochrome c oxidase)	Pathology Outlines - Stains & CD Markers
244	st244	SV-40	Pathology Outlines - Stains & CD Markers
245	st245	SDH	Pathology Outlines - Stains & CD Markers
246	st246	albumin	Pathology Outlines - Stains & CD Markers
247	st247	Alpha-1-antichymotrypsin	Pathology Outlines - Stains & CD Markers
248	st248	Alpha-1 antitrypsin (A1AT/AAT)	Pathology Outlines - Stains & CD Markers
249	st249	ATM	Pathology Outlines - Stains & CD Markers
250	st250	BCL-1	Pathology Outlines - Stains & CD Markers
251	st251	BCL-10	Pathology Outlines - Stains & CD Markers
252	st252	BOB-1	Pathology Outlines - Stains & CD Markers
253	st253	BSEP	Pathology Outlines - Stains & CD Markers
254	st254	CK 10	Pathology Outlines - Stains & CD Markers
255	st255	CK 14	Pathology Outlines - Stains & CD Markers
256	st256	CK 17	Pathology Outlines - Stains & CD Markers
257	st257	CK 18	Pathology Outlines - Stains & CD Markers
258	st258	CK 19	Pathology Outlines - Stains & CD Markers
259	st259	Cytokeratin 34 beta E12	Pathology Outlines - Stains & CD Markers
260	st260	IgA	Pathology Outlines - Stains & CD Markers
261	st261	IgD	Pathology Outlines - Stains & CD Markers
262	st262	IgM	Pathology Outlines - Stains & CD Markers
263	st263	insulin	Pathology Outlines - Stains & CD Markers
264	st264	orcein	Pathology Outlines - Stains & CD Markers

265	st265	L1-CAM	Pathology Outlines - Stains & CD Markers
266	st266	LANA	Pathology Outlines - Stains & CD Markers
267	st267	Leder	Pathology Outlines - Stains & CD Markers
268	st268	Leu-7	Pathology Outlines - Stains & CD Markers
269	st269	LFABP	Pathology Outlines - Stains & CD Markers
270	st270	Luxol fast blue	Pathology Outlines - Stains & CD Markers
271	st271	Lipochrome (lipofuscin) pigments	Pathology Outlines - Stains & CD Markers
272	st272	Lysozyme	Pathology Outlines - Stains & CD Markers
273	st273	Microphthalmia associated transcription factor (MITF)	Pathology Outlines - Stains & CD Markers
274	st274	MLH3	Pathology Outlines - Stains & CD Markers
275	st275	Movat pentachrome	Pathology Outlines - Stains & CD Markers
276	st276	Mucins	Pathology Outlines - Stains & CD Markers
277	st277	MYB	Pathology Outlines - Stains & CD Markers
278	st278	MYC	Pathology Outlines - Stains & CD Markers
279	st279	NB84	Pathology Outlines - Stains & CD Markers
280	st280	NCCT	Pathology Outlines - Stains & CD Markers
281	st281	p75/NGFR	Pathology Outlines - Stains & CD Markers
282	st282	NKX-2	Pathology Outlines - Stains & CD Markers
283	st283	Nonspecific esterase	Pathology Outlines - Stains & CD Markers
284	st284	NPM-1	Pathology Outlines - Stains & CD Markers
285	st285	p21	Pathology Outlines - Stains & CD Markers
286	st286	p62	Pathology Outlines - Stains & CD Markers
287	st287	Parvalbumin	Pathology Outlines - Stains & CD Markers
288	st288	PCNA	Pathology Outlines - Stains & CD Markers
289	st289	SP263	Pathology Outlines - Stains & CD Markers
290	st290	Perforin	Pathology Outlines - Stains & CD Markers
291	st291	PGP9.5	Pathology Outlines - Stains & CD Markers
292	st292	PHLDA1	Pathology Outlines - Stains & CD Markers
293	st293	PTH	Pathology Outlines - Stains & CD Markers
294	st294	ROS-1	Pathology Outlines - Stains & CD Markers
295	st295	Sirius red	Pathology Outlines - Stains & CD Markers
296	st296	Urates/ uric acid	Pathology Outlines - Stains & CD Markers
297	st297	Uroplakin II	Pathology Outlines - Stains & CD Markers
298	st298	Uroplakin III	Pathology Outlines - Stains & CD Markers
299	st299	von Hippel Lindau (VHL)	Pathology Outlines - Stains & CD Markers
300	st300	Villin	Pathology Outlines - Stains & CD Markers
301	st301	VIP	Pathology Outlines - Stains & CD Markers
302	st302	von Willebrand factor (vWF)	Pathology Outlines - Stains & CD Markers
303	st303	VZV (Varicella-Zoster Virus)	Pathology Outlines - Stains & CD Markers
304	st304	HBsAg/Hs Ag	Pathology Outlines - Stains & CD Markers
305	st305	HBcAg /HBC/Hbcore	Pathology Outlines - Stains & CD Markers
306	st306	CD 38	Pathology Outlines - Stains & CD Markers
307	st307	CD 57	Pathology Outlines - Stains & CD Markers
308	st308	CD 21/CR2	Pathology Outlines - Stains & CD Markers
309	st309	CD 2	Pathology Outlines - Stains & CD Markers
310	st310	Fibrinogen / Fibrin	National Library of Medicine
311	st311	Prolactin	National Library of Medicine

312	st312	TSH	Pathology Outlines - Stains & CD Markers
313	st314	CA 125	Pathology Outlines - Stains & CD Markers
314	st315	p 12	Pathology Outlines - Stains & CD Markers
315	st316	CD 22	Pathology Outlines - Stains & CD Markers
316	st317	HPV	Pathology Outlines - Stains & CD Markers
317	st318	BILE - HALL'S BILIRUBIN	Pathology Outlines - Stains & CD Markers
318	st319	NF1/NF2/NF	Pathology Outlines - Stains & CD Markers
319	st320	BerEP4 / EpCAM	Pathology Outlines - Stains & CD Markers
320	st321	TRAP (Tartrate resistant acid phosphatase)	Pathology Outlines - Stains & CD Markers
321	st322	TK-1	Pathology Outlines - Stains & CD Markers
322	stmul	multiple stains	NEDC
323	stxxx	Unknown	Unknown

Big Data Resources for Digital Pathology

S. S. Shalamzari, M. Bagritsevitch, A. Melles, I. Obeid, and J. Picone
The Neural Engineering Data Consortium
Temple University

D. Connolly, C. Wu, B. Schultz, B. Brown, J. James, Y. Gong, and H. Wu
Fox Chase Cancer Center
Temple University

NEURAL ENGINEERING DATA CONSORTIUM
www.nedcdata.org

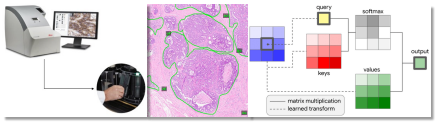
Fox Chase Cancer Center
Temple Health

Abstract

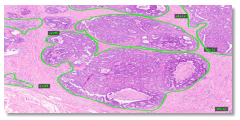
- The Neural Engineering Data Consortium (NEDC), has previously released the Breast Tissue subset of the Temple University Digital Pathology Corpus (TUDP), containing 3,505 partially annotated images.
- In this abstract, we introduce two recent additions to our digital pathology resources:
  - 14,288 images that were collected from Fox Chase Cancer Center's Biosample Repository
  - 97,755 images that are part of the TUDP Corpus
- The FCCC Corpus (FCDP) includes 48 metadata fields that provide information on the history of the sample, the patient's medical history and diagnosis codes.
- The TUDP Corpus includes a wide variety of tissue and stain types. The Breast Tissue subset has been reclassified to be consistent with FCDP.
- These two resources can be used to develop a new generation of machine learning technology that is more robust and can classify a wider range of data and pathology types.

Digital Pathology

- One of the unique challenges of digital pathology is the need to detect small regions of interest (e.g., 32x32 pixels) in extremely high resolution images (50Kx50K pixels):

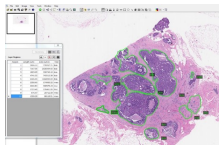


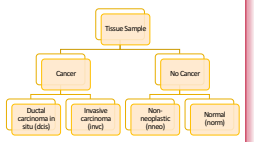
- Localization of information contributing to the diagnosis is a major goal of our work:



Labels		
norm	infl	indc
artf	susp	null
nneo	dcis	bckg

- The TUDP Breast Tissue Subset consists of 3,505 images annotated using a 4-way classification:





- There are 296 patients with 4.3% of the slides showing cancerous features.
- The average number of labels per image is 4.84.
- There are 8,895 cancerous labels, 5,362 labels with carcinogenic signs and 2,714 non-cancerous labels.

The FCCC Digital Pathology Corpus

- There are 14,288 images; the average file size is 244 MB; the entire corpus requires 3.5 TB of storage:

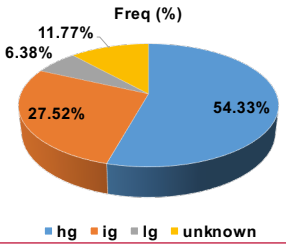
Component	Example	Description
directory name	fcdp/v1.0.0/svs/00026/00026256 4/001010285/c67.9_c67.2	a typical full directory specification
database	fcdp	an acronym for the corpus
version	v1.0.0	the version number
data type	svs	the root node for the image data and annotations
sequential ID	00026	a 5-digit number used to control directory complexity
subject ID	000262564	a 9-digit anonymized subject identifier (col. E)
specimen ID	001010285	a 9-digit specimen identifier (col. F)
ICDO codes	c67.9_c67.2	ICDO codes for tissue type (c67.9) and tissue site (c67.2) (cols. K and M respectively)
Component	**Example**	**Description**
filename	000368678_001013243_st065_xt01_t000.svs	a unique filename
identifier	000368678_001013243	subject ID (col. E) and specimen ID (col. F)
stain code	st065	a code assigned to the type of stain used in the slide (col. G)
slide number	xt1	a three-character code assigned to each slide (col. H)
token number	t000	a four-character number indicated the slide number in a series
extension	svs	indicates the type of file; ".svs" is used for an image data file

- Images are stored in Aperio's popular SVS format.
- There are 13,230 subjects in the corpus.
- Each image is accompanied by a csv file containing 48 metadata fields including diagnostic information:

Idx	Col	Name	Description	Example
11	K	ICDO Code (Tissue Site)	the International Classification of Diseases for Oncology (ICDO) code for the tissue site	C15.9
12	L	Tissue Site.1	the anatomical site of the tissue site corresponding to the ICDO code for the tissue site	Esophagus, NOS
29	AC	Histology	description of the cellular composition and characteristics of the tissue	Ductal Carcinoma NOS
32	AF	Block Level Tissue Histology	description of the block level tissue histology based on an examination and analysis of the tissue sample	Ductal Carcinoma NOS
33	AG	Gradeclin Desc	classification of the tumor grade based on clinical evaluation; indicates the degree of abnormality or aggressiveness of cancer cells compared to normal cells	Grade III Poorly Differentiated
34	AH	Gradepath Desc	description or classification of the tumor grade based on the degree of differentiation of the tumor cells; is a measure of how abnormal the cancer cells look under the microscope	Grade cannot be assessed (GX); Unknown

- The data has been classified according to a universal tumor grading system:

Grade Group	Gleason Score	Risk Group	Class
G1	≤ 6	low / very low	low (lg)
G2	7 (3+4)	intermediate	intermediate (ig)
G3	7 (4+3)		
G4	8	high / very high	high (hg)



The TU Digital Pathology Corpus

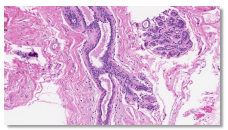
- There are 97,755 images; the average file size is 224 MB; the entire corpus requires 24 TB of storage:

Component	Example	Description
directory name	tudp/v1.0.0/svs/00036/aaaaaaaaa / bbbbbbbbbbbbbbb/c50.9_cxx.x	a typical full directory specification
database	tudp	an acronym for the corpus
version	v1.0.0	the version number
data type	svs	the root node for the image data and annotations
sequential ID	00036	a 5-digit number used to control directory complexity
subject ID	aaaaaaaaa	a 9-digit anonymized subject identifier (ID)
specimen ID	bbbbbbbbbbbbbbb	a 15-character specimen identifier (ID)
ICDO codes	c50.9_cxx.x	ICDO codes for tissue type (c50.9) and tissue site (cxx.x)
Component	**Example**	**Description**
filename	aaaaaaaaa_bbbbbbbbb_st065_a001_lvl000_t002.svs	a unique filename
identifier	aaaaaaaaa_bbbbbbbbbbbbbbb	subject ID (col. E) and specimen ID
stain code	st065	a code assigned to the type of stain used in the slide
slide number	a001_lvl000	a multi-character slide number code (4 + "_" + 6 characters)
token number	t002	a four-character number indicated the slide number in a series
extension	svs	indicates the type of file; ".svs" is used for an image data file

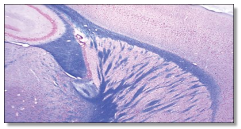
- A unified classification scheme is used to facilitate machine learning experiments using both corpora:

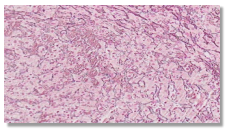
Class	Label	Description / Features
Low Grade (lg)	Normal (norm)	normal ducts and lobules
	Background (bckg)	stroma, no ducts or lobules
	Artifact (artf)	grease pen marks, stitches, foreign bodies, etc.
Intermediate Grade (ig)	Non-Neoplastic (nneo)	fibrosis, hyperplasia, intraductal papilloma, adenosis, ectasia, etc.
	Inflammation (infl)	areas of inflammation
	Suspected (susp)	regions that are at risk of developing into cancerous regions
High Grade (hg)	Ductal Carcinoma in Situ (dcis)	ductal carcinoma in situ, and lobular carcinoma in situ
	Invasive Ductal Carcinoma (indc)	invasive ductal carcinoma, invasive lobular carcinoma, and invasive mammary carcinoma
unknown	Indistinguishable (null)	indistinguishable tissue, normally due to issues with the cut/stain

- There are 192 unique tissue types and 324 unique stain types represented in our combined corpora.



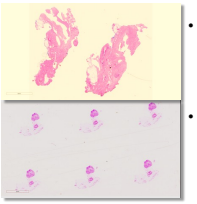
- Hematoxylin and eosin (H&E) stain is preferred for viewing cellular and tissue structure.
- Hematoxylin stains nuclei purplish blue; eosin stains the extracellular matrix pink. H&E demonstrates a broad range of features.
- Myelin is a layer that surrounds nerves. Damaged myelin slows down electrical impulses.
- Myelin cannot be detected with H&E stain. A solocherome cyanine stain is used. It stains cells and tissues blue/violet.



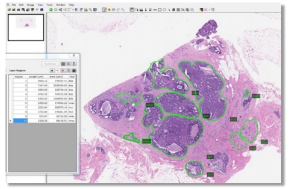


- Reticulin stain uses silver impregnation to detect reticulin fibers, which are made of type 3 collagen.
- It makes it easier to visualize areas of hepatocyte loss or regeneration.

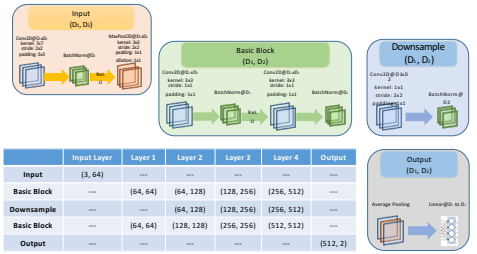
Segmentation and Localization



- Segmentation and localization of information is critical since the number of images and location of an image on a slide can vary.
- Localization requires scanning images at very high resolution, which poses computational and machine learning challenges.
- Annotation of an entire image is slow and expensive.
- We typically only annotate 5 to 10 regions of interest and a few examples of 'background'.



- We also provide a segmentation and classification system based on a ResNet-18 deep learning system:



	Input Layer	Layer 1	Layer 2	Layer 3	Layer 4	Output
Input	(3, 64)	---	---	---	---	---
Basic Block	---	(64, 64)	(64, 128)	(128, 256)	(256, 512)	---
Downsample	---	---	(64, 128)	(128, 256)	(256, 512)	---
Basic Block	---	(64, 64)	(128, 128)	(256, 256)	(512, 512)	---
Output	---	---	---	---	---	(512, 2)

Summary and Future Work

- This abstract introduces the release of two substantial resources pertinent to digital pathology, all of which are publicly available.
- We are currently in the process of annotating approximately 1,400 breast tissue images from FCDP to complement our TUDP release. These additional images will support a number of interesting studies including invariance and robustness.
- Future research is focusing on the use of unsupervised learning techniques to automatically classify and label the data. We are exploring ways to automatically detect stain and tissue type.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under grants nos. CNS-1726188 and 1925494, by the Temple University Office of the Vice President for Research, and by the Temple University College of Engineering Summer Research Experience for Undergraduates program. 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 or Temple University.

To learn more about our resources, please use this URL:
https://isip.piconepress.com/projects/nsf_dpath

Big Data Resources for Digital Pathology

S. S. Shalamzari, M. Bagritsevitch, A. Melles, I. Obeid, and J. Picone
The Neural Engineering Data Consortium
Temple University

D. Connolly, C. Wu, B. Schultz, B. Brown, J. James, Y. Gong, and H. Wu
Fox Chase Cancer Center
Temple University

NEURAL ENGINEERING DATA CONSORTIUM
www.nedcdata.org

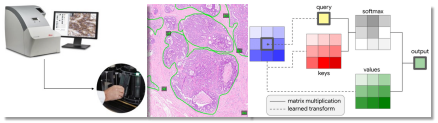
Fox Chase Cancer Center
Temple Health

Abstract

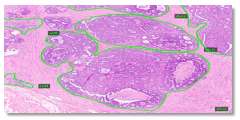
- The Neural Engineering Data Consortium (NEDC), has previously released the Breast Tissue subset of the Temple University Digital Pathology Corpus (TUDP), containing 3,505 partially annotated images.
- In this abstract, we introduce two recent additions to our digital pathology resources:
  - 14,288 images that were collected from Fox Chase Cancer Center's Biosample Repository
  - 97,755 images that are part of the TUDP Corpus
- The FCCC Corpus (FCDP) includes 48 metadata fields that provide information on the history of the sample, the patient's medical history and diagnosis codes.
- The TUDP Corpus includes a wide variety of tissue and stain types. The Breast Tissue subset has been reclassified to be consistent with FCDP.
- These two resources can be used to develop a new generation of machine learning technology that is more robust and can classify a wider range of data and pathology types.

Digital Pathology

- One of the unique challenges of digital pathology is the need to detect small regions of interest (e.g., 32x32 pixels) in extremely high resolution images (50Kx50K pixels):

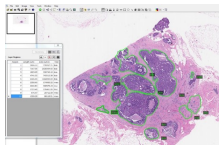


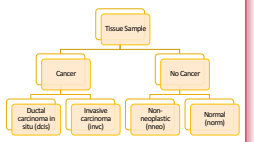
- Localization of information contributing to the diagnosis is a major goal of our work:



Labels		
norm	infl	indc
artf	susp	null
nneo	dcis	bckg

- The TUDP Breast Tissue Subset consists of 3,505 images annotated using a 4-way classification:





- There are 296 patients with 4.3% of the slides showing cancerous features.
- The average number of labels per image is 4.84.
- There are 8,895 cancerous labels, 5,362 labels with carcinogenic signs and 2,714 non-cancerous labels.

The FCCC Digital Pathology Corpus

- There are 14,288 images; the average file size is 244 MB; the entire corpus requires 3.5 TB of storage:

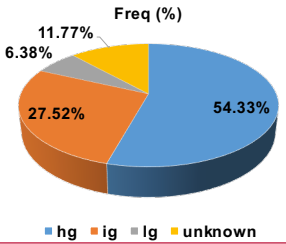
Component	Example	Description
directory name	fcdp/v1.0.0/svs/00026/00026256 4/001010285/c67.9_c67.2	a typical full directory specification
database	fcdp	an acronym for the corpus
version	v1.0.0	the version number
data type	svs	the root node for the image data and annotations
sequential ID	00026	a 5-digit number used to control directory complexity
subject ID	000262564	a 9-digit anonymized subject identifier (col. E)
specimen ID	001010285	a 9-digit specimen identifier (col. F)
ICDO codes	c67.9_c67.2	ICDO codes for tissue type (c67.9) and tissue site (c67.2) (cols. K and M respectively)
Component	**Example**	**Description**
filename	000368678_001013243_st065_xt01_t000.svs	a unique filename
identifier	000368678_001013243	subject ID (col. E) and specimen ID (col. F)
stain code	st065	a code assigned to the type of stain used in the slide (col. G)
slide number	xt1	a three-character code assigned to each slide (col. H)
token number	t000	a four-character number indicated the slide number in a series
extension	svs	indicates the type of file; ".svs" is used for an image data file

- Images are stored in Aperio's popular SVS format.
- There are 13,230 subjects in the corpus.
- Each image is accompanied by a csv file containing 48 metadata fields including diagnostic information:

Idx	Col	Name	Description	Example
11	K	ICDO Code (Tissue Site)	the International Classification of Diseases for Oncology (ICDO) code for the tissue site	C15.9
12	L	Tissue Site.1	the anatomical site of the tissue site corresponding to the ICDO code for the tissue site	Esophagus, NOS
29	AC	Histology	description of the cellular composition and characteristics of the tissue	Ductal Carcinoma NOS
32	AF	Block Level Tissue Histology	description of the block level tissue histology based on an examination and analysis of the tissue sample	Ductal Carcinoma NOS
33	AG	Gradeclin Desc	classification of the tumor grade based on clinical evaluation; indicates the degree of abnormality or aggressiveness of cancer cells compared to normal cells	Grade III Poorly Differentiated
34	AH	Gradepath Desc	description or classification of the tumor grade based on the degree of differentiation of the tumor cells; is a measure of how abnormal the cancer cells look under the microscope	Grade cannot be assessed (GX); Unknown

- The data has been classified according to a universal tumor grading system:

Grade Group	Gleason Score	Risk Group	Class
G1	≤ 6	low / very low	low (lg)
G2	7 (3+4)	intermediate	intermediate (ig)
G3	7 (4+3)		
G4	8	high / very high	high (hg)



The TU Digital Pathology Corpus

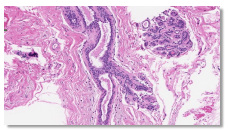
- There are 97,755 images; the average file size is 224 MB; the entire corpus requires 24 TB of storage:

Component	Example	Description
directory name	tudp/v1.0.0/svs/00036/aaaaaaaaa / bbbbbbbbbbbbbbb/c50.9_cxx.x	a typical full directory specification
database	tudp	an acronym for the corpus
version	v1.0.0	the version number
data type	svs	the root node for the image data and annotations
sequential ID	00036	a 5-digit number used to control directory complexity
subject ID	aaaaaaaaa	a 9-digit anonymized subject identifier (ID)
specimen ID	bbbbbbbbbbbbbbb	a 15-character specimen identifier (ID)
ICDO codes	c50.9_cxx.x	ICDO codes for tissue type (c50.9) and tissue site (cxx.x)
Component	**Example**	**Description**
filename	aaaaaaaaa_bbbbbbbbb_st065_a001_lvl000_t002.svs	a unique filename
identifier	aaaaaaaaa_bbbbbbbbbbbbbbb	subject ID (col. E) and specimen ID
stain code	st065	a code assigned to the type of stain used in the slide
slide number	a001_lvl000	a multi-character slide number code (4 + "_" + 6 characters)
token number	t002	a four-character number indicated the slide number in a series
extension	svs	indicates the type of file; ".svs" is used for an image data file

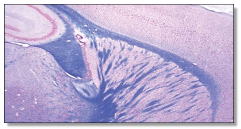
- A unified classification scheme is used to facilitate machine learning experiments using both corpora:

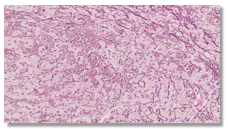
Class	Label	Description / Features
Low Grade (lg)	Normal (norm)	normal ducts and lobules
	Background (bckg)	stroma, no ducts or lobules
	Artifact (artf)	grease pen marks, stitches, foreign bodies, etc.
Intermediate Grade (ig)	Non-Neoplastic (nneo)	fibrosis, hyperplasia, intraductal papilloma, adenosis, ectasia, etc.
	Inflammation (infl)	areas of inflammation
	Suspected (susp)	regions that are at risk of developing into cancerous regions
High Grade (hg)	Ductal Carcinoma in Situ (dcis)	ductal carcinoma in situ, and lobular carcinoma in situ
	Invasive Ductal Carcinoma (indc)	invasive ductal carcinoma, invasive lobular carcinoma, and invasive mammary carcinoma
unknown	Indistinguishable (null)	indistinguishable tissue, normally due to issues with the cut/stain

- There are 192 unique tissue types and 324 unique stain types represented in our combined corpora.



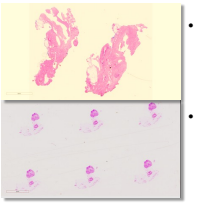
- Hematoxylin and eosin (H&E) stain is preferred for viewing cellular and tissue structure.
- Hematoxylin stains nuclei purplish blue; eosin stains the extracellular matrix pink. H&E demonstrates a broad range of features.
- Myelin is a layer that surrounds nerves. Damaged myelin slows down electrical impulses.
- Myelin cannot be detected with H&E stain. A solocherome cyanine stain is used. It stains cells and tissues blue/violet.



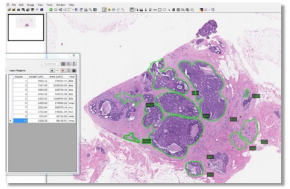


- Reticulin stain uses silver impregnation to detect reticulin fibers, which are made of type 3 collagen.
- It makes it easier to visualize areas of hepatocyte loss or regeneration.

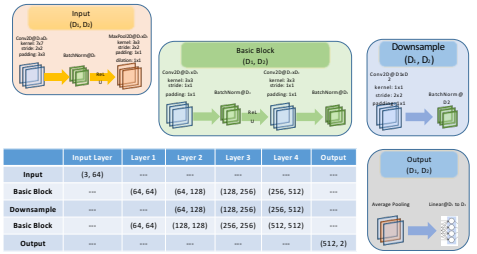
Segmentation and Localization



- Segmentation and localization of information is critical since the number of images and location of an image on a slide can vary.
- Localization requires scanning images at very high resolution, which poses computational and machine learning challenges.
- Annotation of an entire image is slow and expensive.
- We typically only annotate 5 to 10 regions of interest and a few examples of 'background'.



- We also provide a segmentation and classification system based on a ResNet-18 deep learning system:



	Input Layer	Layer 1	Layer 2	Layer 3	Layer 4	Output
Input	(3, 64)	---	---	---	---	---
Basic Block	---	(64, 64)	(64, 128)	(128, 256)	(256, 512)	---
Downsample	---	---	(64, 128)	(128, 256)	(256, 512)	---
Basic Block	---	(64, 64)	(128, 128)	(256, 256)	(512, 512)	---
Output	---	---	---	---	---	(512, 2)

Summary and Future Work

- This abstract introduces the release of two substantial resources pertinent to digital pathology, all of which are publicly available.
- We are currently in the process of annotating approximately 1,400 breast tissue images from FCDP to complement our TUDP release. These additional images will support a number of interesting studies including invariance and robustness.
- Future research is focusing on the use of unsupervised learning techniques to automatically classify and label the data. We are exploring ways to automatically detect stain and tissue type.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under grants nos. CNS-1726188 and 1925494, by the Temple University Office of the Vice President for Research, and by the Temple University College of Engineering Summer Research Experience for Undergraduates program. 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 or Temple University.

To learn more about our resources, please use this URL:
https://isip.piconepress.com/projects/nsf_dpath